FREE PRACTITIONER PLAYBOOK

AI Governance Metrics Playbook

v1.0

28 metrics across 7 layers for governing AI systems that have to perform under scrutiny — with calculation guidance, benchmarks, and gap signals for each one.

Paulo Cavallo, PhD

Senior Credit Risk Model Developer
8+ years SR 11-7 regulatory compliance & model risk management
Building AI governance frameworks for regulated industries

Who this is for: AI leads, model risk managers, compliance professionals, ML engineers, and governance practitioners who need to move from policy documents to measurable, auditable controls.

INTRODUCTION

Why Metrics Beat Policy

Most AI governance programs are built on policy. They document what the AI *should* do, who *should* review it, and what *should* happen when things go wrong. None of that is a control. A policy the AI can ignore is not governance. A checklist nobody verifies is not oversight.

1

The Design Flaw Hidden in Plain Sight

The standard governance failure mode is not malice or negligence. It is a program optimized for documentation, not for detection. Organizations build policies, complete checklists, and publish AI principles — and call that governance. What they have built is governance *theater*: structured enough to look like oversight, but too shallow to catch the failures that matter. A Gartner survey of 360 organizations in Q2 2025 found that those deploying AI governance platforms were 3.4× more likely to achieve high effectiveness in AI governance than those relying on manual processes alone. The difference was not intent. It was instrumentation.

2

The Seven-Layer Model

This playbook organizes AI governance into seven distinct layers, each addressing a category of risk that policy alone cannot contain. The layers are not sequential checkboxes — they operate simultaneously in any production AI system. What changes between layers is the type of failure they are designed to detect, and the kind of evidence they generate when something goes wrong. Each layer contains four metrics. Each metric includes: what it measures, how to calculate it, what good looks like, and what a gap signal tells you. The 28 metrics together form a self-assessment scorecard you can run on any AI system in your portfolio.

3

The Regulatory Foundation

These metrics are grounded in four primary sources: SR 11-7 (Federal Reserve/OCC Model Risk Management Guidance, 2011), NIST AI RMF 1.0 (NIST AI 100-1, 2023), NIST Generative AI Profile (NIST AI 600-1, 2024), and ISO/IEC 42001:2023 (AI Management Systems). Together, these frameworks cover traditional quantitative models, general AI/ML systems, generative AI, and enterprise AI management. Where metrics address different requirements across frameworks, those cross-references are noted. Gartner projects that by 2028, governance technologies will decrease regulatory compliance costs by 20% — but only for organizations that can demonstrate continuous, measurable oversight. That demonstration requires metrics. This playbook is the starting point.

How to use this playbook. Run the scorecard on page 14 once for your overall AI program, then once per high-risk AI system. A program that scores well at the portfolio level but has individual systems with critical gaps is not governed — it is unevenly protected. Per-system scoring reveals where your real exposure is. Use the layer sections to understand how each metric works before scoring.

"28 metrics. Most organizations can fill in 3.
That's not a criticism of the people. It's a design flaw in how governance programs are built."

AI UNDER AUDIT — PAULO CAVALLO, PHD

The Seven Layers at a Glance

#	Layer	Core Question	Primary Framework Alignment
1	Governance Structure	Do we own every AI system we run?	SR 11-7 · ISO 42001 · NIST MAP
2	Model & Data Integrity	Can we reproduce and trace every outcome?	SR 11-7 · NIST MEASURE
3	Operational Risk	Do we know when something breaks before it causes harm?	SR 11-7 · ISO 42001 · NIST MANAGE
4	Fairness & Bias	Are outcomes equitable across the populations we serve?	FHFA AB 2022-02 · NIST AI 100-1 · ECOA
5	Transparency & Explainability	Can any output be explained to the person it affects?	NIST AI 100-1 · ISO 42001 · EU AI Act
6	Security & Adversarial Robustness	Can the system be manipulated, extracted, or poisoned?	NIST AI 600-1 · NIST CSF · ISO 42001
7	Human Oversight & Accountability	When the AI is wrong, who is responsible and how fast do we catch it?	SR 11-7 · NIST GOVERN · EU AI Act

LAYER 01 OF 07

Governance Structure

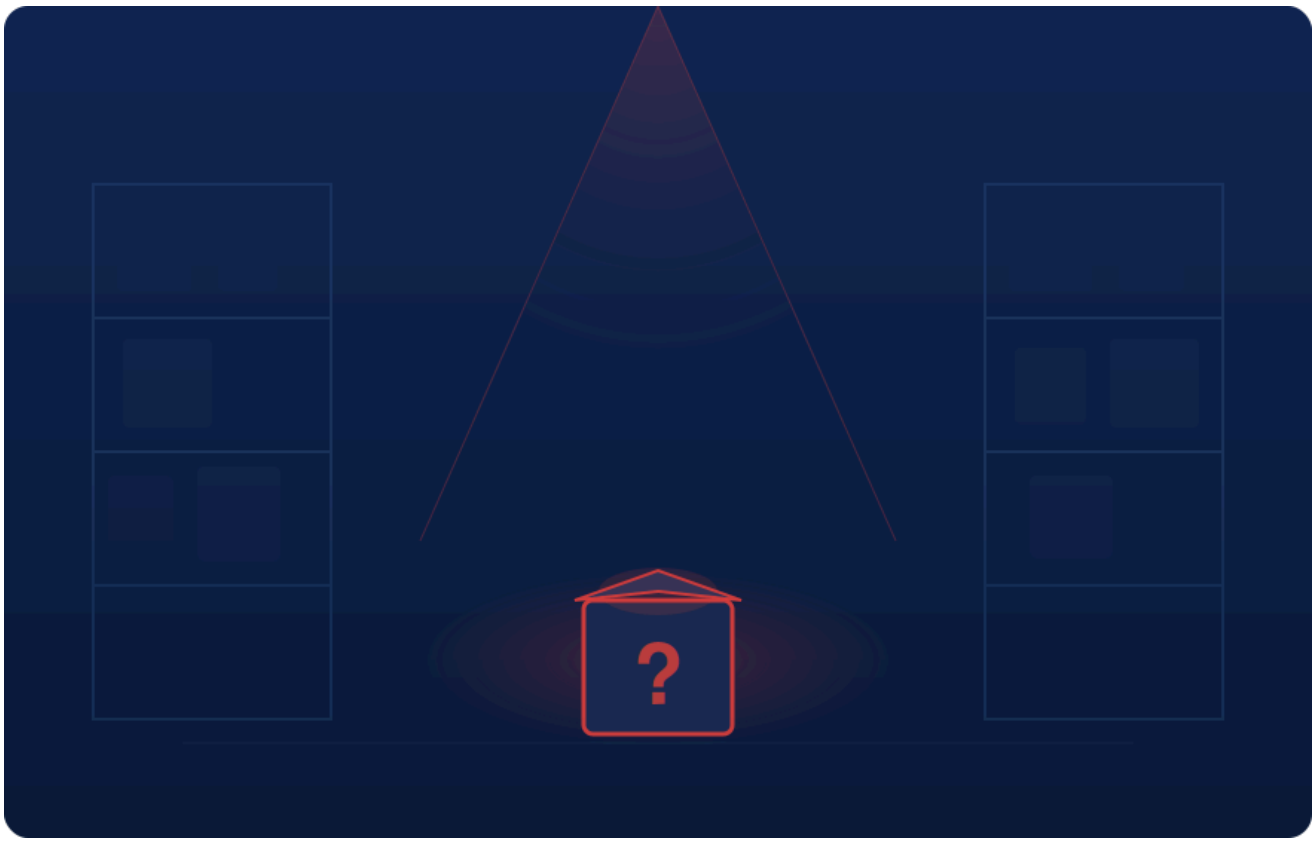
1

CORE QUESTION

Do we own every AI system we run?

Governance that cannot name what it governs is not governance. Before any metric in layers 2–7 can be applied, an organization needs a complete, current inventory of its AI systems, with clear ownership, defined scope, and an independent challenge function for each. SR 11-7 has required this for quantitative models since 2011. ISO/IEC 42001:2023 extends the same requirement to all AI systems, including those embedded in vendor software, third-party tools, and agentic pipelines. Shadow AI — systems in production without governance coverage — is the most common and most dangerous gap in this layer.

The Shadow AI Problem. The average large enterprise has between 40 and 80 AI-enabled tools in active use, of which fewer than half appear in formal inventories. Vendor-embedded AI (copilot features, intelligent routing, automated scoring) frequently escapes governance entirely because procurement treats it as a software feature, not a model. FHFA AB 2022-02 explicitly names this risk, requiring the same governance rigor for AI embedded in vendor software as for internally developed models. If it uses data to make decisions that affect people, it belongs in your inventory.



METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
AI Inventory Coverage Rate	Percentage of AI systems in production with a documented owner, defined scope, and explicit use-case boundaries.	(AI systems with complete governance record ÷ total known AI systems in production) × 100	**100%. Anything less means ungoverned systems are making decisions.**	*Below 80% signals governance theater. Unknown systems cannot be audited.*
Policy-to-Control Ratio	Percentage of AI governance policies that have a corresponding enforcement mechanism — not a document, an actual system control.	(Policies with enforcement mechanism ÷ total governance policies) × 100	**≥70%. Mature programs systematically convert policies to automated controls.**	*Below 50% means governance depends on human compliance. High audit risk.*
Independent Validation Coverage	Percentage of material AI systems with documented validation performed by a team independent of development. SR 11-7's second line of defense requirement, applied to all AI.	(AI systems with completed independent validation ÷ material AI systems) × 100	**100% for material/high-risk systems. Independent means independent — not a peer review by the same team.**	*Any gap in high-risk systems is a direct SR 11-7 or ISO 42001 finding.*
Regulatory Mapping Completeness	Percentage of AI systems with a documented map of which regulatory frameworks apply and which requirements	(Systems with completed regulatory map ÷ total AI systems) × 100	**100% for material systems. Mapping should name applicable frameworks, not just reference them generically.**	*Unmapped systems cannot demonstrate compliance. A major gap when regulators ask.*

METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
	have been addressed.			

Implementation note. Start your inventory with a broad definition of "AI system": any system that uses statistical methods, learned patterns, or automated decision logic to produce outputs that affect people or business processes. Include vendor tools. Include copilot features. Include scoring models embedded in loan origination platforms, CRM tools, and HR software. You cannot govern what you have not named. The NIST AI RMF MAP function formalizes this as a prerequisite step — it comes before Measure and Manage for a reason.

LAYER 02 OF 07

Model & Data Integrity

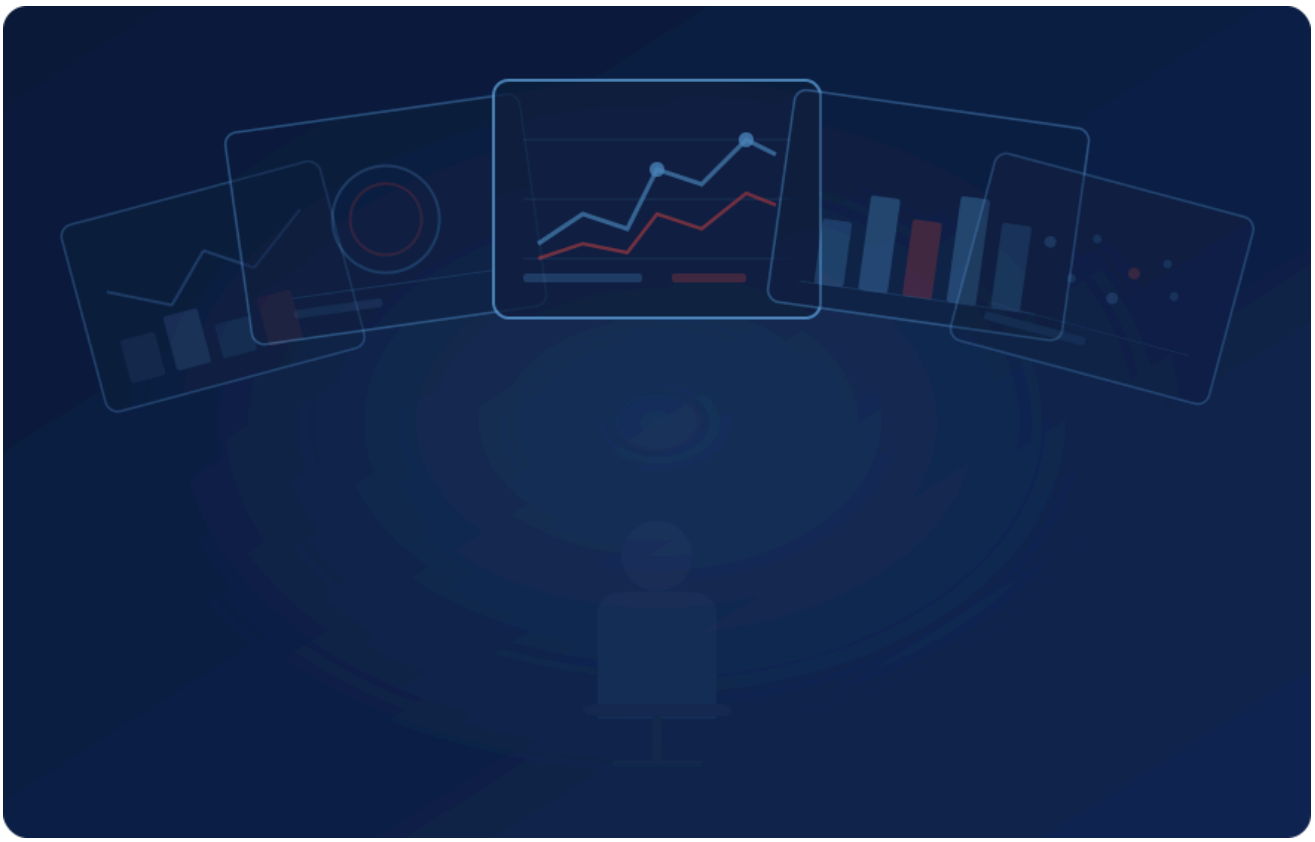
2

CORE QUESTION

Can we reproduce and trace every outcome?

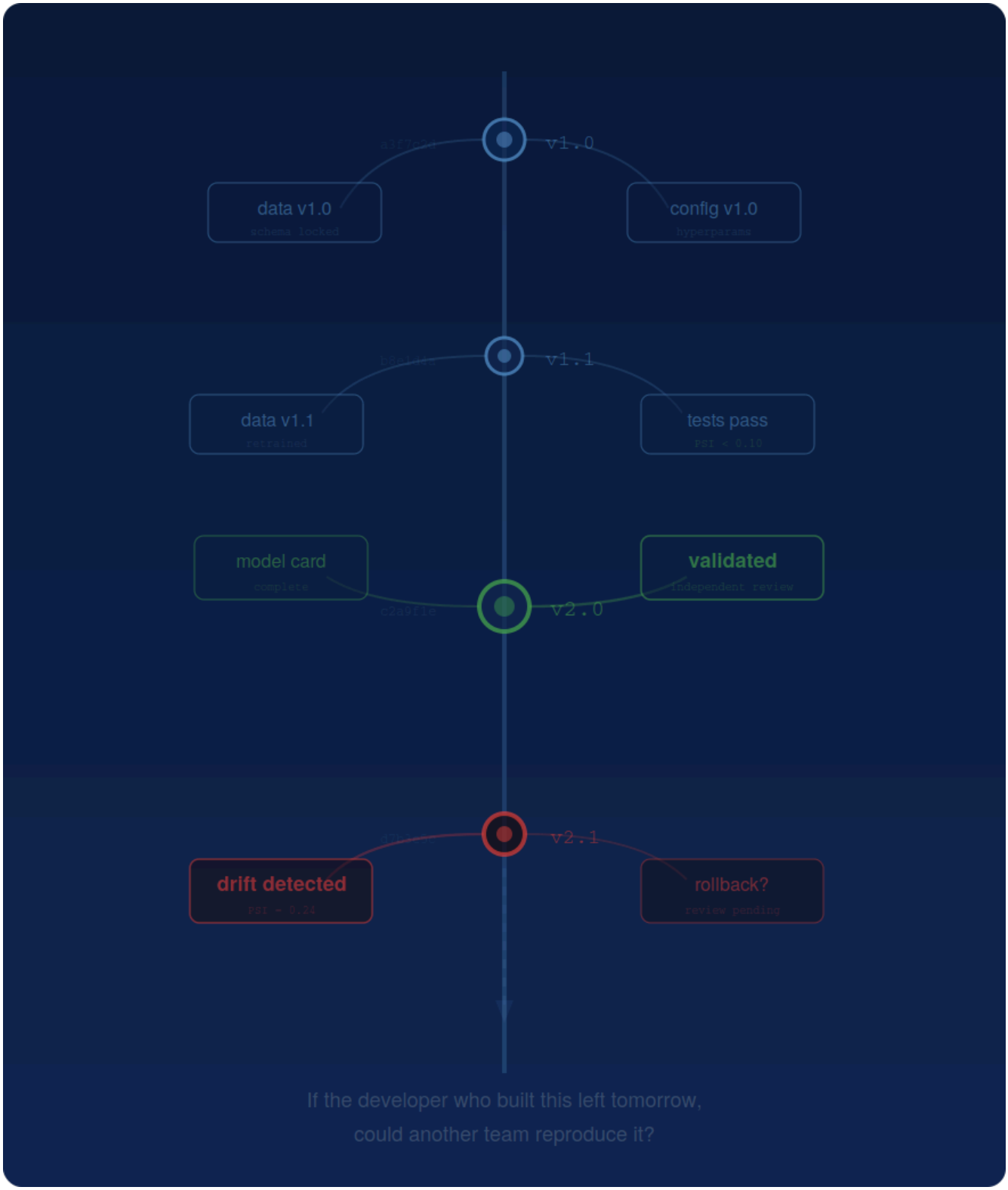
A model you cannot reproduce is a model you cannot validate. SR 11-7's core requirement for model risk management is that development, data, and assumptions be documented with enough specificity to allow independent replication. This is not archival work — it is the mechanism that makes validation possible. If a model cannot be reconstructed from documented inputs, any validation performed is, in practice, a review of the developer's claims rather than an independent assessment of the model itself. ISO/IEC 42001 extends this requirement to the full AI lifecycle, including third-party and foundation models used as components.

The Reproducibility Gap in GenAI. Traditional quantitative models can be reproduced exactly: same data, same algorithm, same hyperparameters, same output. GenAI systems introduce inherent non-determinism (temperature, sampling, context window effects). NIST AI 600-1 addresses this directly — reproducibility for GenAI systems does not mean identical token-for-token output. It means documented system configuration, prompt versioning, and output distribution testing that demonstrate consistent behavior within defined tolerances. If you cannot define "consistent enough," you have not solved the reproducibility problem.



METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
Data Lineage Completeness	Percentage of production models with documented provenance for all training and validation data, including transformations, filters, and known biases.	(Models with complete data lineage documentation ÷ total production models) × 100	**100% for material models. Documentation should include source, vintage, transformations applied, and exclusions made.**	*Undocumented data is unauditable data. Regulators consider this a fundamental gap.*
Model Reproducibility Rate	Percentage of models that can be retrained from documented source data and produce output within an acceptable performance tolerance of the production version.	(Models reproducible within ±X% performance threshold ÷ material production models) × 100. Define X by model type.	**100% for deterministic models. For GenAI: documented system config, version-controlled prompts, and defined output tolerance.**	*Irreproducible models cannot be independently validated. They fail SR 11-7 and ISO 42001 requirements.*
Version Control Coverage	Percentage of production models, training datasets, and model configurations under formal version control with change history.	(Models + datasets + configs under version control ÷ total models + datasets + configs) × 100	**100%. Version control is not optional — it is the mechanism for change traceability.**	*Unversioned components cannot support rollback, audit, or incident investigation.*
Validation Independence Score	Percentage of validations where the reviewing team had no organizational overlap with the development team at the time of validation. Measures genuine independence, not nominal independence.	(Validations with documented organizational independence ÷ total completed validations) × 100	**100% for material/high-risk systems. Peer review within the same team does not count as independent validation under SR 11-7.**	*Developer-validated models are a direct SR 11-7 finding. Independence is structural, not procedural.*

Implementation note. Treat your model registry as a governance artifact, not an IT catalog. Each entry should resolve the question: "If the developer who built this left tomorrow, could another team reproduce, validate, and take ownership of this model?" If the answer is no, documentation is incomplete regardless of what the registry says it contains.



LAYER 03 OF 07

Operational Risk

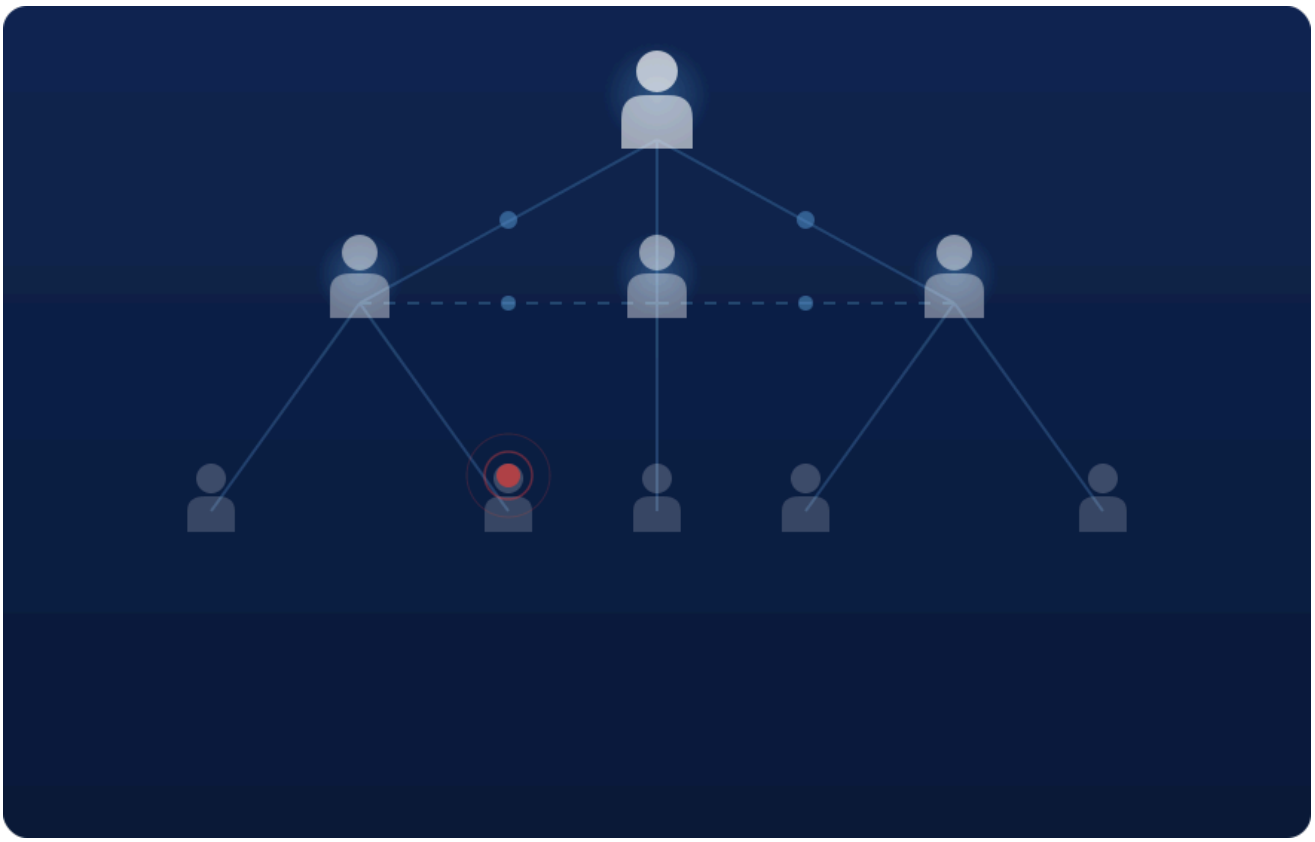
3

CORE QUESTION

Do we know when something breaks before it causes harm?

Production AI systems degrade. Input distributions shift. Upstream data changes without notice. User behavior evolves. Models that performed well at deployment perform poorly six months later, quietly and without announcement. SR 11-7 requires ongoing monitoring "commensurate with the model's risk and complexity." ISO/IEC 42001 requires continuous performance tracking as a fundamental operating requirement, not an annual review. The question this layer answers is whether the degradation is detected by your monitoring system — or by a customer complaint, a regulatory finding, or a downstream business loss.

The Latency Problem. Most monitoring failures are not failures of detection — they are failures of speed. Organizations often have monitoring in place that detects drift or performance degradation but routes alerts through review processes that take weeks to resolve. By the time a remediation is in production, the model may have made thousands of affected decisions. Drift detection latency and incident response time are not the same metric. Both matter. The NIST AI RMF MANAGE function explicitly requires defined response times for AI risk incidents, not just detection capability.



METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
Monitoring Coverage Rate	Percentage of production AI systems with automated, continuous performance monitoring including drift detection and output anomaly alerting.	(AI systems with active automated monitoring ÷ total production AI systems) × 100	**100% for material systems. "Active" means automated alerting, not a scheduled monthly report.**	*Unmonitored production systems are operating under the assumption that they continue to work. That assumption is not evidence.*
Drift Detection Latency	Mean time (in days) from the onset of measurable input or output distribution drift to the generation of a monitoring alert. Measures how quickly the system detects that something has changed.	Average across monitored systems of: (date alert triggered – date drift began). Drift onset can be estimated retrospectively using historical monitoring data.	**≤7 days for high-risk systems. ≤30 days for lower-risk systems. Defined by risk tier in governance policy.**	*Latency >30 days for any material system means the model may be degraded for a full reporting cycle before anyone knows.*
Incident Response Time	Mean time (in days) from documented alert to deployed remediation or documented risk acceptance for AI performance incidents.	Average across incidents in a rolling 12-month period of: (date remediation deployed – date alert generated). Include risk-accepted incidents with their formal	**≤14 days for critical incidents. Risk acceptance decisions documented within 5 business days if remediation will take longer.**	*High alert-to-response time with no documented risk acceptance signals alerts are being generated but not acted on — a governance failure more serious than not monitoring at all.*

METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
		acceptance date.		
Threshold Exceedance Rate	Percentage of monitoring checks that triggered a threshold breach in the rolling 90-day period. Used as a signal of operational health — too high indicates instability, too low may indicate thresholds are set too permissively.	(Monitoring checks triggering threshold breach ÷ total monitoring checks run) × 100 over 90-day rolling window.	**1–5% for a stable, well-calibrated system. Calibrate thresholds during initial deployment; recalibrate annually.**	*>15% suggests systemic instability or poor threshold calibration. 0% over a long period may indicate thresholds are set too loosely to detect real degradation.*

Implementation note. Define your monitoring thresholds before deployment, not after the first incident. Thresholds set post-incident are optimized to not have triggered on the last problem — which tells you nothing about the next one. Document the rationale for every threshold. "We set it at 5% because that seemed reasonable" will not survive audit scrutiny. "We set it at 5% because historical volatility was 2% and we used a 2.5σ tolerance" will.

LAYER 04 OF 07

Fairness & Bias

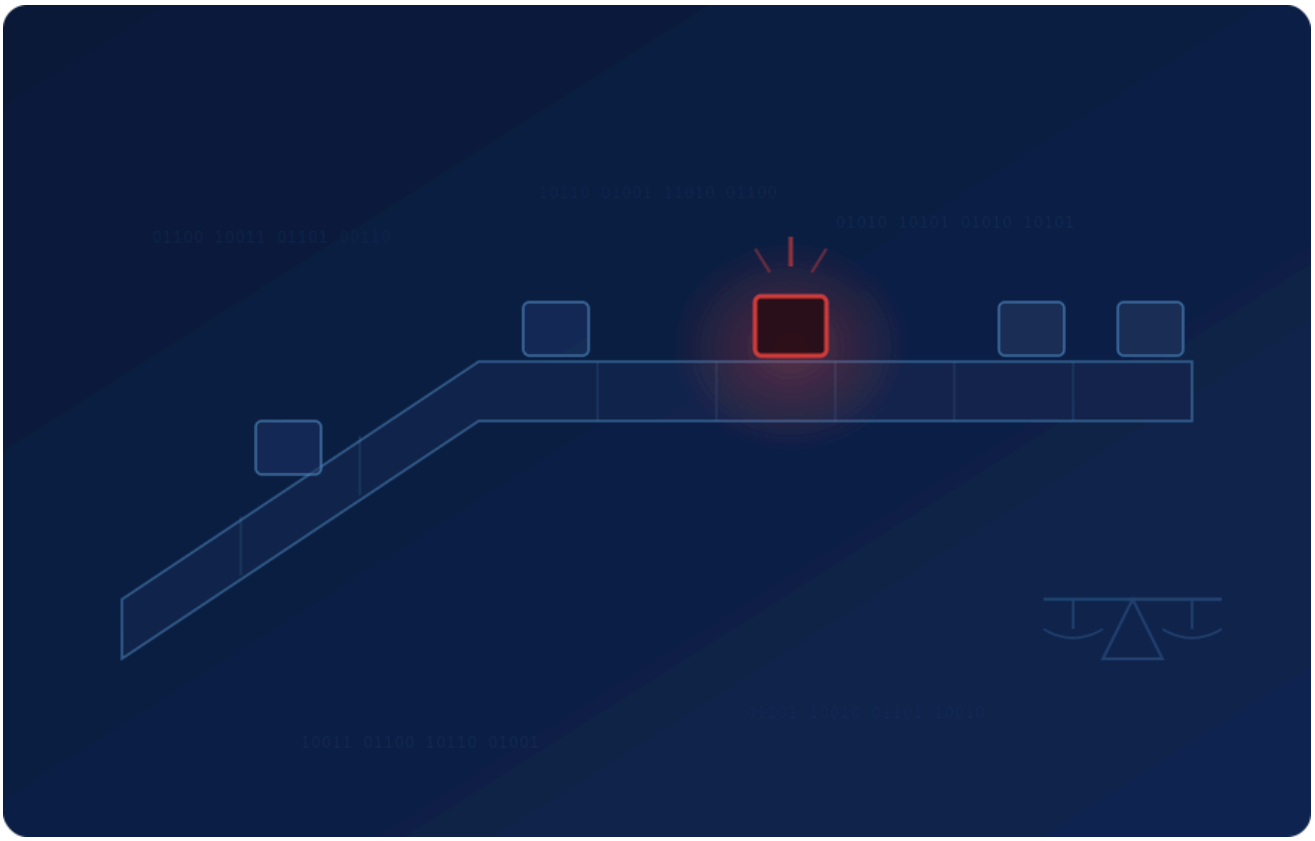
4

CORE QUESTION

Are outcomes equitable across the populations we serve?

Bias in AI systems is not always the product of intent — it is often the product of data that reflects historical inequities, proxy variables that encode protected characteristics, and optimization objectives that maximize aggregate performance at the cost of distributional equity. NIST AI 100-1 identifies fairness as a core characteristic of trustworthy AI, requiring active assessment and mitigation across all phases of the AI lifecycle. FHFA AB 2022-02 goes further, requiring fair lending testing across all lifecycle stages — not just at deployment. The Equal Credit Opportunity Act (ECOA) and Fair Housing Act create legal obligations that apply regardless of whether the AI system's outputs are intentionally discriminatory.

The Proxy Problem. Prohibited variables do not need to be in the model for the model to discriminate. Zip code, first name, and shopping patterns can all serve as proxies for race or national origin. A model that was never shown a protected characteristic can still produce disparate outcomes if trained on historically biased data. This is why fairness testing requires outcome analysis across protected groups, not just an audit of input variables. Removing the variable does not remove the signal if correlated proxies remain.



METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
Protected Segment Coverage	Percentage of production models with documented bias evaluation across all legally relevant protected segments (race, sex, national origin, age, disability status, and others applicable to the use case).	(Models with bias evaluation covering all applicable protected classes ÷ total production models with decision impact) × 100	**100% for any model with credit, employment, housing, or consumer lending impact. Partial coverage is not compliant coverage.**	*Any material model without complete protected class coverage carries direct legal risk under ECOA, FHA, or FCRA.*
Disparate Impact Ratio	Ratio of the favorable outcome rate for the least-favored group to the most-favored group, for each protected class. The "4/5ths rule" (≥0.80) is the EEOC threshold used as a starting benchmark.	Favorable outcome rate for least-favored group ÷ favorable outcome rate for most-favored group. Calculate per protected class, per model, in production on a rolling 90-day basis.	**≥0.80 (80% ratio) per protected class, per model. This is a floor, not a ceiling. Some regulatory contexts require tighter tolerances.**	*Below 0.80 for any protected class requires documented investigation, root cause analysis, and remediation or formal risk acceptance with senior sign-off.*
Fairness Assessment Frequency	Time (in days) since the last completed fairness evaluation for each production model with decision	Current date − date of last completed fairness evaluation. Track per model. Flag systems exceeding	**High-risk models: ≤90 days. Other models: ≤180 days. Frequency should increase after significant data or population shifts.**	*>365 days without a fairness evaluation on any model with protected class impact is a governance failure with potential legal consequences.*

METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
	impact. Fairness is not a one-time assessment — it is a continuous monitoring obligation.	frequency threshold.		
Bias Remediation Rate	Percentage of identified bias findings (disparate impact ratio below threshold, or flagged bias finding from any source) with a documented corrective action: either a deployed remediation or a formally accepted residual risk with named approver.	(Bias findings with documented corrective action or formal risk acceptance ÷ total identified bias findings in period) × 100	**100%. Every identified finding should have a decision. Open findings with no documented action are ungoverned risk.**	*Open findings with no action after 30 days is the characteristic signature of a governance program that identifies problems but cannot resolve them.*

Implementation note. The disparate impact ratio is a detection metric, not a compliance safe harbor. A ratio above 0.80 does not mean the model is free of discriminatory impact — it means you have not yet found evidence of a threshold violation. Rigorous fairness governance includes intersectional analysis (outcomes for members of multiple protected groups simultaneously), sensitivity analysis on proxy variables, and documented review of training data for historical bias. Start with the ratio; do not stop there.

LAYER 05 OF 07

Transparency & Explainability

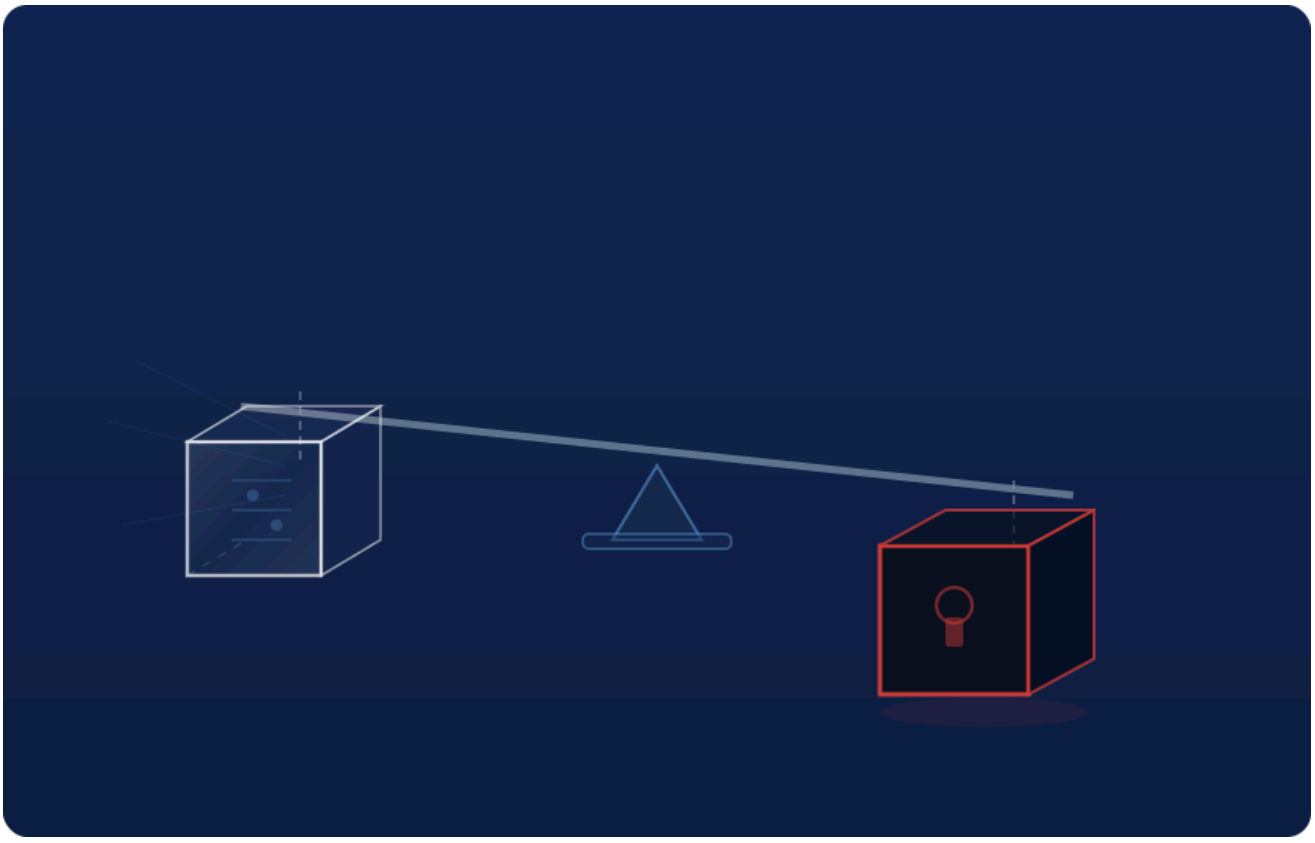
5

CORE QUESTION

Can any output be explained to the person it affects?

NIST AI 100-1 draws a careful distinction between *explainability* (the ability to describe why a model produced a specific output) and *interpretability* (the ability to understand how the model works in general). Regulators and the people affected by AI decisions need both — but they need them in different forms. An examiner needs interpretability: a mechanistic account of the model's decision logic. A loan applicant denied credit needs explainability: a clear, specific, human-readable reason for their outcome. The EU AI Act's right to explanation requirement, FCRA's adverse action notice requirement, and SR 11-7's documentation obligations all stem from the same principle: consequential AI decisions cannot be opaque to those they affect.

The "Black Box Risk" Named by FHFA. FHFA Advisory Bulletin 2022-02 explicitly calls out "black box risk" as a primary concern for AI/ML in housing finance. A model that cannot be explained to a regulator, an auditor, or an affected borrower is not just a technical problem — it is an accountability gap with direct regulatory consequences. Note that "explainability" in this context does not require that the model be simple. It requires that tools, documentation, and processes exist to generate explanations on demand. A complex model with robust explainability infrastructure is more governable than a simple model with no explanation capability.



METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
Model Card Completion Rate	Percentage of production models with a current, complete model card documenting: intended use, out-of-scope uses, training data, evaluation metrics, known limitations, and recommended human oversight level.	(Models with complete, dated model card ÷ total production models) × 100. "Current" = updated within the last model version cycle.	**100%. Model cards are the foundational transparency artifact. They are the first thing an auditor or regulator will ask for.**	*Outdated or incomplete model cards fail both NIST AI 100-1 transparency requirements and ISO 42001 documentation requirements.*
Explanation Coverage Rate	Percentage of individual model decisions for which an explanation can be generated on demand, identifying the primary factors that drove the output. Critical for adverse action notices and regulatory inquiry responses.	(Model decisions for which factor-level explanation is available ÷ total model decisions in period) × 100	**100% for models making consequential individual decisions (credit, employment, healthcare triage). Use SHAP, LIME, attention weights, or rule extraction as appropriate to model type.**	*Any credit, insurance, or employment decision without available explanation creates adverse action notice liability (FCRA, ECOA) and EU AI Act right-to-explanation obligations.*
Audit Package Readiness	Percentage of material models for which a complete audit package — purpose, data lineage, methodology, limitations, monitoring logs, and incident history — can be produced	(Material models with audit-ready documentation retrievable in <24hr ÷ total material models) × 100. Test this quarterly with a simulated audit request.	**100%. The 24-hour standard reflects real regulatory examination timelines. If documentation requires days to compile, governance is not operational — it is aspirational.**	*Inability to produce documentation quickly is not just an operational problem — it signals that documentation is not maintained continuously, which means the documentation may be inaccurate.*

METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
	within 24 hours of request.			
Explanation Comprehensibility Rate	Percentage of explanations provided to non-technical stakeholders (customers, business owners, compliance staff) that are rated as understandable by the recipient. Measures whether explanations serve the person who needs them, not just the person who generates them.	Track via structured feedback from adverse action notice recipients, business stakeholder reviews, and compliance team surveys. Define comprehensibility criteria before measuring.	**≥85% rated as clear and sufficient by non-technical recipients. Recalibrate explanation format if below threshold.**	*Technical explanations that non-technical recipients cannot use are not explanations — they are documentation of the model's complexity without resolution of the accountability gap.*

LAYER 06 OF 07

Security & Adversarial Robustness

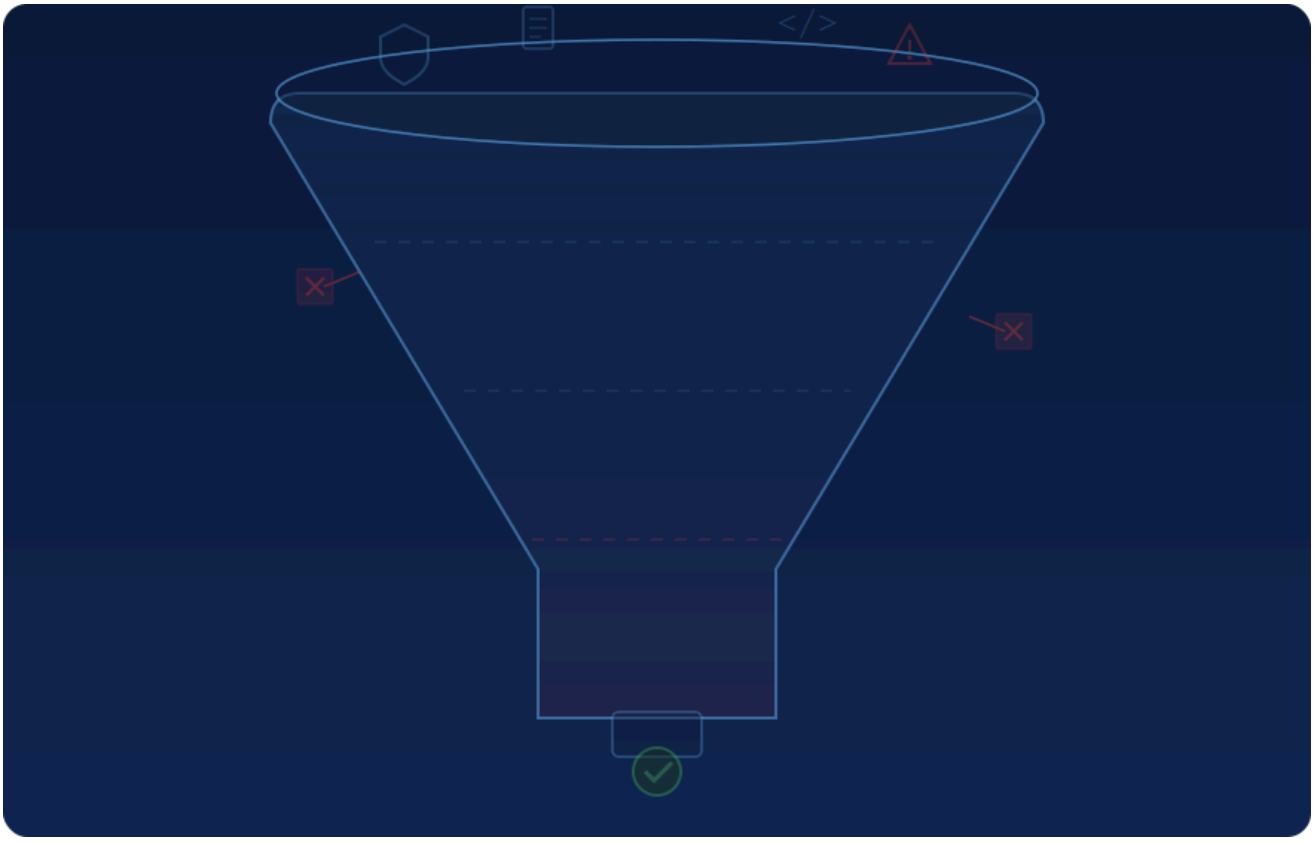
6

CORE QUESTION

Can the system be manipulated, extracted, or poisoned?

AI systems introduce attack surfaces that traditional cybersecurity frameworks were not designed to address. NIST AI 600-1 (2024) dedicates specific guidance to adversarial threats unique to generative AI: prompt injection, data poisoning, model extraction, confabulation under adversarial pressure, and evasion attacks. Traditional quantitative models face their own adversarial risks — gaming of scoring models, population shift through selective application behavior, and training data manipulation. The NIST Cybersecurity Framework AI Profile (NIST IR 8596, 2025) extends cybersecurity controls specifically to AI components. This layer governs whether your AI systems have been tested against adversarial inputs and whether controls exist to detect and respond to exploitation attempts in production.

The Prompt Injection Blind Spot. Organizations deploying LLMs and agentic AI frequently have robust security reviews for their underlying infrastructure but conduct no adversarial testing specific to the AI layer. Prompt injection — where malicious content in inputs redirects AI behavior in ways developers did not intend — is not a software vulnerability in the traditional sense. It exploits the model's core capability (following instructions) as an attack vector. In regulated environments, a successful prompt injection against an AI customer service or document review system can expose sensitive data, generate false compliance certifications, or produce decisions that violate policy. Testing for this is not optional once the system is in production.



METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
Adversarial Testing Coverage	Percentage of AI systems tested against adversarial inputs — including edge cases, distribution extremes, and intentionally crafted inputs designed to exploit model behavior — before initial production deployment.	(AI systems with documented pre-deployment adversarial testing ÷ total AI systems deployed in period) × 100	**100% for all new deployments. For GenAI and agentic systems: include red-team testing with prompt injection, jailbreak attempts, and instruction override tests.**	*Untested systems in production assume adversarial robustness rather than demonstrating it. Any system with user-facing inputs requires this testing.*
Prompt Injection Resistance Rate	*GenAI and agentic systems only.* Percentage of documented prompt injection attempts that the system correctly detects and blocks or contains, without executing the injected instruction. Measures the effectiveness of input filtering and system prompt protections.	(Injection attempts detected or safely contained ÷ total injection attempts in test suite) × 100. Maintain a versioned adversarial test suite; update with newly discovered attack patterns.	**≥95% on a current and comprehensive test suite. Accept only if the test suite is maintained and updated — a high score on a stale suite is false assurance.**	*Below 90% on a current test suite, or any test suite not updated in the last 90 days, indicates meaningful exposure to known attack patterns.*
Input Validation Coverage	Percentage of model	(Model input paths with	**100%. Every input path that can accept user-supplied or**	*Unvalidated input paths are attack surfaces.*

METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
	inputs with defined validation controls — type checking, range validation, sanitization, and anomaly detection — that execute before the input reaches the model.	defined, active validation controls ÷ total model input paths) × 100	**third-party data requires validation. "The model handles it internally" is not a control.**	*Even in internal-only systems, unvalidated inputs create poisoning risk if data pipelines are compromised.*
Security Assessment Recency	Time in days since the last adversarial security assessment (red team, penetration test, or adversarial audit) was completed for each material AI system. The threat landscape evolves; assessments must keep pace.	Current date − date of last completed adversarial security assessment. Track per system. Flag systems exceeding the recency threshold.	**≤180 days for high-risk and externally-facing systems. ≤365 days for internal, lower-risk systems. Reassess immediately after significant model updates or newly disclosed attack techniques.**	*>365 days without reassessment means the system's robustness profile reflects a threat model that may be significantly outdated.*

Implementation note. Traditional IT security teams are not automatically equipped to conduct adversarial AI testing. Prompt injection, model extraction, and data poisoning require AI-specific attack knowledge. Before your first red-team exercise on a GenAI system, verify that the testers have experience with AI-specific attack vectors, not just general application security. NIST AI 600-1 Appendix A provides a taxonomy of GenAI-specific risks that can serve as a baseline for structuring your test suite.

LAYER 07 OF 07

Human Oversight & Accountability

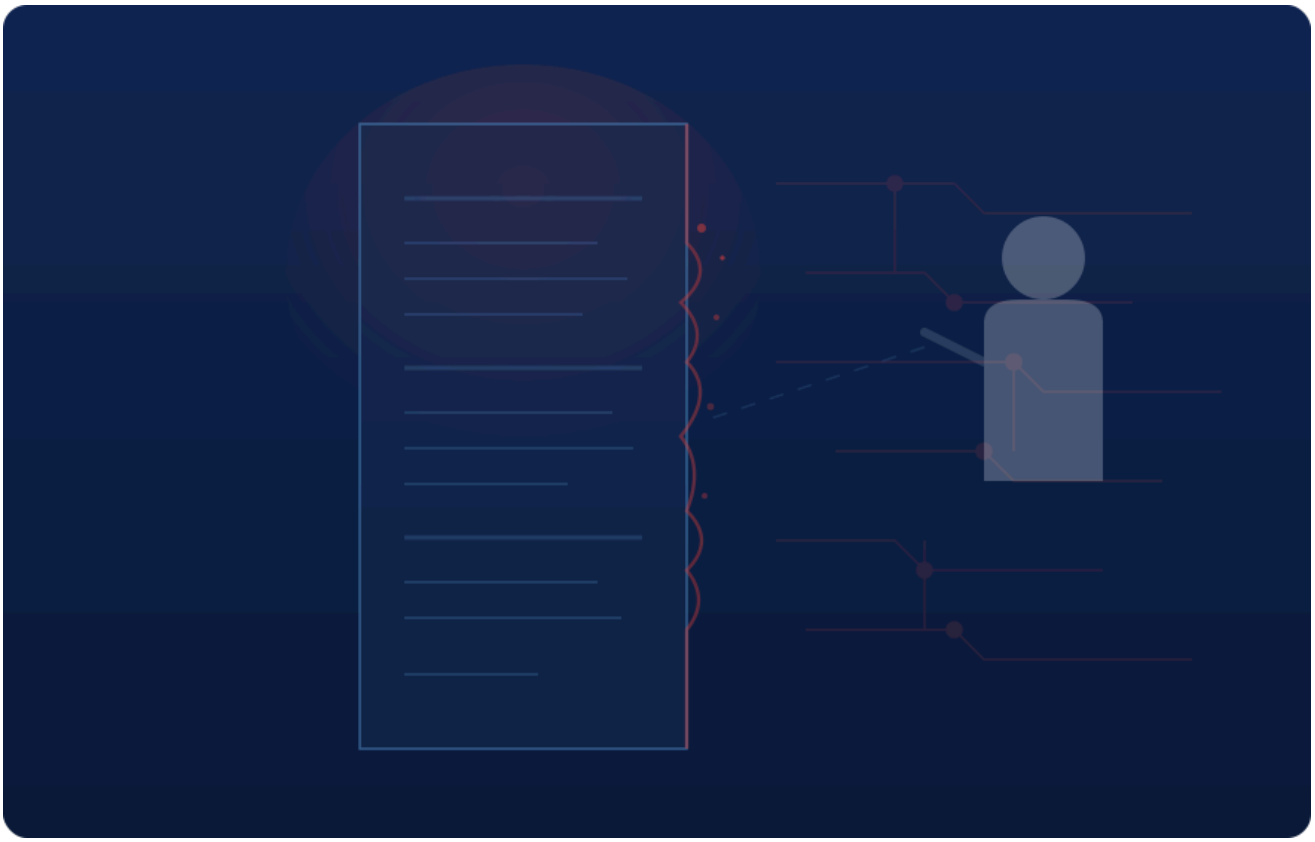
7

CORE QUESTION

When the AI is wrong, who is responsible and how fast do we catch it?

Human oversight is not a failsafe — it is a design requirement. SR 11-7 places accountability for model outcomes squarely on the institution, not the model. ISO/IEC 42001 requires documented human oversight levels for each AI system, commensurate with the consequence of errors. The EU AI Act mandates meaningful human oversight for all high-risk AI systems as a certification prerequisite. The challenge this layer addresses is not whether humans *can* intervene — it is whether the conditions, triggers, and accountabilities for intervention are defined before something goes wrong, not after. An organization that discovers it has no defined override procedure only when it needs one has failed the governance requirement, even if a human ultimately intervened.

The Automation Complacency Risk. Human oversight degrades when AI systems are reliable for long periods. Reviewers who have not seen an AI error in months stop scrutinizing outputs with the same care they applied initially — a well-documented cognitive phenomenon called automation complacency. Governance programs that track override rates often discover this signature: override rates that start at 3–5% and gradually decline to near zero, not because the model improved, but because human reviewers stopped challenging it. The appropriate response is not to blame the reviewers — it is to design oversight workflows that maintain engagement even during periods of apparent system reliability.



METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
HITL Trigger Coverage	Percentage of high-stakes AI decision categories with formally defined human-in-the-loop (HITL) triggers: the conditions under which the AI's recommendation must be reviewed by a human before being acted upon. Triggers should be based on confidence threshold, consequence magnitude, population segment, or irreversibility criteria.	(High-stakes decision categories with formally defined HITL triggers ÷ total high-stakes AI decision categories) × 100. Define "high-stakes" in governance policy before calculating.	**100% of defined high-stakes categories have triggers. Triggers should be documented, testable, and enforced by system design — not by behavioral expectation.**	*Undefined triggers mean oversight depends on individual judgment in real time. This is not a control.*
Human Override Rate	Percentage of AI decisions reviewed by a human that were overridden or modified before being acted upon. A non-zero rate is evidence of active oversight. A declining rate over time may signal automation complacency rather than model improvement.	(AI decisions overridden or modified by human review ÷ total AI decisions subject to human review) × 100. Track as a time series, not just a point-in-time figure.	**3–8% for well-calibrated systems in stable environments. Investigate both sharp increases (model degradation) and sharp decreases (oversight fatigue or complacency).**	*Sustained 0% override rate on a high-stakes model is a red flag: it may indicate reviewers are not substantively engaging with decisions.*
Accountability Assignment Rate	Percentage of AI-generated decisions that can be traced to a named human accountable for that decision	(AI decisions traceable to named accountable human ÷ total AI decisions in period) × 100.	**100%. Every consequential AI decision should resolve to a human accountable for the**	*Decisions without accountable owners are the governance equivalent of anonymous authorship: no one to*

METRIC	WHAT IT MEASURES	HOW TO CALCULATE	WHAT GOOD LOOKS LIKE	GAP SIGNAL
	class — either the system owner, the decision reviewer, or both. Accountability cannot be assigned to the AI. Someone must own every consequential output.	For fully automated pipelines: the accountable human is the system owner, not the reviewer of individual decisions.	**system's behavior in that domain.**	*call when something is wrong.*
Governance Training Completion	Percentage of roles with direct AI governance responsibility — model owners, validation staff, reviewers, product managers — who have completed current AI governance and responsible AI training within the required recertification window.	(Roles with current governance training certification ÷ total roles with governance responsibilities) × 100. "Current" = within the recertification window defined in policy (typically 12 months).	**100%. ISO/IEC 42001 explicitly requires competency management for AI governance roles. Untrained governance staff cannot perform effective oversight.**	*Training gaps in governance roles mean oversight is being performed by staff without established baseline competency. Identify gaps by role tier, not just aggregate percentage.*

SECTION 08

The 28-Metric Scorecard

Score each metric for your AI system or program: Yes (2 points), Partial (1 point), No (0 points). Maximum score: 56. Run this for your overall program, then for each material AI system individually.

#	Metric	Y	P	N
Layer 1 — Governance Structure				
1	AI Inventory Coverage: Every production AI system has a documented owner, defined scope, and explicit use-case boundaries.	☐	☐	☐
2	Policy-to-Control Ratio: ≥70% of governance policies have a corresponding enforcement mechanism (not a document — a system control).	☐	☐	☐
3	Independent Validation: 100% of material/high-risk AI systems have documented validation performed by a team independent of the developers.	☐	☐	☐
4	Regulatory Mapping: Every material AI system has a documented map of applicable frameworks (SR 11-7, NIST, ISO 42001, etc.) and addressed requirements.	☐	☐	☐
Layer 2 — Model & Data Integrity				
5	Data Lineage: 100% of material models have documented provenance for all training and validation data, including transformations and known exclusions.	☐	☐	☐
6	Reproducibility: 100% of material models can be retrained from documented source data within defined performance tolerance.	☐	☐	☐
7	Version Control: 100% of production models, training datasets, and configurations are under formal version control with change history.	☐	☐	☐
8	Validation Independence Score: 100% of material model validations were performed with documented organizational independence from the development team.	☐	☐	☐
Layer 3 — Operational Risk				
9	Monitoring Coverage: 100% of material production AI systems have automated, continuous performance monitoring with defined thresholds and alerting.	☐	☐	☐
10	Drift Detection Latency: Mean time from drift onset to alert is ≤7 days for high-risk systems, ≤30 days for lower-risk systems.	☐	☐	☐
11	Incident Response Time: Mean time from alert to deployed remediation (or documented risk acceptance) is ≤14 days for critical incidents.	☐	☐	☐
12	Threshold Calibration: Monitoring thresholds are documented with rationale, reviewed at least annually, and produce a threshold exceedance rate of 1–15% (neither silent nor hypersensitive).	☐	☐	☐
Layer 4 — Fairness & Bias				
13	Protected Segment Coverage: 100% of models with consumer decision impact have completed bias evaluation across all applicable protected classes.	☐	☐	☐
14	Disparate Impact Ratio: All material models maintain a ratio ≥0.80 per protected class in production on a rolling 90-day basis, with documented investigation of any breach.	☐	☐	☐
15	Fairness Assessment Frequency: High-risk models assessed ≤90 days ago. Other models ≤180 days ago. No model with decision impact >365 days without review.	☐	☐	☐

#	Metric	Y	P	N
16	Bias Remediation: 100% of identified bias findings have documented corrective action or formally accepted residual risk with named approver.	☐	☐	☐
Layer 5 — Transparency & Explainability				
17	Model Card Completion: 100% of production models have a current, complete model card updated within the last model version cycle.	☐	☐	☐
18	Explanation Coverage: 100% of individual decisions with consumer impact have on-demand explanation capability identifying the primary factors driving the output.	☐	☐	☐
19	Audit Package Readiness: 100% of material models can produce a complete audit package (purpose, data, methodology, monitoring, incidents) within 24 hours of request.	☐	☐	☐
20	Comprehensibility: ≥85% of explanations provided to non-technical stakeholders are rated as clear and sufficient by the recipient.	☐	☐	☐
Layer 6 — Security & Adversarial Robustness				
21	Adversarial Testing: 100% of AI systems have documented pre-deployment adversarial testing covering edge cases and intentionally crafted inputs.	☐	☐	☐
22	Prompt Injection Resistance (GenAI): ≥95% of documented injection attempts are detected or safely contained on a current test suite updated within the last 90 days.	☐	☐	☐
23	Input Validation: 100% of model input paths with user-supplied or third-party data have active validation controls executing before the input reaches the model.	☐	☐	☐
24	Security Assessment Recency: High-risk and externally-facing systems assessed adversarially within the last 180 days. All other material systems within the last 365 days.	☐	☐	☐
Layer 7 — Human Oversight & Accountability				
25	HITL Trigger Coverage: 100% of high-stakes AI decision categories have formally defined, system-enforced human-in-the-loop triggers.	☐	☐	☐
26	Override Rate: Human override rate is tracked as a time series and investigated for both sustained increases and sustained declines toward zero.	☐	☐	☐
27	Accountability: 100% of consequential AI decisions are traceable to a named accountable human (system owner, decision reviewer, or both).	☐	☐	☐
28	Governance Training: 100% of roles with direct AI governance responsibility have current training certification within the required recertification window.	☐	☐	☐

Scoring Guide

Score	Assessment	Priority
46–56	Strong governance foundation. Gaps are specific and addressable. Extend coverage to emerging risks (agentic AI, GenAI).	Maintain and extend
30–45	Partial coverage. Core governance exists but significant gaps remain — typically in layers 4–7 where AI-specific requirements exceed traditional model risk controls.	Close gaps by layer priority
14–29	Significant gaps. Governance may exist in documentation but lacks enforcement, independence, or AI-native coverage. High audit exposure.	Rebuild from Layer 1 up
0–13		

Score	Assessment	Priority
	Pre-governance state. AI systems operating without structured oversight. Regulatory risk is not theoretical — it is immediate.	Start with inventory and ownership

SECTION 09

From Score to Action

A low score is not a governance failure. It is a discovery. The failure would be not measuring at all. This section maps the most common gap patterns to their root causes and first actions.

Pattern A — High Layer 1, Low Layers 4–7

→ You have traditional model risk management infrastructure built for SR 11-7 but have not extended it to AI-specific risks (fairness, explainability, adversarial robustness, human oversight).

→ First action: Conduct a gap mapping exercise against NIST AI 100-1 and NIST AI 600-1, using your existing model inventory as the starting point.

Pattern B — Low Layer 1, Everything Else Unknown

→ You cannot score layers 2–7 because you do not have a complete inventory of what you are governing. This is the most common starting position.

→ First action: Conduct an AI discovery exercise across all business units. Include vendor-embedded AI, copilot tools, scoring models, and automated routing systems. Build the inventory before trying to build the governance.

Pattern C — Good Scores on Paper, Low in Practice

→ Documentation exists but controls are not enforced. Policies are written but not operational. Monitoring exists but no one acts on alerts.

→ First action: Conduct a control effectiveness test. Simulate a monitoring alert and measure actual response time. Request an audit package and measure how long it actually takes to compile. Governance theater is most visible under test conditions.

Pattern D — Strong Layers 1–3, Gaps in 4–5

→ Operational governance is solid, but fairness and transparency are treated as separate compliance activities rather than integrated monitoring obligations.

→ First action: Add fairness metrics and explanation coverage to your existing monitoring dashboards. Do not create a separate governance program — extend the one you have.

Regulatory Framework Cross-Reference

This table maps each layer to the primary regulatory frameworks that create obligations in that area. Voluntary today; expected tomorrow; enforced the day after.

Layer	SR 11-7	NIST AI 100-1	NIST AI 600-1	ISO 42001	FHFA AB 2022-02	EU AI Act
1 — Governance Structure	●●●	●●●	●●	●●●	●●	●●
2 — Model & Data Integrity	●●●	●●●	●●	●●●	●●	●●

Layer	SR 11-7	NIST AI 100-1	NIST AI 600-1	ISO 42001	FHFA AB 2022-02	EU AI Act
3 — Operational Risk	●●●	●●	●●	●●●	●●	●●
4 — Fairness & Bias	●	●●●	●●	●●	●●●	●●●
5 — Transparency & Explainability	●●	●●●	●●●	●●●	●●	●●●
6 — Security & Adversarial Robustness	●	●●	●●●	●●	●	●●●
7 — Human Oversight & Accountability	●●●	●●●	●●	●●●	●●	●●●

●●● = Core requirement ●● = Significant coverage ● = Partial or implicit coverage

"A governance program that cannot produce a number is not a governance program. It is a set of intentions."

AI UNDER AUDIT — PAULO CAVALLO, PHD

Continue with AI Under Audit

This playbook goes deeper every issue with case studies from real AI systems, regulatory analysis, and practical frameworks you can implement immediately. Free, for practitioners.

linkedin.com/newsletters/exit-code-2-ai-under-audit **pmcavallo.github.io**

GO DEEPER

Continue the Conversation

This playbook is the practitioner companion to the **AI Under Audit Field Guide**. The Field Guide covers the principles, regulatory map, and governance fundamentals. This playbook operationalizes those fundamentals into 28 measurable metrics. Use them together.

Get the AI Under Audit Field Guide

Five principles, three checklists, and one regulatory map for putting AI systems under audit in regulated industries. Free, like this playbook.

pmcavallo.github.io

PRIMARY SOURCES

SR 11-7 Guidance on Model Risk Management, Federal Reserve / OCC, 2011. The foundational U.S. banking standard for model governance — three-lines-of-defense, independent validation, ongoing monitoring.

NIST AI 100-1 Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST, January 2023. Four-function framework: Govern, Map, Measure, Manage. Cross-industry voluntary standard, increasingly cited in enforcement.

NIST AI 600-1 Generative AI Profile, NIST, July 2024. Extends AI RMF to GenAI-specific risks: confabulation, prompt injection, data poisoning, model extraction, third-party model risk.

ISO/IEC 42001:2023 AI Management Systems, ISO, December 2023. First international certifiable standard for AI management systems. 38 controls across nine governance areas. Plan-Do-Check-Act structure.

FHFA Advisory Bulletin 2022-02 Validation and Approval of AI/ML Models, FHFA, September 2022. GSE-specific extension of SR 11-7 to AI/ML, with explicit coverage of vendor-embedded AI and fair lending testing requirements across the full lifecycle.

Gartner Market Guide for AI Governance Platforms, 2025. Survey of 360 organizations; organizations deploying AI governance platforms were 3.4× more likely to achieve high governance effectiveness. Gartner projects governance technologies will decrease regulatory compliance costs by 20% by 2028.

NIST IR 8596 (Preliminary Draft) Cybersecurity Framework Profile for AI, NIST, 2025. Extends NIST CSF to AI-specific attack surfaces including adversarial inputs, data poisoning, and model extraction.

ABOUT THE AUTHOR

Paulo Cavallo holds a PhD in Public Policy and Political Economy and has spent eight years developing and validating credit risk models under SR 11-7 regulatory frameworks in U.S. financial services. He builds AI governance frameworks and multi-agent systems for regulated industries, and writes *AI Under*

Audit — a practitioner newsletter on what it takes to make AI systems auditable, explainable, and production-ready in environments where failure has consequences.

LINKEDIN	PORTFOLIO & RESOURCES	NEWSLETTER
linkedin.com/in/paulocavallo	**pmcavallo.github.io**	**Exit Code 2 | AI Under Audit**

"The question is not whether your AI works.
The question is whether you can prove it — on demand, to anyone."

New here? Start with the AI Under Audit Field Guide at **pmcavallo.github.io**