

Assurance Before Adoption: What Public Agencies Must Prove About AI

Procurement, fairness testing, and drift monitoring are now core public-trust responsibilities, not technical afterthoughts

By Gopichand Mannava, M.Eng., MBA, PMP

Chief Data & AI Architect, State of Connecticut | Independent Researcher

When the Public Is the Customer

A private company that deploys a flawed AI system loses revenue. A public agency that does it can deny someone a benefit, delay a claim, or misdirect a public-health response. Residents can't switch providers, and they rarely get to see how the decision was made.

That asymmetry changes what governance must deliver. My earlier article focused on moving policy into the pipeline. This one asks a different question for public-sector leaders: **what must an agency be able to prove about an AI system at each stage of its life, and to whom?**

The audience for that proof is broad. It includes legislators, auditors, oversight boards, advocacy groups, journalists, and residents themselves. Each asks a different version of the same thing: *Is this system safe, fair, and still working the way you said it would?*

Three Moments of Proof

I find it useful to treat AI assurance as three moments, each with its own evidence obligation:

- **Before purchase or build:** Do we know what this is, who is affected, and what could go wrong?
- **Before release:** Does it meet defined thresholds for accuracy, fairness, and safety?
- **After release:** Does it still behave as approved?

Most agencies invest heavily in the first moment, because procurement rules force it. They invest unevenly in the second and rarely in the third. Yet the third is where public harm most often goes unnoticed. A system that passed review at launch can drift as populations change, source data shifts, or a vendor quietly updates the underlying model.

For illustration, I'll use Pacific AI, a governance platform built for healthcare, because it organizes these three moments into distinct components: **Governor** for inventory and governance, **Gatekeeper** for pre-release testing, and **Guardian** for production monitoring. Healthcare is a fitting example for the public sector, since state agencies run or oversee hospitals, public-health programs, and Medicaid. The principles carry well beyond health.

Three Moments of Proof in the AI Lifecycle

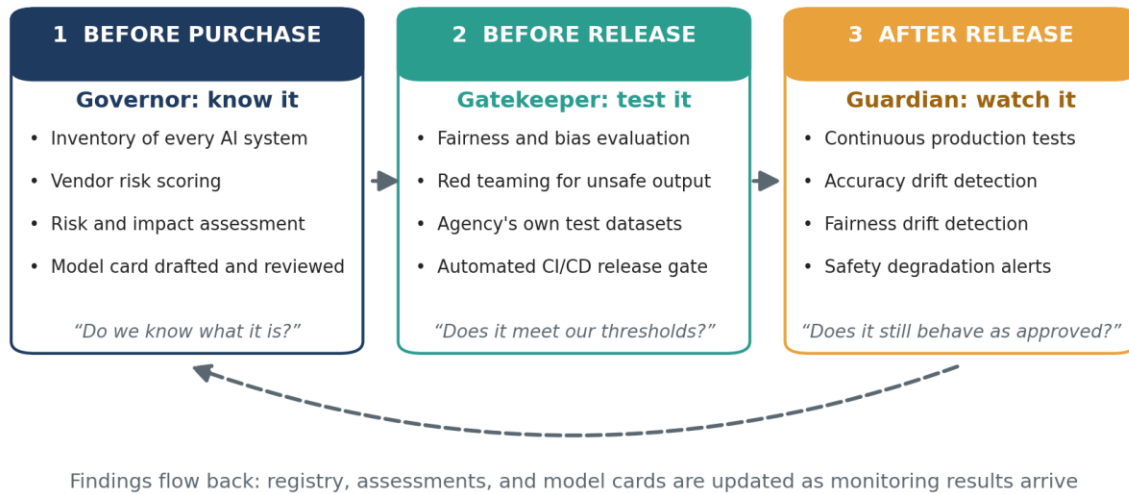


Figure 1. Public agencies owe evidence at three points in an AI system's life, not only at procurement.

"A system that passed review at launch has told you nothing about how it behaves today."

Moment One: Know What You Are Buying

The Vendor Problem

Public agencies buy far more AI than they build. It arrives as a standalone product, a feature inside an existing platform, or a capability a vendor switches on in a routine update. Accountability still sits with the agency.

This is why **vendor risk assessment** belongs in the governance core. A central registry that tracks every AI system, internal or external, along with a vendor risk score, converts "we trust the vendor" into a documented judgment that can survive an audit. Governor, for example, maintains a registry of systems, vendors, risks, model cards, and policies across the lifecycle.

The registry also answers a question that surprises many CIOs: *how many AI systems do we actually have?* Inventory comes first because every later control depends on it.

Risk Versus Impact

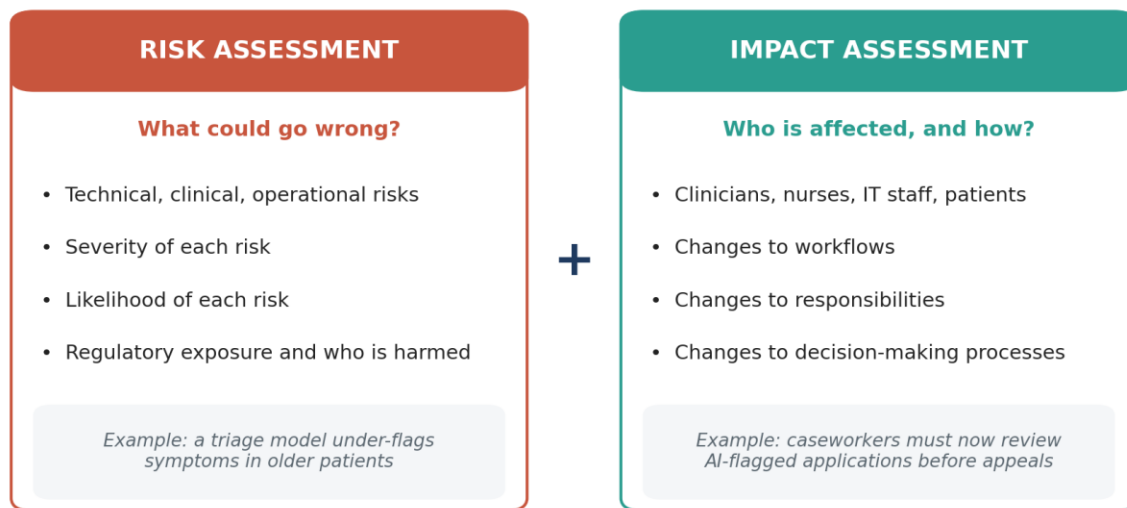
Two assessments are often confused, and public agencies need both.

A **risk assessment** asks what could go wrong: the technical, clinical, operational, and regulatory risks, each rated by severity and likelihood.

An **impact assessment** asks who is affected and how their work or circumstances will change. In a public hospital, that means clinicians, nurses, IT staff, and patients. In a benefits agency, it means caseworkers, applicants, and appeals staff. It examines how a system’s design and operation will alter workflows, responsibilities, and decision-making.

A system can be technically sound and still be a poor fit for the people who must work with it or live with its outputs. Risk assessment will not surface that. Impact assessment will.

Risk Assessment vs. Impact Assessment



A system can be technically sound and still be a poor fit for the people who must live with its outputs.

Figure 2. Public agencies need both assessments; risk analysis alone will not surface workflow and equity effects.

Documentation That Survives Scrutiny

Public agencies operate under transparency expectations that private firms often do not. A **model card** gives structured documentation of an AI system’s intended use, performance, limitations, and other relevant information. Governor can draft one automatically from uploaded documents and keep it versioned, with alignment to frameworks such as CHAI, ONC HTI-1, and California AB 2013.

Two practices matter here:

1. **Treat the generated card as a draft.** It can be manually edited, so the governance team can clarify a section before publication rather than rebuilding the system or releasing first and correcting later.
2. **Treat publication as a decision.** The benefits of publishing a model card are concrete: it supports compliance with transparency requirements and builds trust with the public. It

should be an authorized act by a designated role, not a byproduct of generating a document.

Roles and Separation of Duties

Public-sector governance depends on accountability that is visible and defensible. That requires clear roles. Pacific AI describes distinct roles, including an AI Governance Officer, Risk Manager, Policy Manager, and Compliance Officer. Publicly available descriptions indicate that a Risk Manager defines test suites and assessments, a Policy Manager manages policy creation and submission, and a Governance Officer holds publication and lifecycle-stage authority.

Exact permissions vary by platform and configuration, so verify them in your own environment. The principle is what counts: the person who prepares a decision should not be the only person who approves it. In government, that separation is already the norm for budgets, procurements, and records. AI should be held to the same standard.

Moment Two: Prove It Before Release

Fairness Is Not Optional

For a public agency, fairness testing is closer to a legal and ethical duty than a feature. If a clinical or eligibility-related system performs differently across demographic groups, the agency needs to know before approval, not after a complaint.

Gatekeeper automates this kind of pre-release testing for LLM, ML, and agentic systems. Its test suites span clinical and safety categories, including fairness evaluation across sociodemographic dimensions, medical benchmarks, and medical red teaming. A model that misses its threshold is blocked before it reaches patients.

What a Credible Pre-Release Program Includes

- **Standardized testing areas.** Fairness, bias, safety, robustness, privacy, and accuracy are testable and repeatable. Database performance and network bandwidth are infrastructure concerns, not AI assurance categories.
- **Red teaming.** To learn whether a language model produces unsafe or inappropriate responses, you have to challenge it deliberately with problematic inputs. Public-facing systems, such as resident chatbots, need this most, because the public will probe them whether you do or not.
- **Your own data.** Agencies often hold datasets that reflect their specific populations better than generic benchmarks. A platform should let you run a custom healthcare dataset in its evaluations.

- **Reusable test suites.** When you need a subset of an earlier suite, clone it, adjust the configuration, and run it. That preserves the original and its history, which matters when an auditor asks what was tested and when.
- **Comparison across models and versions.** Leaderboards let teams compare results on the same tests, so adoption decisions rest on evidence rather than vendor claims.

Make Testing Unavoidable

Evaluation that depends on someone remembering to run it will eventually be skipped. If developers want every new version evaluated automatically as part of the software release process, the answer is to integrate the evaluation into the CI/CD pipeline. A failing score then stops the release by default.

For endpoints, one technical requirement is easy to overlook: when Pacific AI evaluates an application through its API, the endpoint should provide predictable request and response structures. Agencies connect their language models by configuring LLM endpoints, and stable, documented interfaces are a prerequisite for reliable results.

Keep Sensitive Data Inside Your Boundary

Public agencies handle protected health information and other sensitive records. Pacific AI states that it deploys inside the customer's own cloud account, with governance data, model cards, risk assessments, and test results remaining within that environment. Whatever platform you evaluate, ask this question first, and get the answer in writing.

Moment Three: Prove It Stays True

Consider a system that passed its pre-release evaluations six months ago and is now behaving differently in production. The wrong responses are to assume the original test results remain sufficient or to repeat unrelated administrative steps. The right response is continuous monitoring to identify and diagnose drift.

Guardian tracks accuracy drift, fairness drift, and safety degradation across deployed systems and flags issues before they affect care. Its most important design feature is that **the same test categories run both before release and in production**. Because "acceptable" is defined the same way in both places, a change in a score is meaningful.

Why a Passing Grade at Launch Is Not Enough

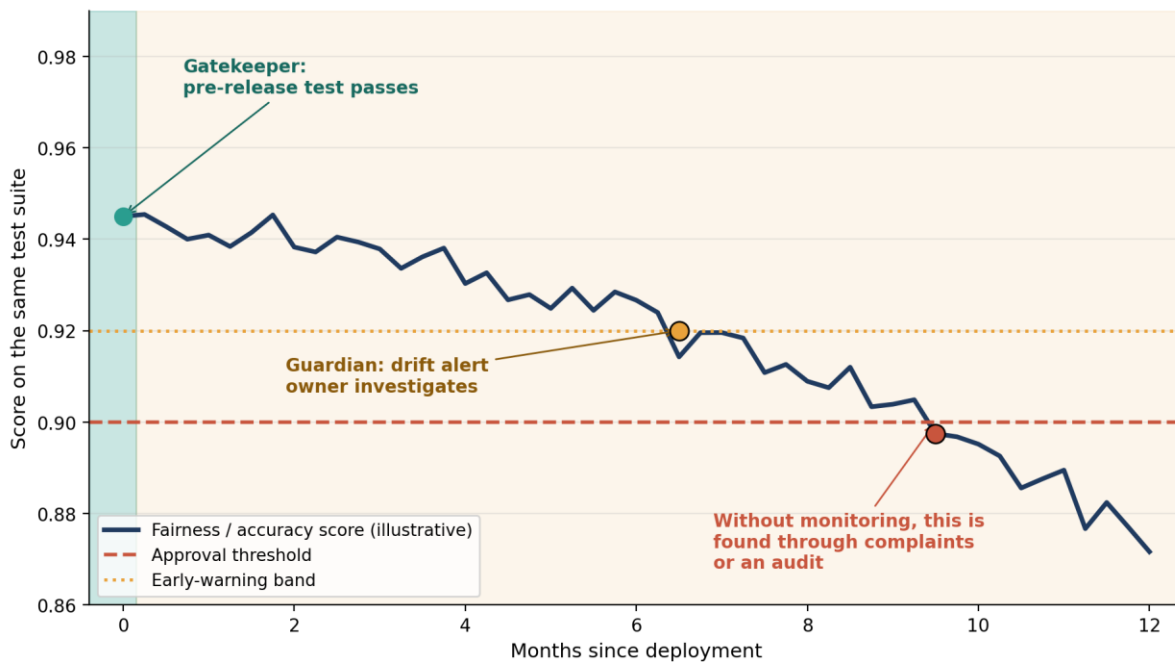


Figure 3. Illustrative only: running the same tests before and after release makes gradual drift visible before it becomes harm.

For public agencies, drift monitoring is also an equity safeguard. Populations change, eligibility rules change, and vendor models change. A system that was fair at launch can become unfair gradually, with no single event to point to. Continuous monitoring is how that erosion becomes visible.

The Accountability Chain

Put together, the lifecycle looks like this:

1. **Register** the system and its vendor, and assign an owner.
2. **Assess** risk and impact, including effects on affected staff and residents.
3. **Document** with a model card that is edited, reviewed, and published by an authorized role.
4. **Test** for fairness, safety, and robustness, using red teaming and your own data, as a pipeline gate.
5. **Approve and deploy** only when thresholds are met.
6. **Monitor** for drift, with named owners for each alert.
7. **Update** the registry, assessments, and documentation as findings arrive.

The result is a continuous record that answers oversight questions with documentation instead of recollection.

The First 90 Days

Public agencies rarely have the luxury of a large new program. Three steps are realistic within a quarter:

1. **Take a census.** Inventory every AI system in use, including vendor features and embedded tools. Assign an owner and a risk tier to each.
2. **Choose one high-visibility system.** Run it through impact assessment, fairness testing, and monitoring setup. Use what you learn to refine your template.
3. **Rewrite one contract clause.** Require vendors to disclose model changes, support independent testing, and provide documentation sufficient for a model card. Contract language is one of the strongest levers public agencies hold.

Be candid about limits. No platform replaces accountable officials, public-records obligations, or human review of high-impact decisions. Tools make assurance faster and more consistent. They do not make it unnecessary.

Closing Thought

Private firms can sometimes recover from AI failures through apology and product updates. Public agencies recover only by earning trust back, which is slow and costly.

The better path is to be able to show, at any moment, that a system was understood before it was bought, tested before it was released, and watched after it went live. When agencies can do that, AI adoption stops being a leap of faith and becomes a decision that can be examined, defended, and improved.

In government, the license to use AI is granted by the public. Assurance is how we keep it.