

**The Five Laws Series**

Communication #11

# **Answerable by Design**



 **Causum**

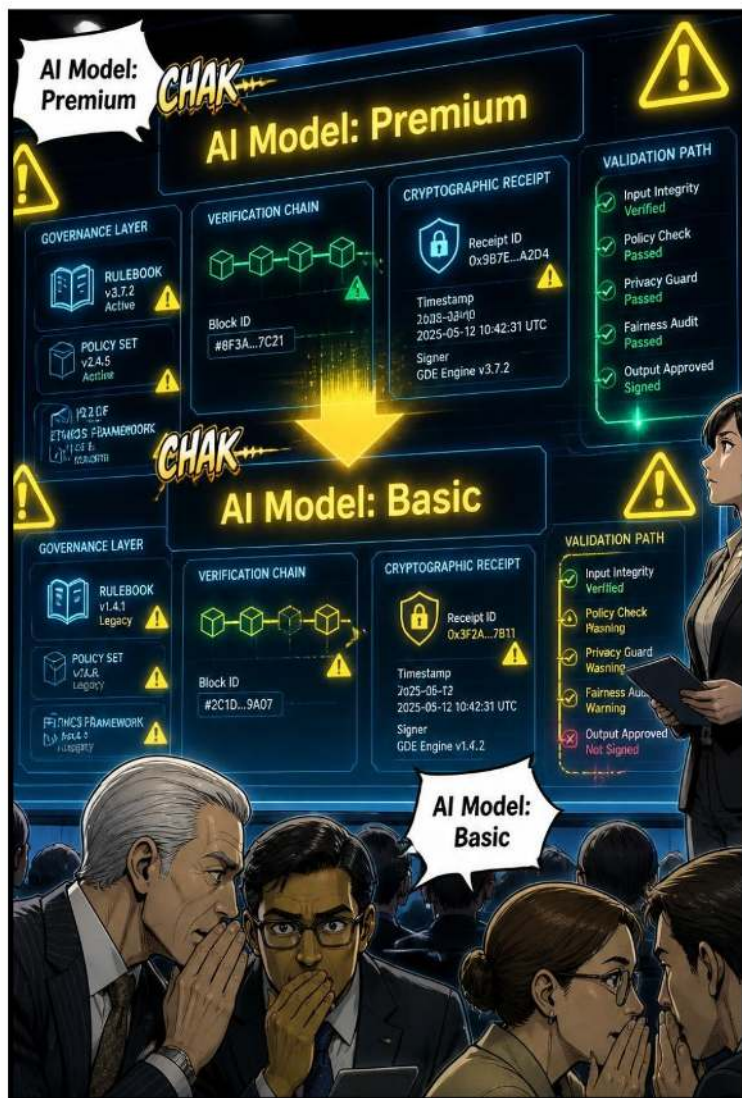
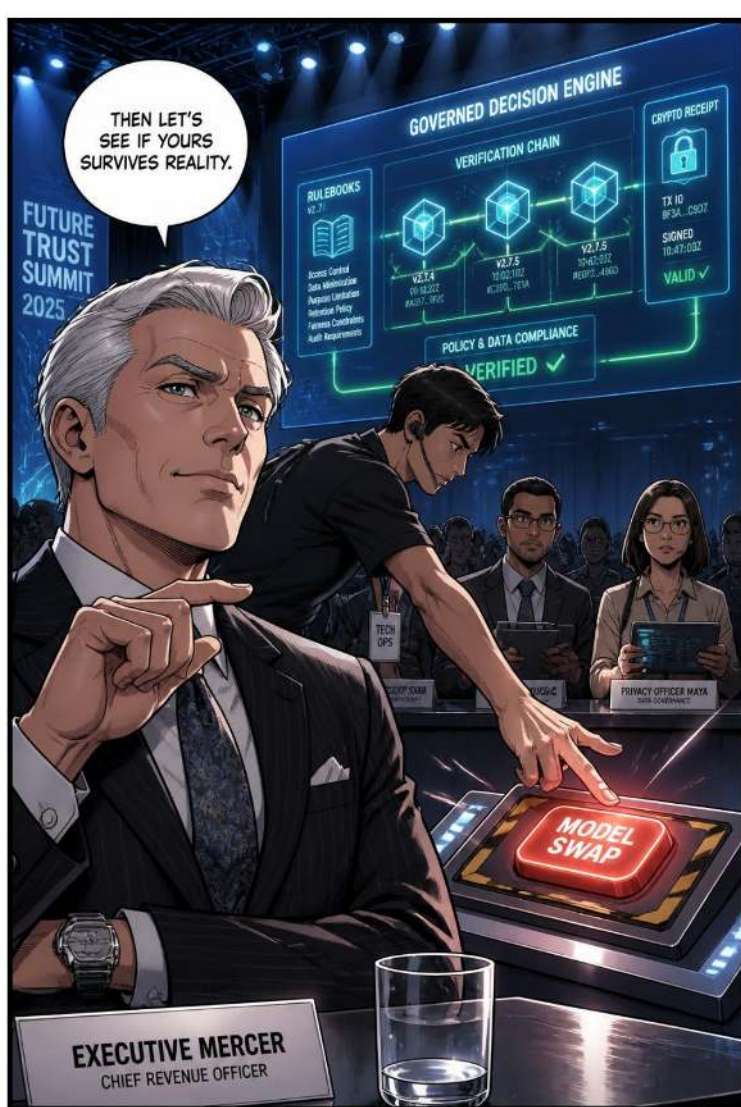
In a world where every AI promises responsibility, one engineer unveils something far more dangerous to the industry: a machine that can prove every decision it makes. As rivals, regulators, and executives try to break it in real time, each attack only exposes a deeper truth—trust isn't a feature. It's architecture.

The fifth law said optimization must answer to those it affects. This episode is about what answerability looks like when it's engineered instead of promised — and why the companies that build it first will treat it as a commercial weapon, not a compliance cost.

Start with the word itself. "*Answerable*" gets used in AI ethics documents the way "*excellence*" gets used in mission statements: warmly and without consequence. Here is a definition with teeth, assembled from everything this book has measured:

**A system is answerable when the rules it operates under are declared, versioned, and owned by a named authority; when any affected party can have a decision re-derived against those rules without trusting the operator; and when the rules can change — but only in the open, as a versioned event, never silently.**

Declared rules. Outsider re-derivation. Change in the open. Each clause is buildable, and each was measured in the study you've been following. Three consequences deserve their own airtime, because each one converts a compliance anxiety into an operating advantage.

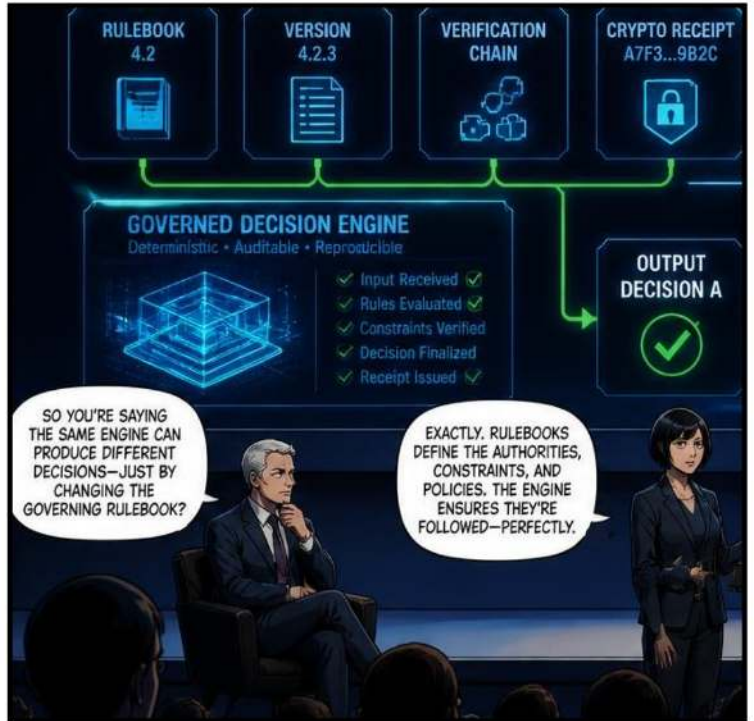


# Upgrades stop being safety events

Episode 8 left Elena on the testing treadmill: every model change voids the binder. Here's the measured alternative.

In the governed evaluation, the researchers did something vendors don't do in public: they made the system *worse* on purpose. The capable model in the pipeline was swapped for a far weaker one, with the rulebook and verification held fixed. Result: **coverage fell from 92% to 15% — correctness didn't move.**<sup>1</sup> The weak system ruled on far fewer cases and returned *"here's what I couldn't establish"* for the rest. Every verdict it did assert still proved out.

Read as an operating principle: ***component quality buys coverage, never correctness.*** In a governed pipeline, the model can't corrupt the answers; it can only change how many questions get answered automatically. A vendor deprecation, a quiet version bump, an ambitious upgrade — throughput events, monitored like throughput, while the correctness guarantee sits with the rulebook and the receipts, which didn't change. Elena gets off the treadmill: recertify the *change*, not the world.



# Jurisdictions stop being rebuilds

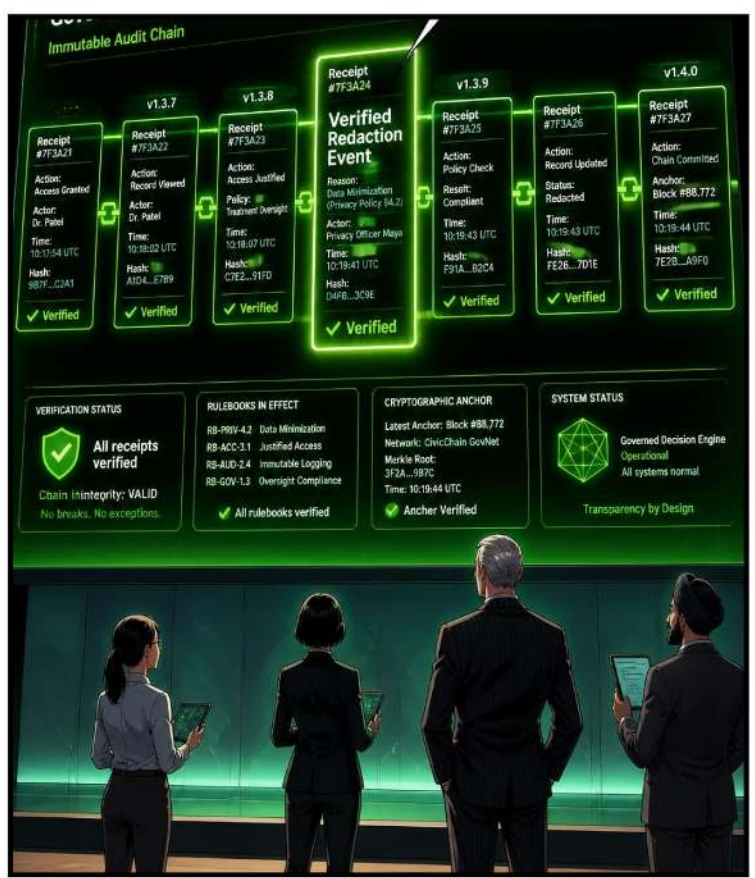
Because the rulebook is a declared, versioned artifact rather than behavior baked into a model, operating under a different authority means declaring a different rulebook. In the study, the specification for a new domain was written in a single pass, reusing thresholds from a prior domain unchanged, and hit the ceiling of episode 10 without tuning.<sup>1</sup>

For anyone selling or deploying across borders, that's the answer to the question that kills expansion timelines: *what happens when the next market's rules differ?* New jurisdiction, new declared rulebook, same engine — and where authorities conflict, the system carries both and says which one governed each decision, rather than blending them into one silent average. Answerability localizes. The alternative — model-embedded behavior — answers to whatever mixture it was trained toward, which is to say: to no one.

# Privacy stops fighting auditability

The standing objection to permanent evidence is the right to erasure: regulators demand records be deletable, auditors demand records be immutable, and every data-governance meeting re-fights that war. The measured system dissolved it: lawfully redacted values were removed while the receipt chain stayed verifiable — the redaction itself became a recorded, provable event, and every check that didn't depend on the erased value still re-derived.<sup>1</sup> Deletion with a receipt.

"Auditable *or* privacy-compliant" was a false choice; you're allowed to stop honoring it.



# The moat

Now step back and look at what a company holds when all three properties are real. Procurement committees, the fastest-growing gate in enterprise AI, are converging on exactly these questions — who owns your rules, can we verify decisions independently, what happens on model change, how do you handle erasure. Most vendors answer with adjectives. An answerable system responds with artifacts: *here is the rulebook and its version history, here is a decision — re-derive it yourself, here is the measured effect of our last model swap, here is a redaction receipt.*

Artifacts beat adjectives in every committee that matters, and the gap widens as regulation converges on the five laws' failure conditions — which is the direction every major framework is already drifting. The declaration's line runs both directions: efficiency without legitimacy is exploitation automated,<sup>2</sup> and in a market that has noticed, legitimacy *with* efficiency is trust automated. Trust, automated and provable, is the scarcest commodity in the entire AI economy right now.

Answerability isn't the tax on the roadmap. On a three-year horizon, it *is* the roadmap.

## Sources

1. Fatahi, R. (2026). *Natural Language Has No Formal Semantics. Verify the Decision Anyway*. doi:10.5281/zenodo.21303485
2. Kanjani AI Research (2026). *Declaration of the Five Laws of Mediated Intelligence Systems*. AIGP v4.0 specification.

GLOBAL  
TRUST  
UMMIT  
2025

TRUST IS  
NO LONGER  
A CLAIM.  
IT'S A  
VERIFIABLE  
SYSTEM.

THEY ANSWERED  
WITH ADJECTIVES.  
WE ANSWER  
WITH ARTIFACTS.

# GOVERNED DECISION ENGINE

 **RULEBOOK**  
Eligibility\_Rules v2.3.1  
Effective: 2025-04-01  
Hash: 7F3A...9C1B  
Status: ACTIVE

 **DECISION POLICY**  
Credit\_Assessment v4.2.0  
Effective: 2025-04-01  
Hash: 86C2...D4E9  
Status: ACTIVE

 **RULEBOOK**  
Adverse\_Action  
Effective: 2025-02-01  
Hash: 3C7B...8F7D  
Status: ACTIVE

 **RULEBOOK**  
Fraud\_Checks v1.9.4  
Effective: 2025-09-15  
Hash: A91D...7622  
Status: ACTIVE

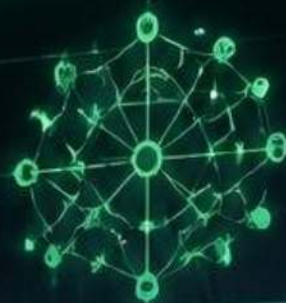
## VERIFICATION CHAIN

-  Request Received  
2025-05-20 10:14:22 UTC  
ID: REQ-8FD...A1
-  Rules Evaluated  
2025-05-20 10:14:23 UTC  
ID: EVF-3CAL...07
-  Decision Computed  
2025-05-20 10:14:34 UTC  
ID: DEC-9011...E3
-  Decision Published  
2025-05-20 10:14:25 UTC  
ID: PUB-7855...C9
-  Receipt Issued  
2025-05-20 10:14:26 UTC  
ID: PCP-2466...D4

# THEY ANSWERED WITH ADJECTIVES. WE ANSWER WITH ARTIFACTS.

## DECISION EVIDENCE (HUMAN-READABLE)

- Applicant meets income threshold. ✓
- No adverse history found. ✓
- Identity verified via KYC provider. ✓
- Debt-to-income within policy limits. ✓
- Decision: **APPROVED**



GOVERNED  
DECISION  
ENGINE