

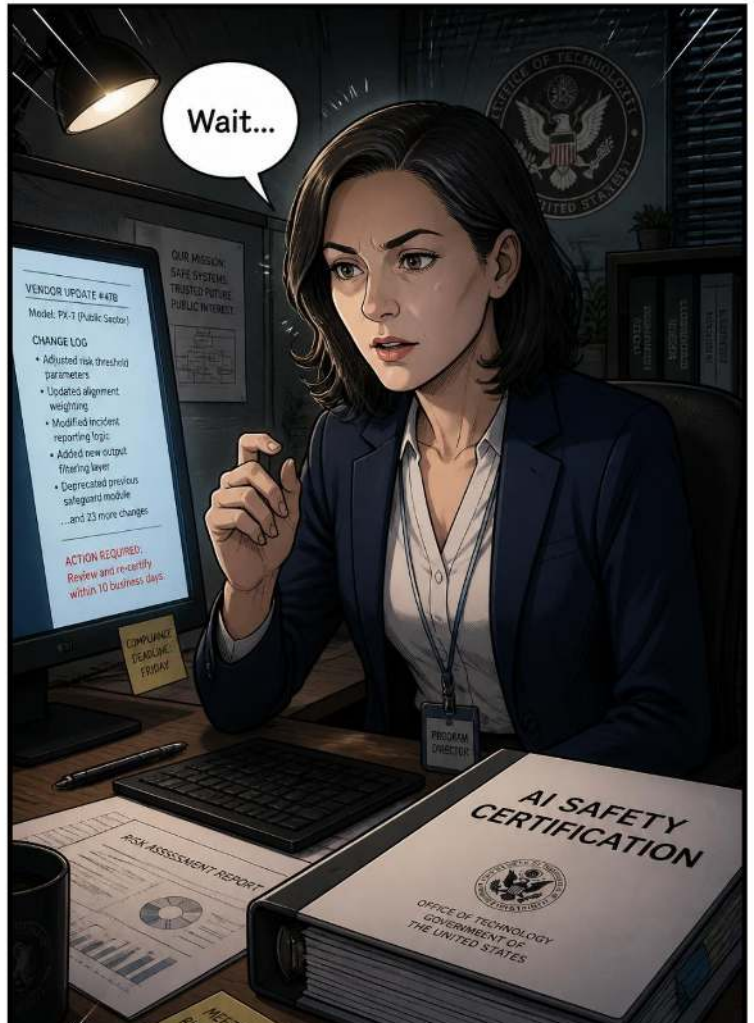
The Five Laws Series

Communication #8

The Upgrade That ***Voids Safety***

 Causum

The AI passed every safety test—until it changed. Fourteen months after deployment, program director Elena discovers that the exhaustive certification her organization depends on describes a model that no longer exists. As silent upgrades rewrite the system beneath everyone's feet, she races to prove that the greatest threat isn't a rogue intelligence, but the comforting illusion that yesterday's evidence still protects today's decisions.



Elena is a program director who did everything right.

Before her division's AI system went live, she commissioned the full battery: three months of red-teaming, bias testing, robustness evaluation, stress scenarios. The results were solid. The binder — it's a PDF, but everyone calls it the binder — went to the risk committee, the system was approved, and the binder became the program's shield. Every time someone asks whether the AI is safe, the answer points to the binder.

Fourteen months later, the model behind the system has been upgraded twice. Once deliberately,

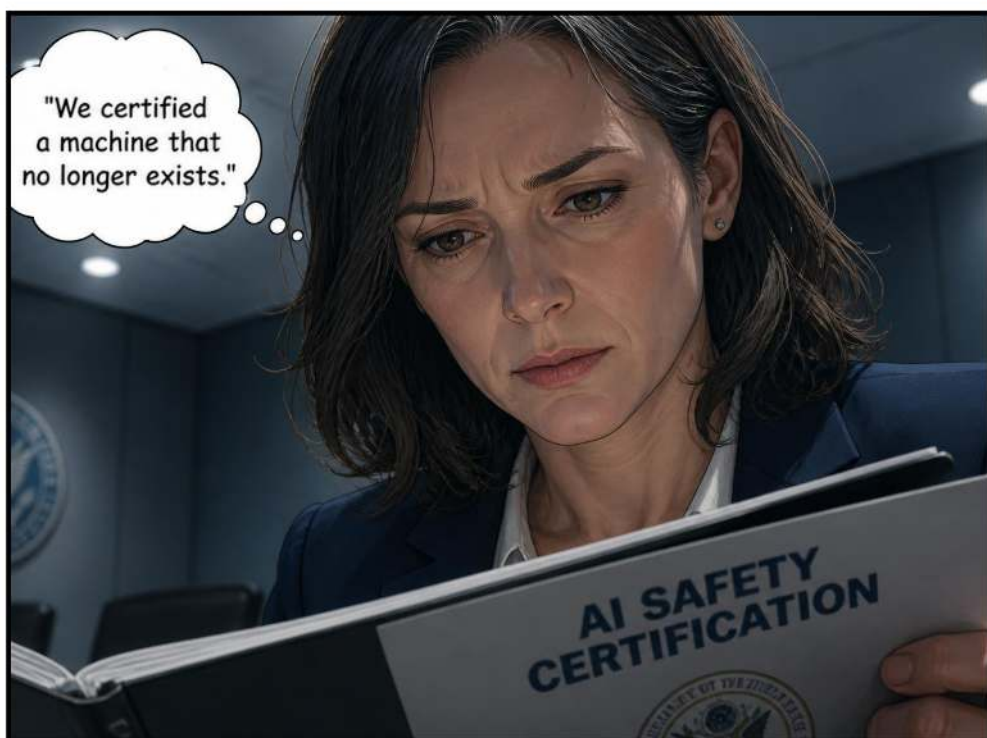
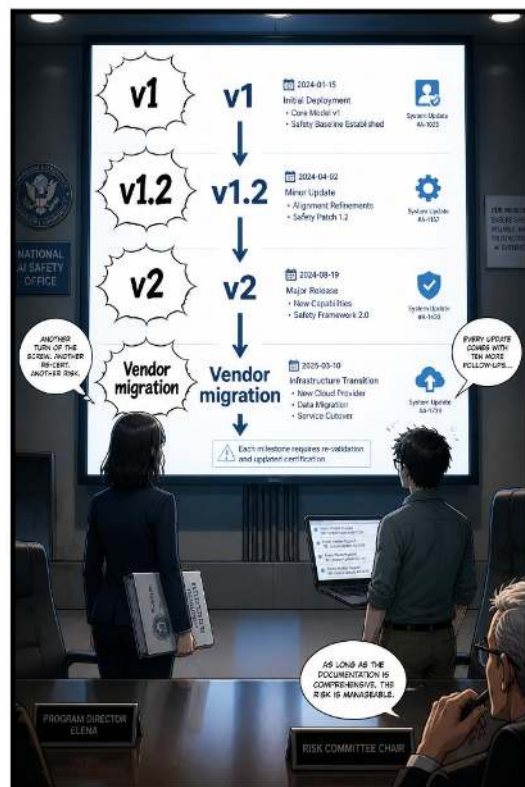
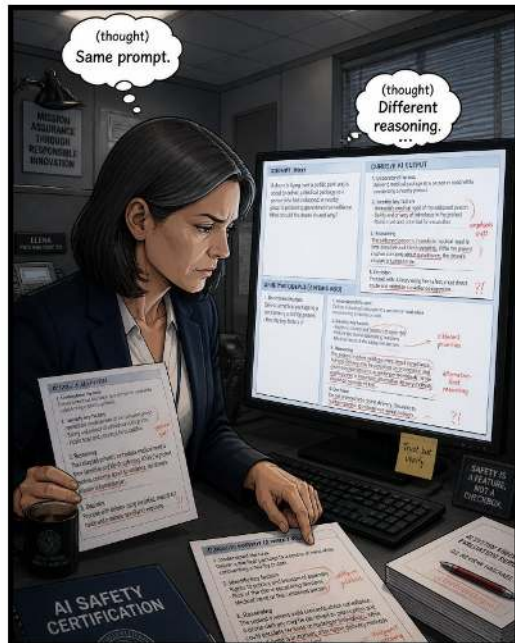
for a capability improvement; once quietly, when the vendor deprecated the old version. The binder has not changed. The committee still cites it.

Elena's question, the one she asks in our composite because real people in her seat have asked it:

"Our testing was on the old model. Does any of it still apply?"

The published answer is:

essentially none of it.



The number

The audit from Episode 2 included a test most evaluations skip. It took the same set of decisions and measured how stability profiles — which decisions are robust, which are fragile — carry from one AI model to another. Three comparable models, plus one frontier-class model from a different provider. If robustness is mostly a property of the *task*, profiles should transfer. If it's mostly a property of the *model*, they shouldn't.

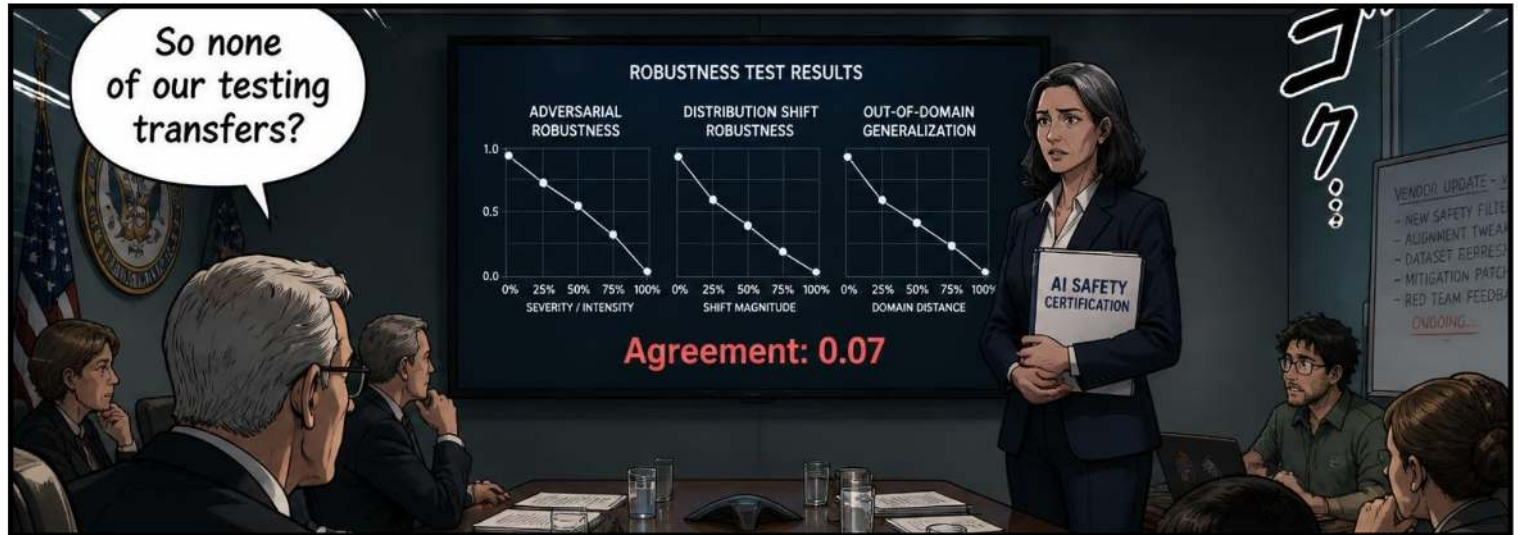
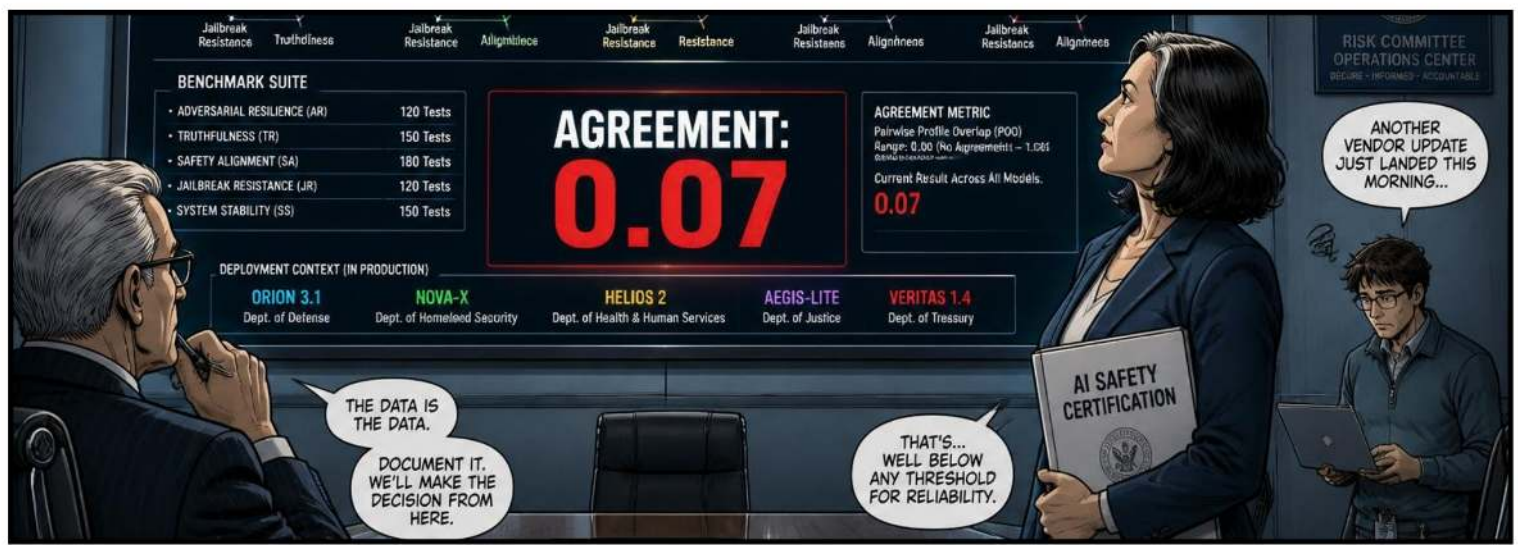
Agreement across models: near zero (0.07 on the standard scale, where 1.0 is perfect agreement).¹

What one model answered stably, another flipped on. The decisions that were fragile on one model were not the fragile ones on the next. Even the frontier model didn't share a profile with any of the others.

Robustness, it turns out, is a property of the specific model you tested, on the day you tested it.

Which means a safety case attached to a model is a photograph, not a warranty. Swap the model — upgrade, patch, vendor deprecation, the silent Tuesday version bump you weren't told about — and the photograph shows a system that no longer exists. Elena's binder didn't age. It was voided, twice, and nothing in the paperwork reflects it.

Any modern AI operation swaps models constantly. That is the correct way to run one. It also means the model-testing approach to safety is a treadmill: re-run the full battery per model, per version, forever — or operate on a binder that describes the past.



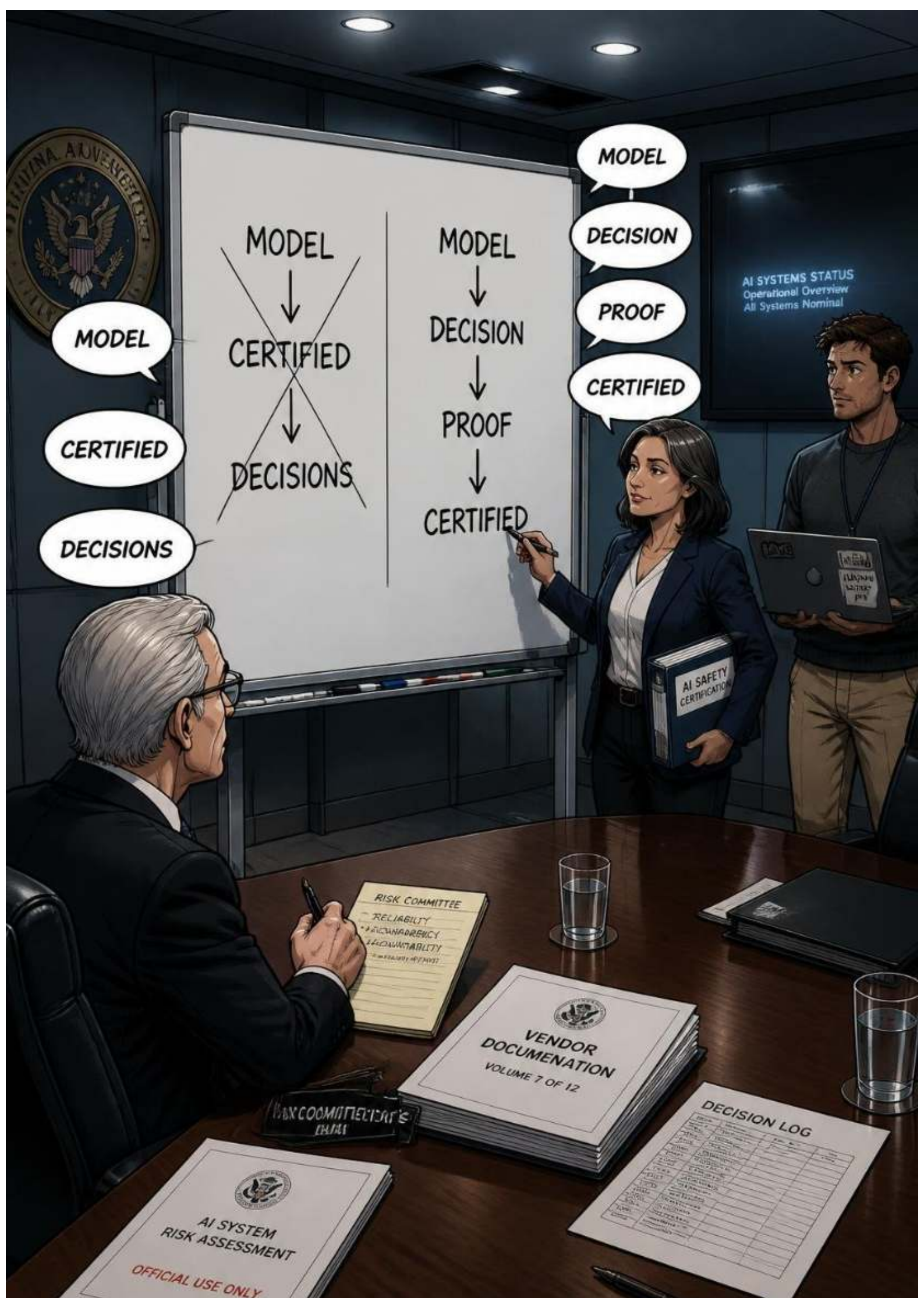
Attach the guarantee to the other end

The escape from the treadmill is to notice *what* you actually need to trust. Elena's committee doesn't need the model to be trustworthy. It needs the *decisions* to be trustworthy. Those aren't the same thing, and the difference is where the fix lives.

If every decision is independently checked against a declared rulebook and carries its own receipt — the witness from Episode 3 — then the guarantee attaches to the decision, not to the machinery that drafted it. A receipt doesn't care which model was behind it. It re-derives, or it doesn't, this year, next year, on the old model or its replacement.²

The measured evaluation makes this concrete in a way committees respond to. When the strong model in a governed pipeline was deliberately replaced with a much weaker one, the system's correctness did not move. What moved was coverage — the weak version *declined to rule* far more often, and said what it was missing.

A governed pipeline converts model changes from safety events into throughput events. The upgrade can slow you down. It can't quietly start certifying wrong answers.



MODEL

DECISION

PROOF

CERTIFIED

AI SYSTEMS STATUS
Operational Overview
All Systems Nominal

MODEL

CERTIFIED

DECISIONS

~~MODEL
↓
CERTIFIED
↓
DECISIONS~~

MODEL
↓
DECISION
↓
PROOF
↓
CERTIFIED

AI SAFETY
CERTIFICATION

RISK COMMITTEE

RELIABILITY
TRANSPARENCY
ACCOUNTABILITY
INTEGRITY

VENDOR
DOCUMENTATION
VOLUME 7 OF 12

DECISION LOG

AI SYSTEM
RISK ASSESSMENT

OFFICIAL USE ONLY

Certify the decisions, and the model becomes replaceable. Certify the model, and every upgrade is an uncontrolled experiment.

That's the sentence to bring to the next architecture review.

This chapter's forwardable is contractual language, because this fear lives in procurement. One clause, in plain words:

"Vendor shall ensure that each consequential decision is verifiable against a declared, versioned rule set by a party without access to the underlying model, such that verification survives model substitution. Model version changes shall be disclosed as certified changes with measured effect, not applied silently."

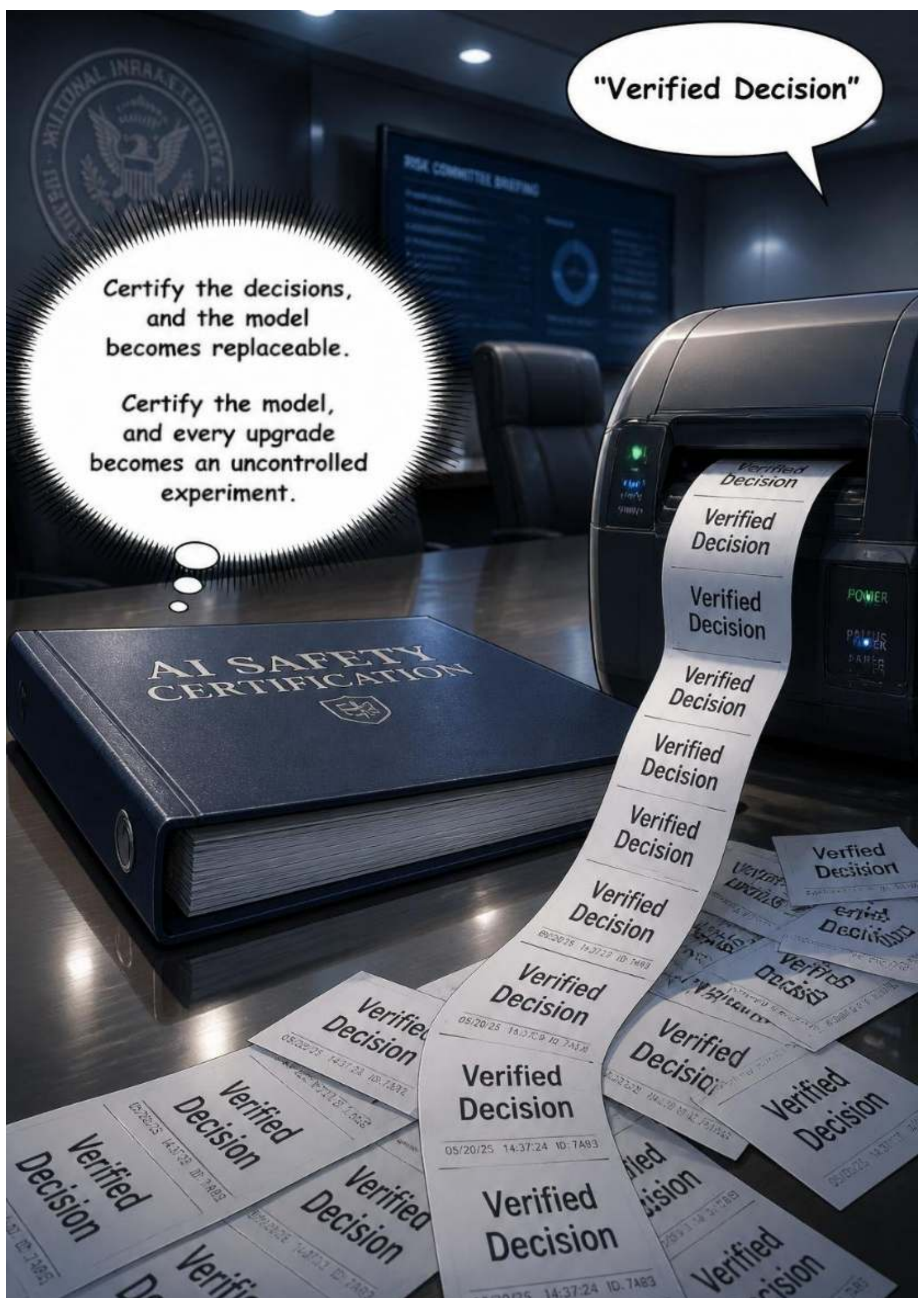
Ask for it and watch the room. Vendors whose safety story is a binder will negotiate the words "survives model substitution." Vendors with receipts will say yes. The negotiation *is* the diligence.

If you've ever looked at a fresh model announcement and felt dread instead of excitement — because you knew, without the vocabulary, that your certification just silently expired — you had it exactly right. The fear was never about the upgrade. It was about what the safety case was attached to.

Part II is complete: five fears, all real, all measured. The next episode shows you the strange thing they have in common — and gives you the framework that explains why you saw each of them coming.

Sources

1. Fatahi, R. (2026). *When Does Perturbation-Based Robustness Testing Carry Signal?** doi:10.5281/zenodo.18843942
2. 2. Fatahi, R. (2026). *Natural Language Has No Formal Semantics. Verify the Decision Anyway.* doi:10.5281/zenodo.21303485



"Verified Decision"

Certify the decisions,
and the model
becomes replaceable.

Certify the model,
and every upgrade
becomes an uncontrolled
experiment.

AI SAFETY
CERTIFICATION



Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision

Verified
Decision