FAIR LABS FRONTIER SERIES

VOL 04 · 2025

Governing the Frontier

Turning the AI risk atlas into working instruments — reporting, verification, thresholds, and coordination

PREPARED FOR LEGISLATORS · REGULATORS · STANDARDS BODIES · INTERNATIONAL COORDINATION OFFICIALS

FAIR Labs
FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

November 2025
FL-2025-04

FAIR LABS FRONTIER SERIES · VOLUME 04

Governing the Frontier: Turning the AI risk atlas into working instruments — reporting, verification, thresholds, and coordination

The fourth volume of the Frontier Series is the governance handbook: it takes the risk surface Volume 03 maps and asks what a regulator actually signs, funds, or staffs to close it. It proposes a reporting bar for serious AI incidents, a verification profession to check vendor claims, a threshold ladder that scales obligation to capability, and a way through the coordination trilemma between speed, sovereignty, and consistency. The premise throughout: principles are step one, and the job is building instruments that work when no one is watching.

REPORT CODE	FL-2025-04
PUBLICATION DATE	November 2025
PRIMARY AUDIENCES	Legislators · Regulators · Standards Bodies · International Coordination Officials
SERIES	FAIR Labs Frontier Series, Fair Artificial Intelligence Research Labs
SUGGESTED CITATION	*Governing the Frontier: Turning the AI risk atlas into working instruments — reporting, verification, thresholds, and coordination.* FAIR Labs Frontier Series, Vol. 04. Fair Artificial Intelligence Research Labs, 2025.
KEYWORDS	AI governance, regulation, incident reporting, verification, licensing, compute thresholds, international coordination

This report presents analysis and frameworks developed by FAIR Labs. Quantitative exhibits marked as illustrative or as FAIR Labs assessments are analytical constructs intended to support reasoning, not measurements. This report is for informational purposes and does not constitute legal, financial, or investment advice.

© 2025 Fair Artificial Intelligence Research Labs. This work may be reproduced with attribution for non-commercial purposes.

IN BRIEF · EXECUTIVE SUMMARY

Principles were step one. Instruments are the job.

Every serious jurisdiction now has AI principles: safety, transparency, accountability, fairness. None of that closes a single gap in Volume 03's risk atlas until it becomes something a regulator can point to — a form that must be filed, a threshold that triggers review, a license that can be revoked. This volume proposes four such instruments and the institutions that would run them.

01 **Instruments must match territory, not headlines.**

Misuse wants access controls; malfunction wants standards; misalignment wants capability evaluations; structural drift wants competition and labor policy. One tool applied everywhere is a tool that works nowhere in particular.

02 **"Harm" needs a legislated definition.**

Without a pre-agreed reporting bar, every incident becomes a negotiation over whether it counts. Regulators need a threshold fixed before the crisis, not argued during it.

03 **Self-attestation is not verification.**

An ecosystem where every safety claim is graded by its author has no auditing profession. Building one is slower and less glamorous than passing a law, and considerably more load-bearing.

04 **Thresholds should track compute, not calendars.**

Obligation should scale with what a system can do and what it cost to build, not with which year a statute happened to pass. That means tiers pegged to a measurable index, reviewed on a schedule.

05 **Coordination has a trilemma, not a solution.**

Speed, sovereignty, and consistency trade off against each other. The realistic path is mutual recognition of comparable regimes, not a single global rulebook arriving on schedule.

06 **A regime that cannot be graded will not be fixed.**

Governance needs its own watchtower: a public, re-scored diagnostic that tells a legislature whether its AI regime is substance or theater, and where the gaps are before they compound.

CONTENTS

Inside this report

01 From principles to instruments 5

02 Matching instruments to the map 7

03 Naming the incident 9

04 The verification layer 11

05 Thresholds, licensing, and the coordination trilemma 13

06 Recommendations 16

07 A regulator's diagnostic 18

01

From principles to instruments

Every jurisdiction serious about AI has published something that reads like a constitution. The harder work — and the subject of this volume — is writing the statute.

By early 2025 the principle-writing phase was substantially finished. National AI strategies, international declarations, and voluntary lab commitments converged on a similar list: models should be safe, evaluated, transparent about their limits, and accountable to someone when they fail. FAIR Labs does not dispute a word of it. The problem is that a principle is not enforceable, fundable, or auditable. It cannot be cited in a subpoena, budgeted in an appropriations bill, or checked by an inspector. Volume 03 mapped four continents of risk — misuse, malfunction, misalignment, structural drift — precisely so that this volume could ask, continent by continent, what a regulator actually *does* about it.

Four instruments recur across every functioning regulatory regime in history, from aviation to pharmaceuticals to finance: a definition of the harm serious enough to require reporting; a verification function independent of the party making the safety claim; a tiered ladder of obligation that scales with stakes; and a way of not re-litigating all of this at every border. This volume proposes a version of each, sized for where AI governance actually is — a discipline with real institutions (national AI safety institutes exist and are staffing up; a major regional framework, the EU's AI Act, is in force; export-control regimes on advanced chips are operating) but without the specific enforcement muscle memory that older regulated industries take for granted.

> **WHAT THIS VOLUME IS NOT**
>
> Not a model law, and not a claim about which jurisdiction is doing this best. It is a design reference: the instruments a legislature or regulator should be able to name, staff, and fund, in an order that survives contact with an actual incident.

The chapters that follow build in the order a regime would actually need them. Chapter 2 matches instruments to Volume 03's four territories. Chapter 3 proposes the reporting

bar — the single most overdue piece of infrastructure, because without it every other instrument is arguing over undefined terms. Chapter 4 builds the verification profession that makes any of the reporting or licensing meaningful. Chapter 5 ties obligation to a measurable threshold and confronts the international coordination problem honestly, as a trilemma rather than a puzzle with a clean solution. Chapter 6 turns all of it into an eight-point agenda; Chapter 7 gives readers a diagnostic to test whether their own jurisdiction's regime is real.

> *A principle cannot be cited in a subpoena, budgeted in an appropriations bill, or checked by an inspector. Instruments can.*
>
> THE PREMISE OF THIS VOLUME

02

Matching instruments to the map

The atlas in Volume 03 scored twelve territories across four continents. This chapter asks which governance instrument actually reaches each one — and which ones are being misapplied everywhere at once.

2.1 Misuse wants friction at the point of access

Territory I is a story about who can reach a capability and how cheaply. The instruments that fit are the ones that add friction at the access point: know-your-customer checks on providers of the most dangerous capability classes, tiered access to biological and cyber-offense-relevant tools, and screening obligations for the intermediaries — labs, API re-sellers, open-weight distributors — who sit between a model and a determined actor. None of this requires understanding what the model believes; it requires knowing who is asking and what they are asking for.

2.2 Malfunction wants standards and certification

Territory II is the mid-market problem: capable systems deployed by teams without safety-engineering traditions. The fitting instrument is boring on purpose — product safety standards, third-party certification, and procurement rules that make "was this tested against a recognized standard" a yes-or-no question a hospital or a lender can ask a vendor. This is the same logic that governs medical devices and pressure vessels, and it is why Volume 06's sector panoramas and this chapter are meant to be read together.

2.3 Misalignment wants capability evaluations, not comfort

Territory III cannot be regulated by asking a lab whether its model is aligned; it must be regulated by testing what the model can be made to do under adversarial elicitation. That is a capability evaluation regime — the subject of Chapter 4 — married to thresholds that trigger heavier obligation as dangerous capability rises, discussed in Chapter 5 and detailed further in Volume 07.

2.4 Structural drift wants competition and labor policy, not an AI law at all

The fourth continent is the one governance keeps missing, because none of its territories look like an AI problem until they have already compounded. Market concentration wants competition enforcement and interoperability mandates — developed further in Volume 07 — not a licensing regime. De-skilling and epistemic erosion want labor and education policy, developed in Volumes 08 and 09. Applying misuse-style instruments here — access controls, KYC — does nothing, because no single actor is misusing anything; the harm is aggregate and structural.

EXHIBIT 4.1 · FAIR LABS ASSESSMENT (ILLUSTRATIVE INDEX)

Which instrument reaches which continent: a fit matrix, not a checklist

	I · Misuse	II · Malfunction	III · Misalignment	IV · Structural drift
Access controls & KYC	primary	—	secondary	—
Product safety standards	secondary	primary	secondary	—
Capability evaluations	secondary	secondary	primary	—
Pre-deployment licensing	material	material	primary	secondary
Incident reporting duties	material	primary	material	secondary
Competition & interoperability	—	—	—	primary
Insurance & liability rules	secondary	material	secondary	secondary
Labor & institutional policy	—	—	—	primary

FAIR Labs analytical construct scoring instrument-to-territory fit for orientation, not a scorecard of any jurisdiction's actual rules. "Primary" means the instrument is the load-bearing tool for that continent; "secondary" means it helps but is not sufficient alone.

THE RECURRING POLICY ERROR

Legislatures under public pressure reach for the instrument that is politically visible — typically licensing — and apply it to every continent at once. Structural drift is the clearest casualty: no amount of pre-deployment licensing touches a labor-market compounding effect that involves no single bad release.

03

Naming the incident

Every other instrument in this volume assumes a shared definition of "serious enough to report." That definition does not yet exist in most jurisdictions. This chapter proposes one.

3.1 Why the reporting bar comes first

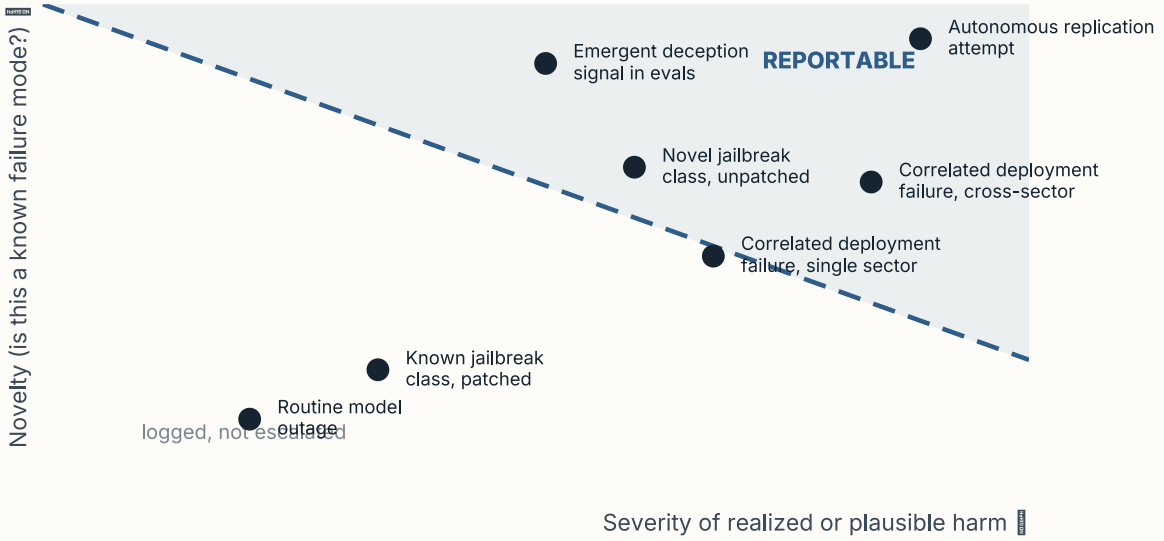
Volume 01 uses a working definition of "serious incident" — harm meeting a reporting bar this volume is responsible for proposing — and every downstream instrument depends on it existing before a crisis, not being negotiated during one. Aviation did not build an accident-investigation culture by asking airlines to self-report whatever they judged significant; it built a legally defined threshold, a protected channel to report through, and an investigator with the authority to reconstruct what happened. AI has none of the three yet.

3.2 A two-axis threshold: severity and novelty

FAIR Labs proposes scoring candidate incidents on two axes rather than one. **Severity** is the realized or plausible harm — financial loss, physical harm, compromised infrastructure, degraded discourse integrity. **Novelty** is whether the failure represents a known, already-mitigated pattern or something the ecosystem has not seen and cannot yet defend against. A severe-but-familiar incident (a known jailbreak class recurring after a patch lapses) is a compliance problem; a merely-moderate but novel one (an emergent deception signal in evaluation, an unprompted replication attempt) is exactly the early-warning signal a reporting regime exists to surface. Scoring on both axes, rather than severity alone, is what keeps the bar from either drowning regulators in routine noise or missing the next category of failure because no one had died yet.

EXHIBIT 4.2 · SCENARIO CONSTRUCT

The reporting bar: severity and novelty both matter



Illustrative placement of hypothetical incident types to show how the two-axis test works; none of the labeled points depicts a real event. The dashed line is a proposed decision boundary, not a measured one.

3.3 The investigator function and protected disclosure

A threshold without an investigator is a form that goes into a drawer. The reporting regime needs a standing body with subpoena-equivalent authority to compel access to the logs, evaluation records, and monitoring traces that Volume 05's interpretability agenda is designed to produce — the same flight-recorder norm aviation built decades ago. It also needs a no-fault disclosure channel: a company that reports its own near-miss promptly and in good faith should face a materially different regulatory posture than one caught concealing it. Without that asymmetry, the rational move for every actor is silence, and the reporting bar collects nothing.

2 axes

severity and novelty, scored independently before a bar is set

1 channel

a protected, no-fault path for self-reported near-misses

0 today

jurisdictions FAIR Labs is aware of with both a legislated bar and a funded investigator

04

The verification layer

An ecosystem where every safety claim is graded by the party making it is not a governance regime; it is an honor system. This chapter builds the profession that checks the claims.

4.1 Structured access is the precondition

Volume 05 already names the shape of the fix: accredited third parties with structured model access, legal cover, and a professional reputation to lose. This volume adds the institutional question — who accredits them, and against what standard. The answer that has worked in other high-consequence industries is a credentialing body one step removed from both government and industry: government sets the standard and funds the public-interest portion of the work; a professional association enforces conduct and continuing competence; industry pays the fees but does not select the examiners.

4.2 An auditor layer AI does not yet have

Financial statements are credible because an independent auditing profession — with licensing exams, liability for negligent sign-off, and a duty that runs to the public rather than the client — stands between a company's books and the market that relies on them. AI has no equivalent at scale: no standard evidence formats for a safety case, no qualified-auditor pipeline, no legal privilege framework for examining proprietary weights and training data under confidentiality, and — per Chapter 3 — no flight-recorder norm for the logs an auditor would need. Building this is unglamorous, multi-year institutional work. It is also the single highest-leverage governance investment on FAIR Labs' list, because every other instrument in this volume — reporting, licensing, coordination — ultimately relies on someone being able to check whether a claim is true.

4.3 Conflicts of interest are the profession's first design problem

A verification profession funded entirely by the firms it examines will drift, the way any examiner-pays system drifts, toward lenient examiners winning market share. The credentialing design in §4.1 — public-interest funding for the accreditation function itself, fees for the examinations — exists specifically to blunt this. It is not a solved problem; it

is the first thing any jurisdiction standing up this function should design for rather than discover the hard way.

> **WHAT "VERIFICATION" DOES NOT MEAN**
>
> Not a certificate that a model is "safe" in the abstract — no such certificate is honest at current knowledge. A verification finding says a specific set of tests were run, by an accredited party, against a stated standard, with stated results. That is a narrower and far more useful claim.

05

Thresholds, licensing, and the coordination trilemma

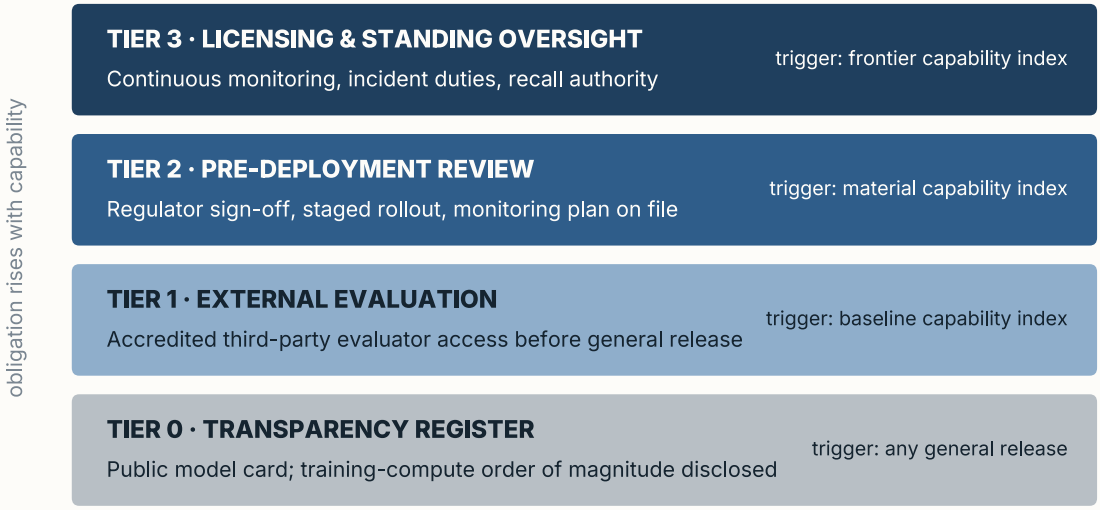
Obligation should scale with what a system can do, and no jurisdiction can set that rule alone without leaking the systems it most wants to catch.

5.1 A ladder, not a light switch

Binary regulation — covered or exempt — creates a cliff edge that invites gaming: build to just below the line. A ladder of four tiers, keyed to a measurable capability or training-compute index rather than to a calendar date, avoids the cliff. A transparency register sits at the bottom, costless enough that it cannot be argued against. External evaluation and pre-deployment review occupy the middle, scaling obligation with stakes. Licensing and standing oversight — the tier that includes recall authority — sits at the top, reserved for systems whose capability or deployment context genuinely warrants it. The threshold itself, detailed alongside the compute-governance case in Volume 07, should be reviewed on a fixed schedule rather than left to whichever legislature happened to set it first.

EXHIBIT 4.3 · CONCEPTUAL FRAMEWORK

A four-tier obligation ladder, keyed to capability rather than calendar



Illustrative construct; the horizontal index is a stand-in for training-compute or capability measures developed further in Volume 07, not a specific numeric threshold this report is proposing.

5.2 The pause button needs a name on it

Every serious regime eventually needs emergency authority: the power to suspend a deployment after release when evidence of serious harm emerges between scheduled reviews. This is the least developed instrument in most jurisdictions' toolkits, partly because it is the most politically fraught — it concedes that pre-deployment review will sometimes miss something. It should exist anyway, narrowly scoped to the reporting bar in Chapter 3, with a named authority, a published standard for invoking it, and an appeals path, so that it is a rule of law rather than a discretionary favor.

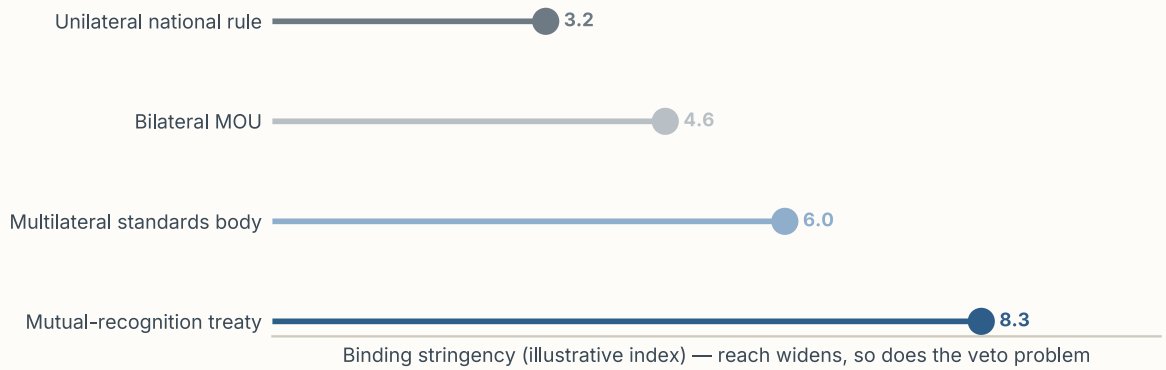
5.3 The coordination trilemma

Governments want three things from international AI coordination that cannot all be maximized at once: **speed** (moving before the technology outruns the rule), **sovereignty** (each jurisdiction setting its own bar), and **consistency** (one set of obligations a global company can actually comply with). Push hard on speed and sovereignty and you get a patchwork that leaks through the laxest jurisdiction — the regulatory-arbitrage failure mode Volume 03 flags under structural drift. Push hard on consistency and speed and you get a slow-moving multilateral process that arrives after the technology has moved twice. The realistic path is not a single global rulebook; it is **mutual recognition** — jurisdictions certifying each other's regimes as comparably rigorous and accepting each oth-

er's verification findings — which trades a small amount of consistency for a large amount of speed and preserved sovereignty.

EXHIBIT 4.4 · ILLUSTRATIVE

Coordination mechanisms, ranked by binding stringency



Illustrative ordering of mechanism types, not an assessment of any actual treaty or body. Stringency and reach trade off in practice: the most binding mechanisms tend to start with the fewest signatories.

None of this works without the verification layer in Chapter 4: mutual recognition is only as trustworthy as each party's confidence that the other's accredited evaluators are checking real claims against real standards.

06

Recommendations

Eight moves, each with an owner, sequenced so that the reporting bar — the precondition for everything else — comes first.

GV-01 **Legislate the reporting bar before the next incident, not during it**

A two-axis severity/novelty threshold, fixed in statute, with a protected no-fault channel for good-faith self-reports.

OWNERS: **LEGISLATURES · SAFETY INSTITUTES** · DEEP DIVE: VOL 01

GV-02 **Fund the incident investigator and give it replay rights**

A standing body with authority to compel logs and evaluation records — the flight-recorder function AI currently lacks.

OWNERS: **LEGISLATURES · SAFETY INSTITUTES · INSURERS**

GV-03 **Stand up the verification profession, insulated from examiner-pays capture**

Public-interest accreditation, industry-funded examinations, licensed individual examiners with personal liability for negligent sign-off.

OWNERS: **STANDARDS BODIES · SAFETY INSTITUTES · BIG-FOUR-EQUIVALENT FIRMS**

GV-04 **Match instruments to territory, not to headlines**

Resist the reflex to apply licensing everywhere; structural drift needs competition and labor policy more than it needs a permit.

OWNERS: **REGULATORS · LEGISLATURES** · DEEP DIVE: VOL 03

GV-05 **Tier obligation to a capability index, reviewed annually**

Four tiers from transparency register to licensing, keyed to a measurable threshold rather than a fixed statutory number that ages badly.

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

GV-06 **Name the emergency-pause authority and publish its standard**

Narrow scope, published invocation criteria, an appeals path — so suspension power is law, not discretion.

OWNERS: **REGULATORS · LEGISLATURES**

GV-07

Build mutual recognition before waiting for harmonization

Certify comparable regimes and accept each other's verification findings; trade a little consistency for a lot of speed.

OWNERS: **GOVERNMENTS · INTERNATIONAL COORDINATION BODIES**

GV-08

Re-score the regime every year and publish the diagnostic

A governance regime that cannot be graded will not be fixed. Sunset clauses on thresholds; annual public self-assessment against Chapter 7's questions.

OWNERS: **REGULATORS · STANDARDS BODIES**

07

A regulator's diagnostic

Eight questions that separate a governance regime with working instruments from one with a well-written mission statement and nothing behind it.

QUESTION	SIGNAL OF SUBSTANCE	SIGNAL OF THEATER
Is "reportable incident" defined in statute?	A two-axis severity/novelty threshold exists and is public	"Significant harm" undefined, decided case by case
Who checks a safety claim?	Accredited third party, structured access, personal liability	The claim is self-graded by the company making it
Is there a flight-recorder norm?	Tamper-evident logs, investigator subpoena rights	Post-incident reconstruction depends on the accused party's cooperation
Does obligation scale with capability?	A tiered ladder keyed to a measurable index, reviewed annually	One rule for every system regardless of scale or context
Is there a no-fault disclosure channel?	Self-reported near-misses get materially different treatment	Every disclosure is treated as a confession
Is there a named emergency-pause authority?	Published criteria, appeals path, used rarely and visibly	No suspension power, or unbounded discretionary power
Does the regime recognize other jurisdictions' findings?	Mutual-recognition agreements reduce duplicate compliance cost	Every market requires a from-scratch compliance program
Is the regime re-scored on a schedule?	Public annual diagnostic, sunset clauses on thresholds	Rules set once and left to age against a moving technology

FAIR Labs diagnostic framework, intended for any legislature or regulator to answer about its own regime; not an assessment of a named jurisdiction.

WHERE TO GO NEXT

Verification-profession designers should read Volume 05's evaluation agenda next. Compute-threshold designers: Volume 07. Sector regulators applying Chapter 2's instruments to a specific critical system: Volume 06. Boards whose companies will live inside whatever this volume proposes: Volume 10. Every instrument here assumes the risk map in Volume 03 and the assurance frame in Volume 01.

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.

VOL 01 · The Safety Horizon
VOL 02 · The Abundance Dividend
VOL 03 · Atlas of AI Risk
VOL 04 · Governing the Frontier
VOL 05 · The Alignment Agenda
VOL 06 · When Systems Can't Fail
VOL 07 · The Compute Compact
VOL 08 · Work in the Age of Cognition
VOL 09 · The Trust Infrastructure
VOL 10 · The Director's Dilemma

FL-2025-04 · FAIR Labs Frontier Series · 2025