FAIR LABS

FAIR ARTIFICIAL INTELLIGENCE RESEARCH LABS

POLICY REPORT ◆ AI SAFETY ◆ 2026

AI Safety Framework for Public Institutions

A structured approach to evaluating, procuring, deploying, and overseeing artificial intelligence in government and enterprise environments.

PROGRAM	DOCUMENT TYPE	AUDIENCE	PUBLISHED
AI Safety	Policy Report	Government & Enterprise	2026

FAIR Labs · A 501(c)(3) nonprofit research institute · Washington, DC

CONTENTS

Table of Contents

EXECUTIVE SUMMARY

The bottom line for decision-makers

IN ONE PARAGRAPH

Institutions do not need to become AI research labs to deploy AI safely. They need a disciplined way to decide *which* systems warrant *which* level of scrutiny, a catalog of controls tied to the stages of a system's life, a procurement process that extracts honest evidence from vendors, clear lines of human accountability, and a monitoring regime that treats deployment as the beginning of oversight rather than the end. This report supplies each of these as a ready-to-adapt instrument.

3
risk tiers that determine the depth of every downstream control

6
lifecycle stages, each with named owners and exit criteria

5
maturity levels charting the path from ad-hoc to institutionalized

What we recommend

1. **Tier before you build or buy.**
Classify every candidate system by consequence-of-failure and reversibility before any technical evaluation. The tier sets the evidentiary bar for everything that follows.

2. **Make the vendor prove it.**
Shift the burden of demonstrating safety onto suppliers through structured documentation requirements, acceptance testing, and contractual commitments to post-deployment cooperation.

3. **Name a human who is answerable.**
Every deployed system must have a designated accountable official with the authority to pause it. Oversight that no one owns is oversight that does not happen.

4 **Treat deployment as day one.**
Fund monitoring, drift detection, and incident response as recurring line items, not one-time project costs. Safety is a subscription, not a purchase.

The remainder of this report develops each recommendation in operational detail, with matrices and checklists an institution can lift directly into policy.

SECTION I

The Governance Gap

Every wave of institutional technology has arrived with a lag between capability and control. Mainframes were installed years before information-security practice matured; the web reached agencies long before records-management law caught up. Artificial intelligence has compressed that lag into something more dangerous, because the technology does not merely execute instructions—it produces judgments. When a benefits system, a hiring screen, or a fraud detector is powered by a learned model, the institution is delegating discretion that once belonged to officials. Discretion delegated without governance is discretion abandoned.

The gap is not primarily a gap in enthusiasm. Public leaders understand that AI can shorten queues, surface patterns in overwhelming caseloads, and extend scarce expertise. The gap is in the scaffolding that turns enthusiasm into responsible practice: the shared vocabulary for talking about risk, the decision rights that determine who may approve a deployment, the evidence standards that separate a vendor's marketing from a vendor's proof. Without that scaffolding, institutions oscillate between two failure modes—reckless adoption that ships harm at scale, and reflexive prohibition that forfeits genuine public benefit.

Why conventional IT governance is not enough

Institutions already govern software. They have change-management boards, security reviews, and procurement rules. It is tempting to assume these suffice for AI. They do not, for three structural reasons that recur throughout this report.

Learned behavior resists specification. Traditional software is governed against a specification: the system is correct if it does what the requirements say. A learned model has no complete specification of its behavior; it has a training objective and a distribution of inputs it was optimized over. Correctness gives way to *calibration*—how well the system performs across the range of situations it will actually encounter, including those its builders never imagined. Governance must therefore ask empirical questions the change-management board is not equipped to pose.

Failure is often silent. A crashed server announces itself. A model that has quietly grown less accurate for one neighborhood, one dialect, or one disability does not. The harm accrues in the space between the metric the institution watches and the outcome the public experiences. Governance must actively look for harm rather than wait for it to page someone.

The system keeps changing after purchase. Vendors update models. The population shifts. Upstream data pipelines are re-plumbed. A control regime that certifies a system once, at acceptance, and then trusts it indefinitely is certifying a photograph of a river.

A WORKING DEFINITION

Throughout this report, an **AI system** means any deployed capability whose behavior is materially shaped by a statistical model learned from data, together with the data pipelines, human procedures, and interfaces that surround it. The unit we govern is the *system*, not the model file.

The cost of getting it wrong—in both directions

It is conventional to frame AI risk in terms of spectacular failures, and those matter. But the more common institutional harm is mundane and cumulative: a translation tool that subtly mangles the languages of the communities least able to complain; an eligibility model that is right on average and wrong for the vulnerable tail; a "decision-support" tool that operators learn to defer to because overriding it is more work than accepting it. These failures rarely make headlines. They erode the one asset a public institution cannot easily rebuild—legitimacy.

Equally real is the cost of paralysis. An agency that cannot distinguish a low-stakes productivity tool from a high-stakes adjudication system will tend to govern both as if they were the latter, smothering harmless efficiency in process while, paradoxically, still lacking the capacity to scrutinize the deployments that truly warrant it. Proportionality is not a convenience; it is what makes rigorous oversight affordable where it counts.

The remainder of this report is an attempt to close the governance gap without falling into either ditch. We begin by making the central term precise.

SECTION II

What "AI Safety" Means for an Institution

Safety is not a score a model earns in a lab. It is a property of the whole sociotechnical system operating in a specific institutional context, judged against the harms that context makes possible.

Vendors and researchers often speak of model safety as though it were intrinsic—an attribute measured once and carried with the model like a crash rating. For an institution this framing is actively misleading. The same language model is safe as an internal drafting aid and unsafe as an unsupervised source of legal determinations. Safety lives at the join between what the system can do and what the institution lets it decide.

Five dimensions of institutional AI safety

We find it useful to decompose the diffuse notion of "safety" into five concrete dimensions. Each is measurable, each maps to controls introduced later, and each can fail independently of the others.

THE FIVE SAFETY DIMENSIONS AND THE QUESTION EACH ONE ASKS

Dimension	The question it forces	Typical failure when neglected
Validity	Does the system actually do what it claims, at the accuracy it claims, on our population?	Impressive demos that degrade on real caseloads.
Robustness	How does it behave under unusual, adversarial, or out-of-distribution inputs?	Confident, wrong answers on edge cases and manipulation.
Fairness	Are error rates and outcomes equitable across the groups we serve?	Uniform averages hiding concentrated harm in a subgroup.
Transparency	Can we explain a given decision to the person it affects and to an auditor?	Unappealable "computer says no" determinations.
Controllability	Can a human understand, override, and if necessary halt the system in time?	Automation the staff cannot practically contest.

The value of the decomposition is that it turns an argument about whether a system is "safe" into a structured inventory of specific, answerable questions. A procurement panel that asks these five questions, in order, will surface more than a panel debating safety in the abstract.

Safety as a sociotechnical property

Consider a resume-screening tool that a vendor has tuned to statistical parity across protected groups. On the model's own terms it is "fair." Now place it in an institution where recruiters see its ranking first, treat the top of the list as a shortlist, and lack the time to review candidates it buried. The human process has re-introduced exactly the disparity the model was tuned to remove, because the people around the model behave in predictable ways under time pressure. No amount of model-level fairness fixes this; the remedy is a change to the workflow, not the weights.

> *A model is not safe or unsafe. A deployment is. The institution owns the deployment.*
>
> — The governing premise of this framework

This is why our framework refuses to certify models. It certifies deployments, and it insists that the assessment of any deployment account for the humans in the loop: what they see, what they are incentivized to do, how much time they have, and what happens when they disagree with the machine. A safety case that omits the operators is a safety case about a system that does not exist.

What safety is not

Two clarifications prevent common misuse. Safety is *not* the same as accuracy: a highly accurate system can be unsafe if its rare errors are catastrophic or land on the same vulnerable people every time. And safety is *not* the same as compliance: meeting a checklist can leave real hazards untouched, while a thoughtful institution can be safe in ways no checklist anticipated. We use checklists in this report as scaffolding for judgment, never as a substitute for it.

SECTION III

The FAIR Risk-Tiering Matrix

Proportionality is the engine of an affordable governance program. If every system receives the same scrutiny, scrutiny becomes a rubber stamp, because no institution can afford to examine a meeting-transcription tool as rigorously as a system that decides who receives medical priority. The first act of governance, therefore, is triage: assigning each system to a risk tier that sets the depth of every control that follows.

Two axes that determine the tier

We tier along two axes that, in practice, capture most of what matters. The first is **consequence of failure**: how badly a wrong or manipulated output can hurt a person, a community, or the institution's mission. The second is **reversibility**: whether a mistake can be caught and corrected before it causes durable harm, or whether it lands irrevocably. A system whose errors are grave but easily reversed can be governed more lightly than one whose errors are moderate but final.

THE RISK-TIERING MATRIX — CONSEQUENCE × REVERSIBILITY

	Easily reversible	Correctable with effort	Effectively irreversible
Severe consequence	Tier 2 — Elevated	Tier 3 — Critical	Tier 3 — Critical
Moderate consequence	Tier 1 — Standard	Tier 2 — Elevated	Tier 3 — Critical
Minor consequence	Tier 1 — Standard	Tier 1 — Standard	Tier 2 — Elevated

The tier is deliberately coarse. Three levels are enough to drive meaningfully different treatment while remaining simple enough that a program manager can assign a tier in an afternoon and defend it to a reviewer. Finer gradations tend to invite disputes that cost more than the precision they buy.

What each tier requires

1. **Tier 1 — Standard.**
   Productivity and convenience tools whose errors are minor and caught quickly: internal search, drafting assistants, meeting summaries. Lightweight review, clear usage guidance, and an easy way for staff to report problems. The goal is to avoid strangling low-risk value.

2 **Tier 2 — Elevated.**
Systems that inform consequential decisions but leave a human materially in control: triage prioritization, investigative lead generation, translation of official communications. Full documentation, acceptance testing, subgroup performance analysis, and a named owner.

3 **Tier 3 — Critical.**
Systems that determine or heavily steer decisions about rights, benefits, safety, liberty, or livelihood. The full weight of this framework applies: independent evaluation, a written safety case, individualized explanation and appeal, live monitoring, and executive sign-off.

DESIGN RULE

When a system sits on a boundary, tier it up, not down. The cost of over-governing a Tier 2 system is some wasted process. The cost of under-governing a Tier 3 system is borne by the public, who did not consent to the gamble.

Tiering is a living judgment

A tier is assigned at intake but revisited whenever the system's use expands, its population changes, or an incident reveals that the original assessment underestimated the stakes. A chatbot launched to answer directory questions (Tier 1) that begins advising the public on eligibility has quietly become a Tier 2 or Tier 3 system, and its governance must follow it up the ladder. We return to this "scope creep" problem in the monitoring section, because it is one of the most common ways safe deployments become unsafe without anyone deciding that they should.

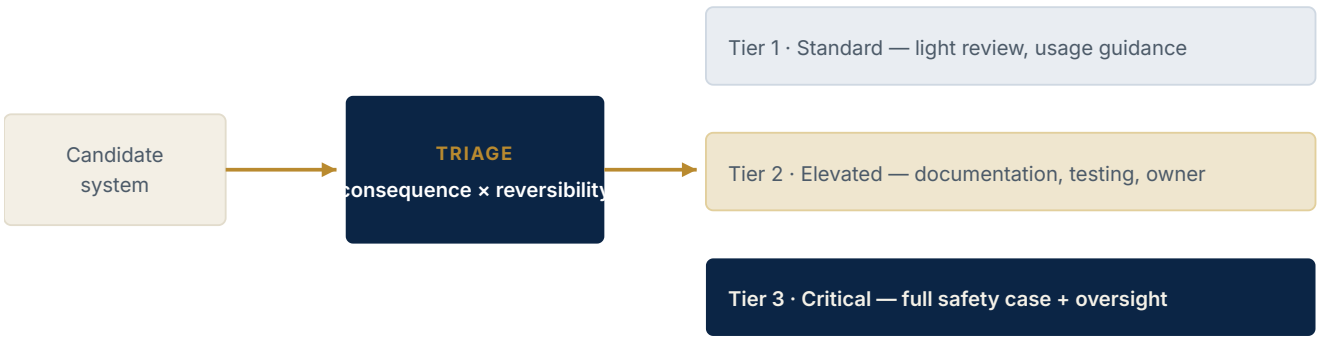

Figure 1. Every system passes through a single triage gate that sets the depth of all downstream controls. Tiering is cheap; it is what makes the expensive controls affordable where they are needed.

SECTION IV

Lifecycle Controls

Controls attach to stages, not to systems in the abstract. A safe deployment is one that has passed the right gate at each stage of its life and can prove it.

We organize institutional controls around six lifecycle stages. The stages are sequential but the lifecycle is a loop: monitoring feeds back into re-evaluation, and re-evaluation can send a system back for redesign or retirement. Each stage has an owner, an exit criterion, and an artifact that records the decision to proceed. The artifacts are the connective tissue of the whole program—they are what an auditor, an inspector general, or a court will ask to see.

THE SIX-STAGE LIFECYCLE AND ITS CONTROL GATES

Stage	Primary owner	Exit criterion (gate)	Artifact
1. Problem framing	Program owner	Documented need, non-AI alternatives considered, tier assigned	Intake & tiering memo
2. Data & design	Technical lead	Data provenance and representativeness assessed; design reviewed	Data & design dossier
3. Evaluation	Independent evaluator	Performance, robustness, and fairness meet tier thresholds	Evaluation report
4. Deployment approval	Accountable official	Written safety case accepted; oversight plan funded	Authorization to operate
5. Operation	Operations lead	Monitoring live; staff trained; incident path tested	Operations & monitoring log
6. Review & retirement	Accountable official	Periodic re-authorization or orderly decommissioning	Review & retirement record

Stage 1 – Problem framing

The cheapest safety intervention is asking whether the system should be built at all. Many proposed deployments dissolve under a direct question: what decision are we trying to improve, how is it made today, and would a simpler change—clearer forms, more staff, a rule change—serve better? When AI is the right answer, framing fixes the scope narrowly enough that the later stages are tractable. Scope defined loosely at stage one metastasizes into ungovernable ambition by stage five.

Stage 2 – Data and design

A model inherits the strengths and the pathologies of its data. At this stage the institution establishes where the data came from, whom it represents and whom it omits, how labels were produced, and what historical decisions—possibly biased ones—are baked into the training signal. For systems the institution builds or fine-tunes, design choices are documented here; for systems it buys, this stage becomes a demand for the vendor's equivalent evidence, discussed in Section V.

Stage 3 – Evaluation

Evaluation is where claims meet reality. The depth is set by tier: a Tier 1 system may need only a structured trial with staff, while a Tier 3 system requires evaluation by a party independent of the team that wants to deploy it. Whatever the depth, evaluation tests the five safety dimensions on data that resembles the real operating population—not a vendor's curated benchmark—and reports performance disaggregated across the groups the institution serves.

NON-NEGOTIABLE

For Tier 3 systems, the party that evaluates the system must be independent of the party that champions it. Self-graded homework is not evidence. Independence can be an internal unit with a separate reporting line, but it cannot be the deploying team.

Stage 4 – Deployment approval

Approval is a decision, made by a named human, recorded in writing. The central artifact is the *safety case*: a structured argument that the system is safe enough for its intended use, supported by the evaluation evidence and honest about residual risk and unknowns. The accountable official is approving the argument, not a checklist. A safety case that cannot state what it does not know is not finished.

Stage 5 – Operation

Operation is where most of a system's life is spent and where most governance programs go quiet. The controls here—monitoring, staff training, a tested path for reporting and escalating problems—are the subject of Section VII. The essential point at the lifecycle level is that the authorization to operate is conditional: it is granted on the promise that these controls are live, and it lapses if they are not.

Stage 6 – Review and retirement

Every authorization has an expiry. Re-authorization forces a periodic, deliberate answer to the question the institution answered at approval: is this still safe enough, given what we now know? And when a system is retired—superseded, defunded, or found wanting—retirement is itself a governed event, with attention to the records it produced, the decisions it made, and the people who may need to contest those decisions after the system is gone.

SECTION V

Procurement & Vendor Due Diligence

Most public-sector AI is bought, not built. This makes procurement the single most powerful point of leverage an institution has—and the one most often squandered. A well-designed procurement extracts, before any money changes hands, the very evidence that Sections III and IV demand. A poorly designed one accepts a demonstration and a slide deck, and inherits a decade of unaccountable operation.

Shift the burden of proof to the supplier

The governing principle of AI procurement is simple: the institution should not have to prove a system is unsafe; the vendor should have to prove it is safe enough. This inverts the usual dynamic, in which a polished vendor overwhelms a resource-constrained buyer. The instruments below are ways of writing that inversion into the acquisition itself, where the institution's leverage is greatest.

PROCUREMENT EVIDENCE DEMANDS, BY RISK TIER

Evidence requirement	Tier 1	Tier 2	Tier 3
System & model documentation	Summary	Detailed	Detailed + independent review
Training-data provenance statement	—	Required	Required + representativeness analysis
Disaggregated performance results	—	Required	Required on buyer's population
Acceptance testing on buyer data	Optional	Required	Required, independently observed
Security & privacy attestation	Required	Required	Required + third-party audit
Incident cooperation & update terms	Basic	Required	Required with SLAs
Exit & portability provisions	Basic	Required	Required with data return

Acceptance testing: trust, then verify

A vendor's benchmark numbers describe the vendor's benchmark. What matters is performance on the institution's own data, tested before acceptance. Acceptance testing need not be elaborate to be powerful: a held-out sample of real cases, scored against the vendor's claims, will expose the gap between demonstration and deployment more reliably than any amount of documentation. For Tier 3 systems this test is observed by a party independent of the vendor and the deploying team.

> *The most expensive words in public-sector AI are "the vendor said it works."*
>
> — Field maxim, adopted here without apology

Contract terms are safety controls

Several risks can only be managed in the contract, because they concern what happens after deployment, when the institution's leverage has evaporated. Four terms deserve non-negotiable status for Tier 2 and Tier 3 systems.

Change notification

The vendor must notify the institution before materially changing the model, and the institution must have the right to re-test before the change takes effect on consequential decisions. A silent model update is an ungoverned redeployment.

Incident cooperation

When something goes wrong, the vendor must cooperate with investigation on a defined timeline, including providing logs and technical access the institution cannot generate itself. This must be written down while the institution still has bargaining power.

Auditability

The institution, or a party acting on its behalf, must have the contractual right to inspect the system to the depth the tier requires. "Trade secret" is a legitimate interest but not a universal shield against a public body's duty to understand what it has deployed.

Exit and portability

The institution must be able to leave—to recover its data, migrate to an alternative, and decommission cleanly—without punitive cost. Lock-in is not merely a commercial hazard; it is a safety hazard, because it raises the price of pausing a system that should be paused.

PROCUREMENT RULE

Write the off-ramp before you drive onto the highway. Every term that governs failure, update, audit, and exit is far cheaper to secure in the contract than to litigate in the incident.

SECTION VI

Accountability & Human Oversight

Oversight that belongs to everyone belongs to no one. Every consequential system needs a specific human who is answerable for it and empowered to stop it.

Institutions are fond of "human-in-the-loop" as a reassurance. But a human in the loop who cannot understand the system, lacks the time to review its outputs, or fears the consequences of overriding it is not oversight—she is a liability shield. Meaningful human oversight has preconditions, and this section makes them explicit.

The accountable official

For every Tier 2 and Tier 3 system, the institution names an **accountable official**: a specific person, senior enough to have authority, who owns the deployment's safety. This official approves the authorization to operate, ensures the monitoring is funded and functioning, and—critically—holds the power to suspend the system without needing anyone else's permission. If no one can pause a system on their own authority, the institution does not control it.

A **Authority to halt.**
The accountable official can suspend the system immediately on credible evidence of harm, and is protected for doing so in good faith. Halting must never require assembling a committee.

B **Visibility.**
The official receives monitoring reports at a defined cadence and is notified of incidents without delay. Accountability without information is a setup for blame.

C **Standing.**
The role is documented, resourced, and durable across staff turnover. When the person leaves, the role transfers; it does not evaporate.

Preconditions for meaningful operator oversight

Below the accountable official sit the operators who work with the system daily. For their oversight to be real rather than nominal, four conditions must hold. Operators must *understand* the system well enough to know when to doubt it. They must have the *time* to exercise judgment rather than rubber-stamp a queue. They must face incentives that do not punish careful override. And they must have a

usable path to escalate concerns that does not run only through the people who championed the deployment.

THE AUTOMATION-BIAS TRAP

People defer to machines, especially confident ones, especially when overriding is extra work. An oversight design that ignores this predictable tendency will produce the appearance of human control and the reality of automated decision-making. Countermeasures—selective review, friction on high-stakes overrides, tracking of override rates—must be designed in, not hoped for.

Explanation and appeal for the affected person

Accountability runs outward to the public as well as upward to leadership. A person subject to a consequential automated decision has a legitimate claim to know that a system was involved, to receive an explanation intelligible to a non-specialist, and to contest the outcome through a route that reaches a human with the authority to change it. For Tier 3 systems these are requirements, not courtesies. A determination that cannot be explained cannot be justly appealed, and a determination that cannot be appealed should not be automated.

Figure 2. A minimal accountability structure. The through-line from the affected person's appeal up to an official with the authority to halt is what distinguishes oversight from decoration.

SECTION VII

Monitoring, Incidents & Continuous Assurance

A system certified safe at deployment is a claim with an expiration date. The population shifts, the vendor updates the model, an upstream data feed changes format, and the careful calibration established at acceptance quietly decays. Continuous assurance is the discipline of noticing this decay before the public does. It is also the stage institutions most often underfund, because monitoring produces no launch to celebrate—only the steady, unglamorous work of watching.

What to monitor

Effective monitoring watches four things at once. It watches **performance**—is the system still as accurate as it was, overall and for each group we serve? It watches **drift**—are the inputs the system now sees still like the ones it was evaluated on? It watches **usage**—is the system being used for what it was authorized to do, or has its scope crept? And it watches **human interaction**—are operators overriding it, and when they do, who is right?

A MINIMUM MONITORING REGIME BY TIER

Signal	Tier 1	Tier 2	Tier 3
Aggregate performance	Periodic	Scheduled review	Continuous with alerts
Subgroup performance	—	Scheduled review	Continuous with alerts
Input drift	—	Periodic	Continuous
Override / appeal rates	Feedback channel	Tracked	Tracked with thresholds
Scope / usage audit	Annual	Semi-annual	Quarterly

Incident response

When monitoring—or a member of the public, or a journalist—surfaces a problem, the institution needs a rehearsed response, not an improvised one. A workable AI incident process mirrors mature security practice: **detect** the issue and log it; **triage** its severity and reach; **contain** it, up to and including pausing the system; **remediate** the root cause; and **learn**, feeding the lesson back into design, evaluation, and procurement. The authority to contain—to pause—must sit close to the monitoring, not three approval layers away, because harm compounds while permission is sought.

THE PAUSE MUST BE REAL

An institution that cannot actually turn a system off—because it is load-bearing, because the fallback was never built, because the vendor controls the switch—has no meaningful containment. The ability to revert to a safe manual fallback is a control that must be established at deployment and tested, not assumed at the moment of crisis.

Scope creep: the quiet path from safe to unsafe

The most common way a well-governed system becomes dangerous is not a dramatic failure but a gradual expansion of use. A tool approved to draft internal notes starts drafting public correspondence; a model that flags cases for review starts effectively deciding them because reviewers defer. Each step seems small; none triggers a re-assessment; and the system slides up the risk tiers without ever passing back through the gate that would have demanded new evidence. Usage auditing exists precisely to catch this, and it is why "how is this actually being used?" is a standing monitoring question rather than a one-time design assumption.

SECTION VIII

Implementation Roadmap & Maturity Model

No institution adopts this framework whole on a Monday. The realistic path is incremental, and a maturity model lets an institution locate itself honestly and choose the next achievable step.

A five-level maturity ladder

AI GOVERNANCE MATURITY — FROM AD-HOC TO INSTITUTIONALIZED

Level	Character	What it looks like in practice
1 · Ad-hoc	Unmanaged	AI enters through individual projects; no inventory, no shared rules, no one accountable across systems.
2 · Aware	Reactive	The institution knows AI is a distinct risk; a policy exists on paper; application is inconsistent.
3 · Defined	Proportionate	Tiering, lifecycle gates, and procurement standards are documented and applied to new systems.
4 · Managed	Assured	Monitoring and incident response operate for consequential systems; accountability is real and resourced.
5 · Institutionalized	Self-improving	Governance is routine and funded; lessons flow back into standards; the program improves itself.

A pragmatic first year

1. **Inventory and tier (months 1–3).**
You cannot govern what you cannot see. Build a register of AI systems already in use and assign each a provisional tier. This alone typically surfaces surprises.

2. **Adopt the gates for new systems (months 2–6).**
Apply lifecycle gates and procurement standards to everything new, where the cost of doing so is lowest and the leverage highest. Do not try to retrofit everything at once.

3 **Name accountable officials for Tier 3 (months 3–6).**
Assign ownership for the highest-consequence systems first. Ownership is the control that makes every other control enforceable.

4 **Stand up monitoring for the riskiest systems (months 6–12).**
Start where failure is most consequential. A modest monitoring capability on Tier 3 systems beats an elaborate plan that never launches.

What good looks like

An institution operating this framework well is not one that has eliminated risk—that is not on offer. It is one that can answer, for any AI system it operates, a short list of questions without scrambling: *What tier is it, and why? Who is accountable? What evidence supports its use? How would we know it it started causing harm, and how fast could we stop it?* An institution that can answer these has closed the governance gap this report opened with. The instruments here exist to make those answers routine.

CLOSING NOTE

The aim of AI safety governance is not to slow institutions down. It is to let them move quickly on the many low-stakes uses of AI precisely because they have the discipline to move carefully on the few that can hurt people. Proportionality is not the enemy of ambition; it is its precondition.

APPENDIX

Notes & Further Reading

This framework synthesizes established practice in risk management, safety engineering, and public-sector technology governance. The sources below are entry points into the literature that informs it; they are indicative rather than exhaustive.

1. National Institute of Standards and Technology. *AI Risk Management Framework (AI RMF 1.0).* A voluntary framework organizing AI risk into govern, map, measure, and manage functions.
2. Organisation for Economic Co-operation and Development. *OECD AI Principles.* Intergovernmental principles for trustworthy AI, including accountability and transparency.
3. Leveson, N. *Engineering a Safer World: Systems Thinking Applied to Safety.* On safety as a control problem across a sociotechnical system rather than a component property.
4. Mitchell, M., et al. *Model Cards for Model Reporting.* On structured documentation of model performance across conditions and groups.
5. Gebru, T., et al. *Datasheets for Datasets.* On documenting data provenance, composition, and intended use.
6. Raji, I. D., et al. *Closing the AI Accountability Gap: internal algorithmic auditing.* On institutional audit processes across the development lifecycle.
7. Selbst, A. D., et al. *Fairness and Abstraction in Sociotechnical Systems.* On why fairness cannot be resolved at the level of the model alone.
8. Parasuraman, R., & Riley, V. *Humans and Automation: Use, Misuse, Disuse, Abuse.* On automation bias and the conditions for meaningful human oversight.
9. U.S. Government Accountability Office. *Artificial Intelligence: An Accountability Framework.* On governance, data, performance, and monitoring for federal use.
10. Amodei, D., et al. *Concrete Problems in AI Safety.* On robustness, distributional shift, and safe operation as engineering problems.