FAIR LABS FRONTIER SERIES

VOL 01 · 2026

The Safety Horizon

The State of AI Safety 2026 — an annual assessment of how fast the frontier is moving, and whether our ability to assure it is keeping up

PREPARED FOR POLICYMAKERS · RESEARCH LEADERS · EXECUTIVES · CIVIL SOCIETY

FAIR Labs
FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

July 2026
FL-2026-01

FAIR LABS FRONTIER SERIES · VOLUME 01

The Safety Horizon: The State of AI Safety 2026 — an annual assessment of how fast the frontier is moving, and whether our ability to assure it is keeping up

The flagship volume of the Frontier Series takes stock of artificial intelligence safety at mid-decade. Its central finding is a widening capability–assurance gap: the pace at which frontier systems gain autonomy and reach has outrun the pace at which we can evaluate, interpret, and control them. The report inventories what has gone right, maps where the gaps are widening, sketches four trajectories to 2030, and closes with an eight-point agenda that assigns concrete work to every stakeholder with a hand on the wheel.

REPORT CODE	FL-2026-01
PUBLICATION DATE	July 2026
PRIMARY AUDIENCES	Policymakers · Research Leaders · Executives · Civil Society
SERIES	FAIR Labs Frontier Series, Fair Artificial Intelligence Research Labs
SUGGESTED CITATION	*The Safety Horizon: The State of AI Safety 2026 — an annual assessment of how fast the frontier is moving, and whether our ability to assure it is keeping up.* FAIR Labs Frontier Series, Vol. 01. Fair Artificial Intelligence Research Labs, 2026.
KEYWORDS	AI safety, frontier models, assurance, evaluations, alignment, governance, safety maturity

IN BRIEF · EXECUTIVE SUMMARY

Capability has become infrastructure. Assurance is still a craft.

Frontier AI systems now write production code, negotiate workflows, operate as agents across the open internet, and sit inside hospitals, banks, and ministries. The systems crossed from demonstration to dependency faster than any general-purpose technology on record — and faster than our tools for evaluating, interpreting, and controlling them have matured. That mismatch, not any single model, is the central safety fact of 2026.

01 **The capability–assurance gap is the master risk.**

Every specific concern this series examines — misuse, systemic fragility, labor shock, epistemic erosion — grows in the space between what systems can do and what anyone can verify about them. Closing that gap is the organizing goal of the safety field.

02 **Real progress happened, and it is worth defending.**

Frontier safety frameworks, national AI safety and security institutes, third-party evaluation ecosystems, and interpretability breakthroughs moved safety from aspiration toward engineering discipline. The scaffolding of a mature assurance regime now exists in outline.

03 **Agentic autonomy is outrunning the test bench.**

Evaluations built for chat-era models — static prompts, scored answers — cannot certify systems that plan, use tools, spend money, and act over hours. Control and monitoring of agents is the least mature assurance domain in our index, and the fastest-growing source of exposure.

04 **The gap is now a deployment problem, not just a lab problem.**

Most incidents in the coming years will originate downstream — in the thousands of businesses, agencies, and startups wiring capable models into consequential processes without the safety engineering the labs apply upstream.

05 **Verification is the missing institution.**

Aviation has flight recorders and independent crash investigators; finance has auditors and stress tests. AI still largely asks the world to take developers' word. Independent, adversarial, well-resourced verification is the highest-leverage institutional investment of the decade.

06 **The window is open, and it is a window.**

Nothing in our assessment says a safe transition is out of reach. Everything in it says the choices that determine which 2030 we get — Managed Ascent or Hard Landing — are being made now, mostly by people who do not think of themselves as making them.

CONTENTS

Inside this report

01

The year the frontier became infrastructure

For most of its history, AI safety was a debate about the future. In 2026 it is a description of the present: the systems are deployed, the dependencies are real, and the question is no longer whether to run powerful AI inside society, but on what terms.

Three shifts define the moment. The first is **ubiquity**. Frontier models stopped being destinations — a chat window you visit — and became substrates that other software is built on. They draft contracts inside law firms, triage tickets inside utilities, review radiology queues, tutor students, and write a meaningful share of the world's new code. Dependence arrived quietly, the way it did with electricity and cloud computing: one integration at a time, each individually sensible, collectively irreversible.

The second shift is **agency**. The defining products of the past two years are not better answers but longer leashes — systems that decompose goals, call tools, browse, purchase, write and execute code, and coordinate with other agents over hours or days. Agency multiplies usefulness, and it multiplies every safety question: an assistant that only speaks can at worst mislead; an agent that acts can transact, replicate workflows, and touch real systems. The unit of concern has moved from *outputs* to *behavior over time*.

The third shift is **stakes**. Frontier training runs now command national-scale capital, compute, and electricity. States treat model weights as strategic assets and datacenter siting as industrial policy. Safety debates that once lived in research workshops now run through export-control regimes, security services, and treaty diplomacy. AI safety has become a subset of statecraft — which brings resources, and brings racing dynamics with them.

WHY THIS REPORT EXISTS

Every stakeholder now holds a piece of the safety problem: labs hold the training runs, deployers hold the integrations, insurers hold the tail risk, regulators hold the rulebook, workers and citizens hold the consequences. Yet each group reads different documents, attends different conferences, and reasons from different maps. The Frontier Series is FAIR Labs' answer: one shared evidence base, rendered ten ways. This volume is the map; Volumes 02–10 are the territory, one stakeholder at a time.

Honest stocktaking requires holding two truths at once. The optimistic truth: capable AI is already delivering real value — compressing scientific workflows, extending expert scarcity in medicine and education, and beginning to pay what Volume 02 calls the abundance dividend. The sober truth: our collective ability to say *what these systems will do, why they did what they did, and how to stop them mid-mistake* has not kept pace with what they can do. The rest of this report measures that distance.

02

The capability–assurance gap

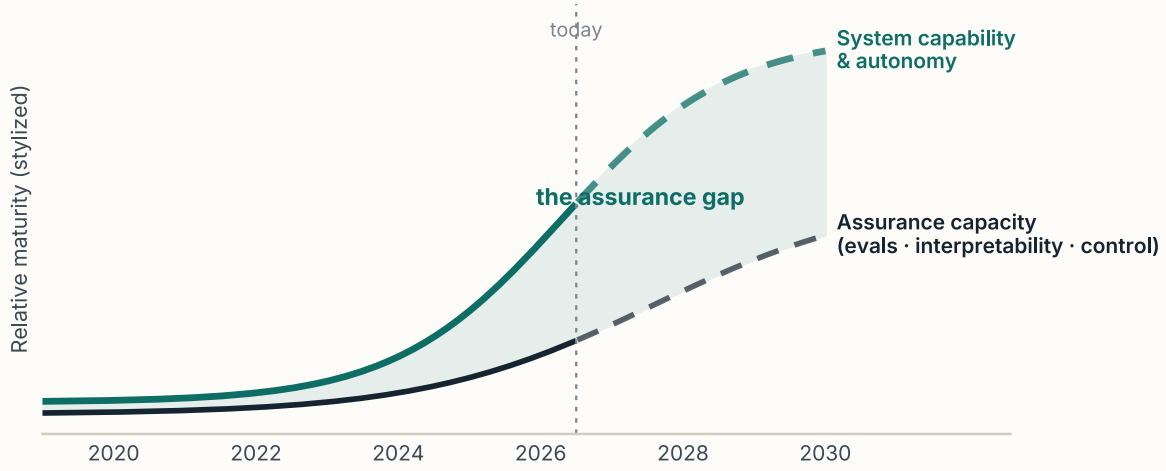
A single frame organizes everything in this series: capability compounds on an industrial curve, while assurance — the science of evaluating, interpreting, and controlling systems — compounds on an academic one. The distance between the curves is where risk lives.

Capability compounds industrially because it is fed by convergent inputs: capital seeking returns, compute scaling with fabrication cycles, algorithmic efficiency improving on top of hardware, and — increasingly — AI systems accelerating their own development pipeline. Each input reinforces the others. Progress is jagged and unevenly distributed, but the trend needs no heroics to continue: it is the default output of the market.

Assurance compounds academically. Evaluation science, interpretability, and control research advance through peer review, replication, and the slow accumulation of method — activities with long feedback loops and a global expert community that would fit in a mid-sized lecture hall. Assurance also faces a structural headwind capability does not: its target moves. Every capability jump partially invalidates the previous generation of tests, the way a new aircraft class invalidates old wind-tunnel data.

EXHIBIT 1.1 · CONCEPTUAL FRAMEWORK

Two curves, one gap: why the safety problem grows even as safety work succeeds

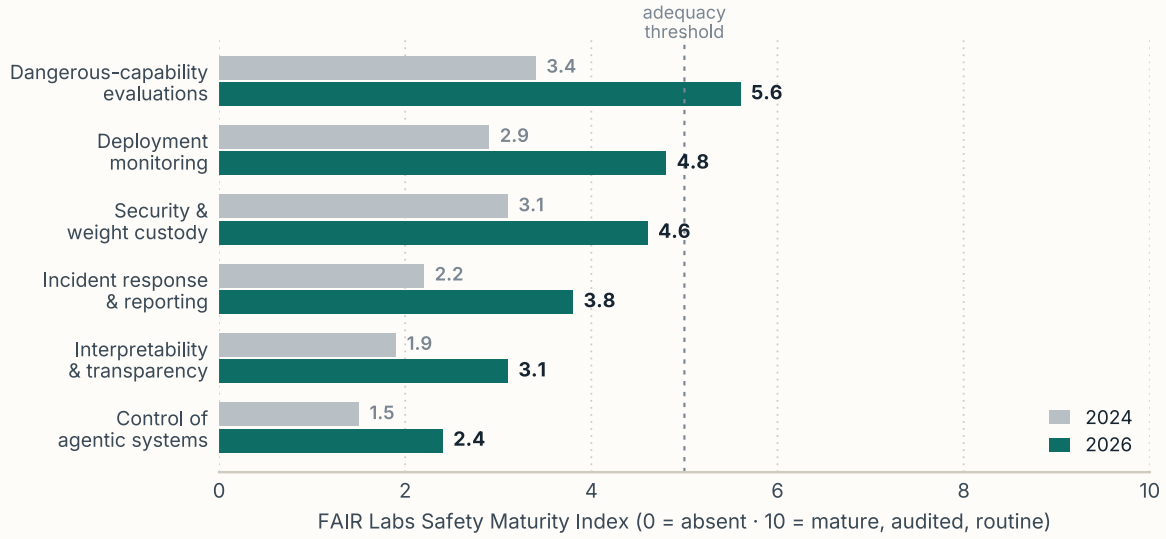


Stylized rendering of the report's core frame; axes are relative, not measured. Solid segments reflect FAIR Labs' reading of 2019–2026; dashed segments show momentum trajectories absent policy change. The gap is a claim about relative rates, not absolute levels.

To make the gap tractable, FAIR Labs assesses six assurance domains — the load-bearing walls of any credible safety regime — on a 0–10 maturity index scoring whether practices are *absent, emerging, standardized, or audited-and-routine*. Exhibit 1.2 shows the 2026 snapshot against 2024. The pattern is encouraging and alarming at once: every domain improved; none has crossed the adequacy threshold; and the domains improving slowest — interpretability and agent control — are precisely the ones the next capability generation leans on hardest.

EXHIBIT 1.2 · FAIR LABS ASSESSMENT (ILLUSTRATIVE INDEX)

Safety maturity by assurance domain, 2024 vs 2026



FAIR Labs analytical construct synthesizing published safety frameworks, evaluation practice, and expert consultation; scores support comparison and discussion, not measurement. "Adequacy" = practices sufficient for the capability level they must assure, independently verified.

6/6
assurance domains improved since 2024 — the work is working

0/6
domains past the adequacy threshold for today's frontier

2.4/10
maturity of agent control — the weakest wall, bearing the newest load

Two properties make this gap dangerous rather than merely untidy. It is **silent**: nothing visibly breaks while it widens, because assurance debt, like deferred maintenance, presents as savings until the day it presents as collapse. And it is **correlated**: the same untested failure mode ships simultaneously into thousands of deployments that share the same handful of base models — a monoculture dynamic explored in Volume 06. Gaps with those two properties do not announce themselves; they are discovered.

> *The gap will not be closed by asking capability to slow down and hoping it listens. It will be closed by making assurance compound industrially too — funded, staffed, instrumented, and institutionalized at the scale of the thing it must assure.*
>
> THE CENTRAL ARGUMENT OF THIS REPORT

03

What went right: the safety inventory

Pessimism is cheap and inaccurate. The period since 2023 built more genuine safety infrastructure than the previous two decades combined. An honest ledger starts with the assets.

3.1 Safety became a published, inspectable practice

The major frontier developers now operate under public safety frameworks — documents that define capability thresholds, commit to pre-deployment evaluation, and specify escalating safeguards as thresholds approach. These frameworks are imperfect, self-administered, and unevenly honored. They are also revolutionary in a precise sense: they convert safety from a private virtue into a public commitment that journalists, researchers, and regulators can hold against observed behavior. The gap between promise and practice became measurable — which is the first step to closing it.

3.2 States built institutional muscle

National AI safety and security institutes, first established in 2023–24 and since networked internationally, gave governments something they never had during previous technology transitions: in-house technical capacity to test frontier systems before and after release, rather than depending entirely on vendor assertions. The network remains under-resourced relative to its mandate, and its access to models is negotiated rather than guaranteed — but the institutional form now exists, and institutional forms are what survive political weather.

3.3 Evaluation grew an ecosystem

A functioning market of third-party evaluators, red-team firms, and benchmark consortia now probes frontier models for dangerous capabilities — from biosecurity uplift to cyber-offense to autonomous replication. Methodology matured visibly: from static question sets toward uplift studies with human subjects, agentic task suites, and pre-registered thresholds. Evaluation is becoming a discipline with careers, standards bodies, and reputations to lose — the sociological signature of a field that has begun to professionalize.

3.4 Interpretability produced usable instruments

Mechanistic interpretability graduated from curiosity to early toolchain. Feature-extraction techniques let researchers decompose model internals into human-auditable concepts and, in production settings, monitor for deception-adjacent activations and steer behavior directly. Coverage remains a small fraction of any frontier model's cognition — an instrument panel with most gauges still dark — but the proof of concept matters: the black box is not a law of nature. It is an engineering backlog.

THE ASSET WORTH PROTECTING

None of these gains was inevitable, and none is self-sustaining. Safety frameworks can decay into marketing; institutes can be defunded in a budget cycle; evaluator access can be quietly narrowed. The 2026 safety inventory is best understood not as a foundation but as scaffolding — load-bearing only if the permanent structure gets built while it stands.

04

Where the gaps are widening

Against the assets, five liabilities — each a place where the distance between capability and assurance grew over the assessment period, and each a preview of a later volume in this series.

4.1 Agents broke the test bench

Chat-era evaluation asks: *given this input, is the output acceptable?* Agentic systems invalidate the question. Their risk lives in trajectories — chains of dozens of decisions, tool calls, and environmental feedback in which no single step is alarming and the aggregate can be. Certifying trajectory-space requires methods closer to aviation certification and control theory than to exam grading: bounded operating envelopes, tripwires, interruptibility guarantees, and post-hoc flight recorders. Every element exists in prototype. None exists as standard. Meanwhile, agent deployment doubles on a cadence measured in quarters.

4.2 The frontier is no longer one thing

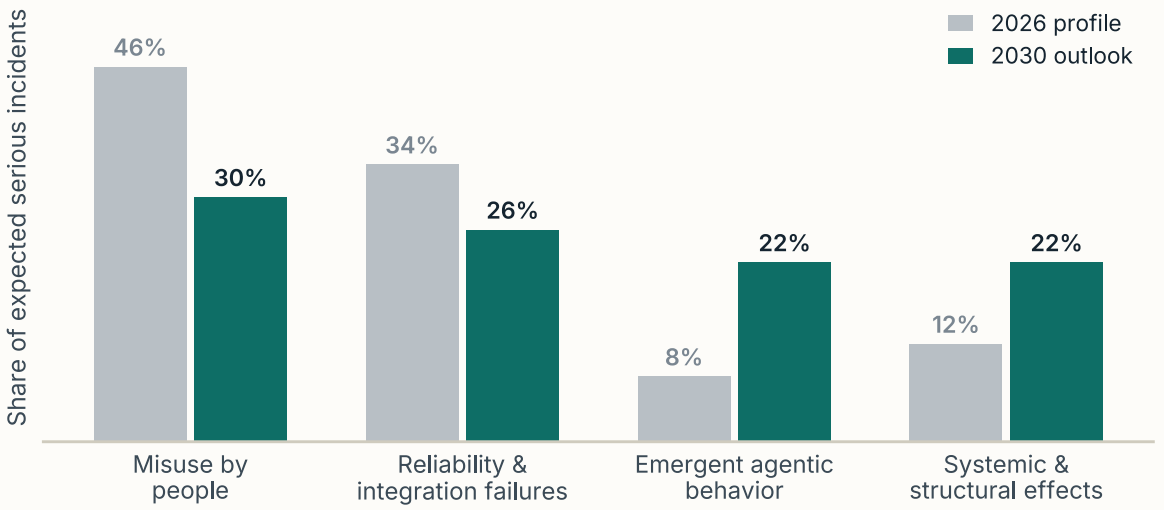
Highly capable open-weight models — released deliberately, leaked, or fine-tuned downstream — put near-frontier capability permanently outside any single actor's control. The genuine benefits are large: transparency, sovereignty, competition, and research access. So is the irreversibility. Safety strategies premised on a small number of controllable chokepoints must now coexist with a world where capability, once published, cannot be recalled — shifting the defensive burden from the model layer toward the compute, deployment, and societal-resilience layers examined in Volumes 07 and 06.

4.3 Deployment outran deployers

The marginal risk in 2026 is less the flagship model than the mid-market integration: capable systems wired into lending decisions, medical triage, industrial control, and government services by teams with no safety engineering tradition, thin monitoring, and vendor-supplied confidence. The labs' safety practices, whatever their limits, are the best in the ecosystem; the ecosystem's median practice is closer to nothing. Exhibit 1.3 sketches how FAIR Labs expects the incident profile to shift as a result.

EXHIBIT 1.3 · FAIR LABS OUTLOOK (ILLUSTRATIVE SHARES)

Where serious AI incidents come from: 2026 profile vs 2030 outlook



Analytical outlook expressing FAIR Labs' expectation of directional change, not a forecast with confidence intervals. "Serious incident" = harm meeting the reporting bar proposed in Volume 04, §3. Note the growth of emergent-behavior and systemic categories: the incidents hardest to attribute are the ones set to grow fastest.

4.4 Verification still runs on trust

Nearly every load-bearing safety claim in the ecosystem — training-compute figures, evaluation results, deployment safeguards, incident counts — is self-reported. The auditing profession that makes financial statements credible has no AI equivalent at scale: no standard evidence formats, no qualified-auditor pipeline, no legal privilege framework for examining proprietary systems, and no flight-recorder norm producing tamper-evident logs for post-incident reconstruction. Until verification is institutionalized, safety competition rewards the best narrative rather than the best practice — and honest actors cannot prove they are honest.

4.5 The talent asymmetry compounds

The researchers capable of frontier-relevant safety work remain concentrated inside the handful of organizations whose products they would assure — an arrangement with obvious strengths (access, context, resources) and an obvious structural flaw: societies do not let any other high-consequence industry employ essentially all of its own examiners. Building the independent bench — in institutes, universities, and evaluator firms, at compensation that survives contact with lab offers — is slow, unglamorous, and prerequisite to everything else in this agenda.

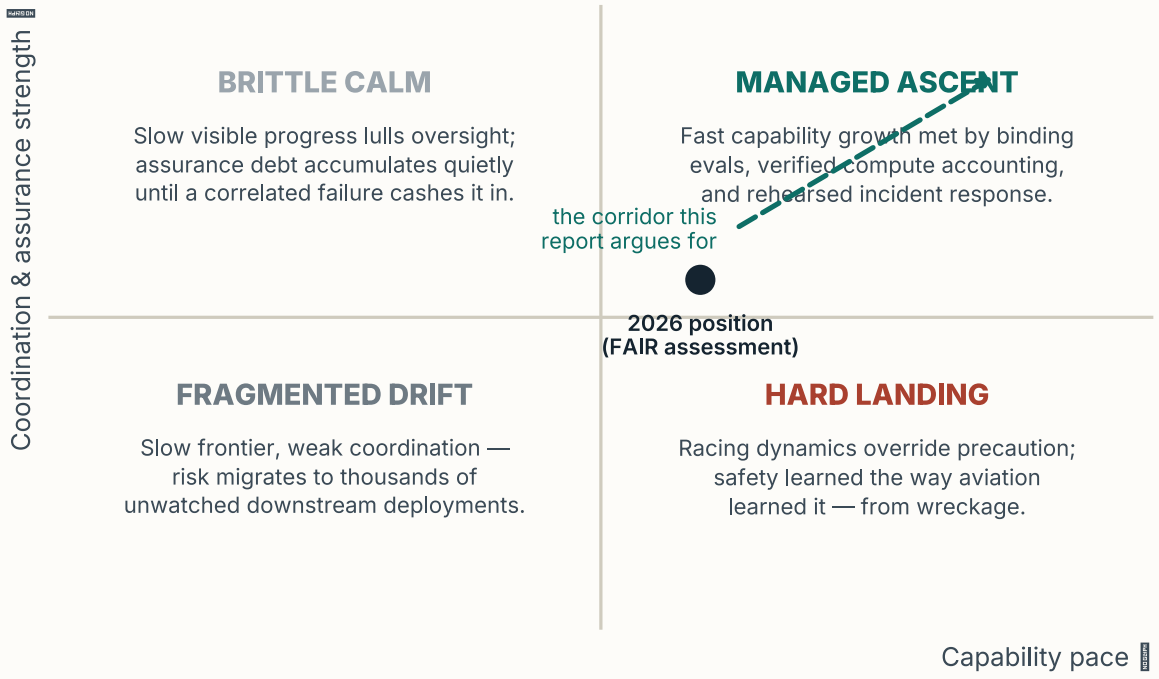
05

Four roads to 2030

Scenarios are not predictions; they are instruments for noticing which world you are entering while there is still time to change lanes. Two uncertainties dominate the road to 2030 — how fast capability moves, and how strongly coordination and assurance respond.

EXHIBIT 1.4 · SCENARIO FRAMEWORK

Capability pace × coordination strength: the 2030 possibility space



FAIR Labs scenario construct. The 2026 position marker reflects the report's qualitative assessment: capability pace moderately high and rising; coordination real but under-institutionalized. Small institutional investments move the vertical coordinate; little short of coordinated restraint moves the horizontal one.

Managed Ascent — the corridor this report argues for — is not a slow world. It is a fast world with instruments: binding evaluation gates at capability thresholds, verified compute accounting, rehearsed incident response, and insurance markets pricing assurance

quality. Its economy captures most of the abundance upside modeled in Volume 02, because trust is what lets capable systems be deployed into consequential domains at all.

Brittle Calm is the seductive failure: capability appears to plateau, urgency drains, safety budgets rotate to growth, and assurance debt accumulates behind stable-looking dashboards — until a correlated failure across shared infrastructure cashes it in all at once. Its warning signs are cultural, not technical: safety teams reporting to marketing, evaluation waivers becoming routine, incident disclosures becoming press releases.

Fragmented Drift disperses risk rather than concentrating it: coordination frays, capability diffuses through open weights and mid-tier labs, and no single actor is worth regulating because every actor is small. Aggregate harm arrives as chronic background noise — fraud, instability, erosion of epistemic trust — rather than as headline events, which is precisely why this scenario resists correction.

Hard Landing is the scenario the safety field was founded to prevent: racing dynamics — commercial or geopolitical — override precaution at exactly the capability level where precaution matters most, and the world learns AI safety the way it learned aviation safety, from wreckage. The report's judgment is blunt: Hard Landing is not the most likely road, but it is the only one from which recovery is not guaranteed, which is why its probability — not its inevitability — justifies the entire agenda that follows.

> **DASHBOARD FOR THE DESCENT**
>
> Five observable indicators tell readers which road they are on: (1) whether evaluation gates bind — do frontier deployments ever actually wait; (2) whether independent verifiers gain or lose model access year over year; (3) whether incident reporting grows more candid or more curated; (4) whether safety R&D spending tracks capability R&D spending or decouples from it; (5) whether international technical coordination survives its first genuine geopolitical stress test. This series will re-score all five in the 2027 edition.

06

The FAIR agenda: eight moves that close the gap

Each recommendation names an owner, because agendas without owners are moods. Fuller treatments appear in the volumes indicated; this is the whole program on one spread.

FA-01

Make assurance compound: fund it like the capability it must match

Treat evaluation science, interpretability, and control research as public-good infrastructure — multi-year funding lines, national-lab scale facilities, and procurement that pays for verified safety properties, not just capability benchmarks.

OWNERS: **LEGISLATURES · SCIENCE FUNDERS · FRONTIER LABS** · DEEP DIVE: VOL 05

FA-02

Build the verification profession

Standard evidence formats, qualified-auditor accreditation, legal privilege frameworks for examining proprietary systems, and tamper-evident logging norms — the flight-recorder-and- auditor layer AI currently lacks.

OWNERS: **STANDARDS BODIES · SAFETY INSTITUTES · BIG-FOUR-EQUIVALENT FIRMS** · DEEP DIVE: VOL 04

FA-03

Certify agents like operations, not essays

Shift agentic assurance to operating envelopes, interruptibility guarantees, authority budgets, and trajectory audits — with high-stakes domains gated on demonstrated control, not demonstrated cleverness.

OWNERS: **FRONTIER LABS · SECTOR REGULATORS · DEPLOYERS** · DEEP DIVE: VOLS 05–06

FA-04

Institutionalize incident learning

Mandatory, blameless-by-default reporting of serious AI incidents to an independent body with publication duties — because fields that cannot share their failures repeat them.

OWNERS: **LEGISLATURES · SAFETY INSTITUTES · INSURERS** · DEEP DIVE: VOL 04

FA-05

Use compute as the honest ledger

Verified training-compute accounting and hardware-level attestation as the backbone of threshold-based governance — the one input that is measurable, material, and hard to hide.

OWNERS: **GOVERNMENTS · CLOUD PROVIDERS · CHIPMAKERS** · DEEP DIVE: VOL 07

FA-06

Raise the deployment floor

Sector-specific minimum-assurance standards for consequential integrations — monitoring, fallback, human authority — so the ecosystem's median practice rises toward its best practice.

OWNERS: **SECTOR REGULATORS · INSURERS · ENTERPRISE BOARDS** · DEEP DIVE: VOLS 06, 10

FA-07

Keep the independent bench solvent

Career structures, salary parity mechanisms, and guaranteed model access for researchers outside the labs — the examiner class every other high-consequence industry maintains.

OWNERS: **PHILANTHROPY · UNIVERSITIES · SAFETY INSTITUTES · LABS** · DEEP DIVE: VOL 05

FA-08

Tie the safety story to the abundance story

Safety sustains political support only as the enabler of widely shared benefit — assurance is what permits deployment into medicine, energy, and education at all. Argue them together, always.

OWNERS: **EVERYONE IN THIS SERIES** · DEEP DIVE: VOLS 02, 08, 09

The through-line is institutional, not heroic: none of the eight moves requires a scientific miracle, a treaty renaissance, or anyone's unilateral disarmament. They require budgets, statutes, standards, and staffing — the boring machinery by which civilizations make dangerous things reliable. That is the good news. The machinery only works if it is built before it is needed. That is the deadline.

07

Reader's toolkit: terms that matter

Twelve terms, defined the way this series uses them — because half of every AI governance disagreement is two people using one word for three things.

Frontier model
A general-purpose AI system at or near the most capable end of the current distribution, typically identified by training compute, benchmark position, or developer designation.

Assurance
The umbrella discipline of establishing justified confidence in system behavior: evaluation, interpretability, monitoring, control, and verification, taken together.

Capability–assurance gap
This report's master frame: the distance between what systems can do and what anyone can verify, explain, or guarantee about what they will do.

Dangerous-capability evaluation
Structured testing for capabilities whose misuse or malfunction would be severe — bio-uplift, cyber-offense, autonomous replication, large-scale persuasion — ideally with pre-registered thresholds and consequences.

Agentic system
An AI system that pursues goals over multiple steps with tool use and limited supervision; the relevant unit of evaluation is the trajectory, not the response.

Operating envelope
The bounded set of actions, resources, and authorities within which an agent is certified to act — aviation's answer, imported.

Interpretability
Methods for explaining model internals in human-auditable terms; "mechanistic" variants aim at the circuit and feature level rather than post-hoc rationale.

Alignment
The property that a system's objectives and behavior track its operators' and society's intent — including under distribution shift, delegation, and optimization pressure.

Safety case
A structured, evidence-backed argument — borrowed from nuclear and aviation practice — that a specific system is acceptably safe for a specific deployment.

Open-weight model
A model whose parameters are publicly downloadable; capability released this way is irreversible, with benefits and burdens analyzed in §4.2.

Compute governance
Using the measurability of large-scale computation — chips, clusters, energy — as the verifiable substrate for AI oversight (Volume 07's subject).

Assurance debt
The accumulated liability of deploying capabilities faster than they are assured; invisible on dashboards, payable in incidents.

HOW TO READ THE REST OF THE SERIES

Executives should go next to Volume 10 (governance duties) and Volume 06 (deployment risk). Policymakers: Volumes 04 and 07. Researchers and funders: Volume 05. Economists and strategists: Volumes 02 and 08. Communicators, educators, and civil society: Volumes 09 and 03. Every volume is written to stand alone — and every volume assumes the frame established here.

Fair Artificial Intelligence Research Labs (FAIR Labs) is an independent research institute working to ensure that advanced artificial intelligence is developed safely, governed wisely, and shared fairly. Through rigorous technical research, policy analysis, and public engagement, FAIR Labs works with scientists, governments, companies, and communities to widen the circle of people who benefit from — and have a say in — the most consequential technology of our time.

The **Frontier Series** is FAIR Labs' flagship collection of stakeholder briefings on advanced AI — one shared evidence base, ten vantage points. Each volume stands alone; together they form a common map for scientists, executives, policymakers, workers, and citizens navigating the same transition.
