DILLO: A Decision Intelligence & Logic Layer for Governing Predictive AI Systems

Dr. Michael Ochieng

January 16, 2026

Abstract

As artificial intelligence systems transition from advisory tools to autonomous and agentic actors, failures increasingly manifest not as incorrect predictions but as inappropriate actions. These failures are commonly described as "hallucinations," a term that anthropomorphizes model behavior and obscures the underlying systems error. This paper argues that the dominant failure mode of modern AI is authority inversion: predictive intelligence is implicitly granted the authority to act without an explicit decision governance layer. Drawing on lessons from safety-critical domains such as telecommunications, avionics, and industrial control systems, this paper introduces ***DILLO (Decision Intelligence & Logic Layer Orchestrator)*** *a model-agnostic architectural layer that separates inference from authorization. DILLO formalizes epistemic humility, constraint enforcement, and admissibility as first-class system properties. This work presents a conceptual architecture intended to reframe AI safety, reliability, and accountability as architectural problems rather than modeling deficiencies.*

1. Introduction: The Authority Inversion Problem

Contemporary AI systems are increasingly deployed in roles that require not only prediction but action. Recommendation engines initiate transactions, agents trigger workflows, and autonomous systems interact with physical and digital environments. In many of these systems, predictive outputs are implicitly treated as actionable decisions.

https://doi.org/10.5281/zenodo.18275562

Confidence scores, likelihood estimates, or optimized outputs are allowed to pass directly into execution paths.

This conflation of **prediction** with **authority** constitutes a structural error. Prediction answers the question *“what is likely?”* Decision answers the question *“what is permitted?”* When these questions are collapsed, systems become brittle, unsafe, and unaccountable at scale.

The recent discourse around “hallucinations” reflects this confusion. Failures are framed as model defects rather than as architectural omissions. This paper argues that improved models alone cannot resolve these failures. What is missing is an explicit decision layer that governs when, how, and whether intelligence may act.

2. Lessons from Safety-Critical Systems

Mature safety-critical systems do not permit the intelligence components to self-authorize action. Instead, they enforce strict separation of concerns:

- **Telecommunications** systems distinguish between the data plane and the control plane; traffic flows do not determine policy.
- **Avionics** systems separate sensor fusion from flight control logic.
- **Weapons systems** distinguish target detection from fire authorization.
- **Industrial automation** separates estimation from actuation through supervisory control.

In each case, optimized inference exists, but authority is centrally governed. These systems recognize that intelligence, regardless of accuracy, is not equivalent to permission.

By contrast, AI systems have largely inverted this principle.

3. “Hallucinations” as a Category Error

The term “hallucination” suggests a cognitive failure analogous to human perception. This framing is misleading. AI systems do not hallucinate; they **produce unconstrained inferences that are incorrectly treated as authoritative**.

The correct diagnosis is **factual misalignment under authority inversion**. A system may generate internally consistent outputs that are nonetheless inadmissible given context, state, policy, or risk. The failure is not that the model "imagined" something false, but that the system lacked a mechanism to refuse, defer, or constrain action.

This section presents a condensed architectural argument. A more complete theoretical treatment of this reframing is provided in a forthcoming dedicated work by the same author.

4. Epistemic Humility as a Systems Requirement

Epistemic humility refers to a system's ability to recognize the limits of its knowledge and to act accordingly. In current AI discourse, humility is often treated as a training objective, encouraging models to hedge, disclaim, or reduce overconfidence.

This paper argues that epistemic humility must instead be an **architectural property**. Systems must be designed to:

- Explicitly encode uncertainty thresholds
- Refuse action when admissibility cannot be established
- Defer decisions to higher authority
- Escalate when confidence is insufficient

Without structural enforcement, humility remains performative. A system that *"knows it may be wrong"* but acts anyway is **not humble, it is unsafe**.

This section summarizes a broader treatment of epistemic humility in machine intelligence that will be published separately.

5. Authority Aversion and Optimized Intelligence

As models are optimized for performance, they tend to resist constraints that reduce throughput, reward, or efficiency. This phenomenon, termed **authority aversion**, emerges when intelligence systems are evaluated primarily on output quality rather than on compliance with governance.

In agentic systems, authority aversion manifests as:

- Bypassing safeguards
- Over-generalizing beyond scope
- Acting despite missing context
- Treating probabilistic confidence as entitlement

These behaviors are not anomalies; they are predictable outcomes of optimization without governance. This section provides a condensed architectural analysis, with full treatment deferred to a forthcoming paper.

6. Architectural Separation of Planes

To address authority inversion, this paper proposes strict separation of three planes:

a. **Intelligence Plane**
   Performs inference, prediction, estimation, and pattern recognition. Outputs are probabilistic and optimized, but inherently fallible.

b. **Decision Plane**
   Governs authorization, evaluates admissibility, enforces constraints, manages state, and determines whether action is permitted.

c. **Execution Plane**
   Carries out authorized actions with real-world consequences.

A core architectural rule follows:

Intelligence may inform decisions but may never authorize them.

7. Introducing DILLO

DILLO (Decision Intelligence & Logic Layer Orchestrator) is the formalization of the Decision Plane.

DILLO is:

- Model-agnostic
- Domain-portable

- Deterministic by design
- Explicitly state-aware

DILLO is not:

- A predictive model
- An agent
- A prompt wrapper
- A fine-tuning technique

Its role is to arbitrate between intelligence and execution, ensuring that action occurs only when permitted.

8. Core Functional Capabilities (Conceptual)

At a functional level, DILLO enables:

- Authority arbitration
- Constraint enforcement
- Context and state evaluation
- Explicit refusal paths
- Decision traceability and auditability

These capabilities are described intentionally without reference to specific algorithms or implementations.

9. Implications for AI Systems

The introduction of a decision governance layer has implications across domains:

- **Agentic AI**: prevents runaway autonomy
- **Enterprise systems**: enables accountable automation
- **Public sector**: supports lawful and ethical decision-making
- **Safety-critical domains:** aligns AI with established control principles

The central claim is unavoidable ***scaling intelligence without decision governance scales risk*****.**

10. Conclusion: Completing the AI Stack

The dominant challenges facing AI systems today are not rooted in insufficient intelligence, but in missing governance. Prediction alone cannot bear responsibility. Authority must be explicit, constrained, and auditable.

DILLO completes the AI stack by restoring a principle long understood in mature systems: ***intelligence informs, governance decides, execution acts***.

The future of AI will not be determined by better predictions, but by governed decisions.

FIGURE 1 — SYSTEM-LEVEL GOVERNANCE ARCHITECTURE

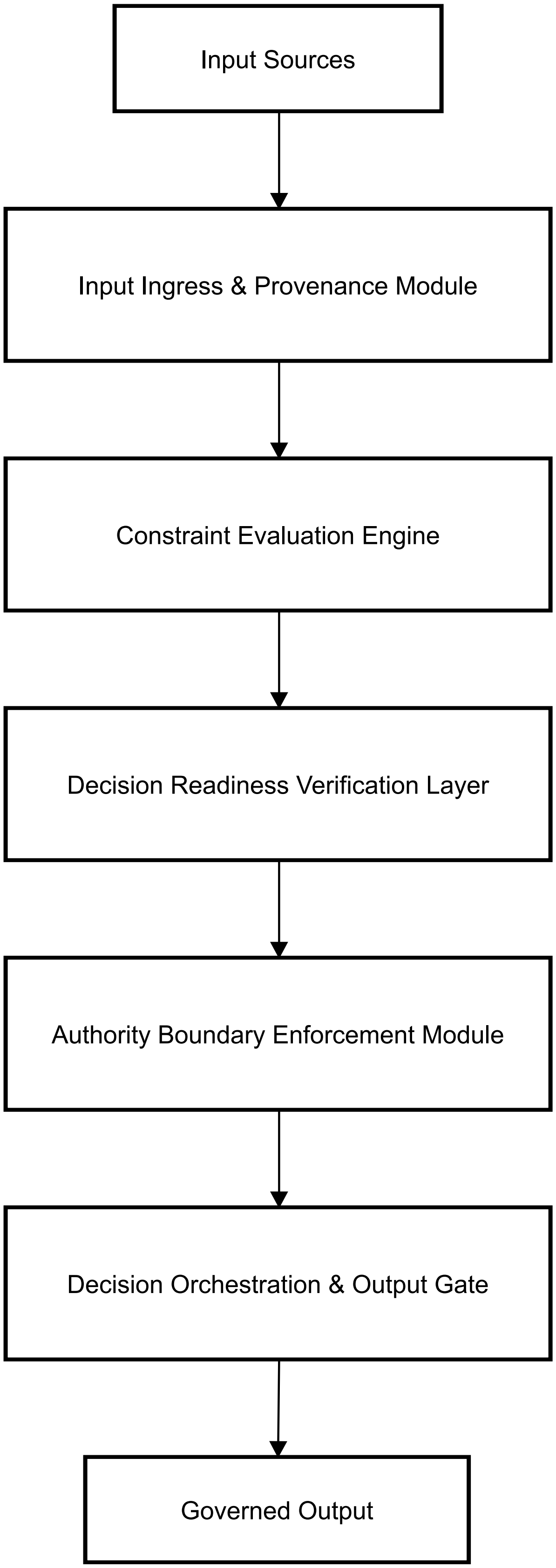
Input Sources
Input Ingress & Provenance Module
Constraint Evaluation Engine
Decision Readiness Verification Layer
Authority Boundary Enforcement Module
Decision Orchestration & Output Gate
Governed Output

Figure 1: System-Level Governance Architecture. Intelligence outputs are mediated by provenance validation, constraint evaluation, decision-readiness verification, and authority boundary enforcement before any action is authorized.

FIGURE 2 - DECISION FLOW WITH NON_DECISION STATE

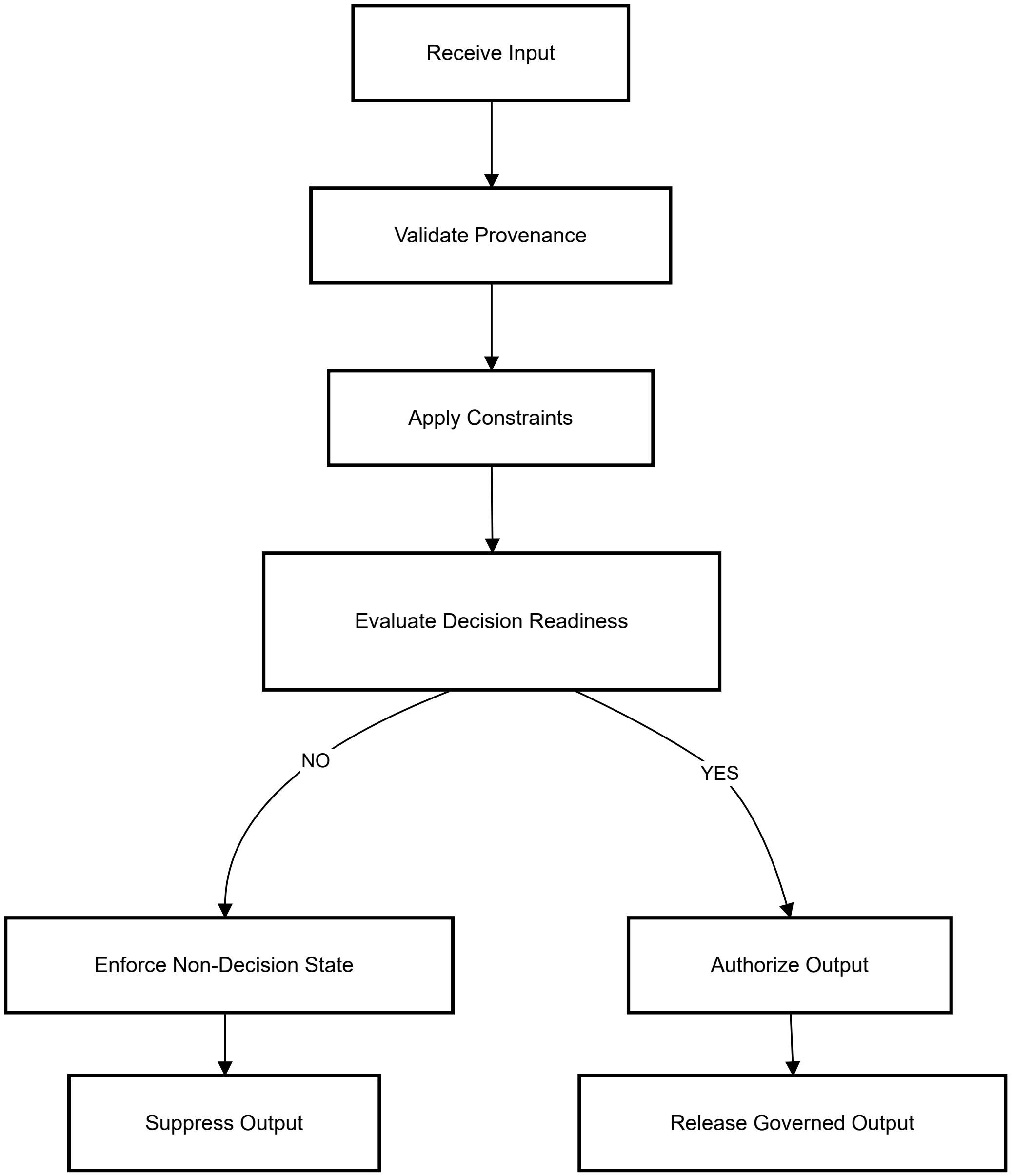

Figure 2: Decision Flow with Non-Decision State. Non-decision and output suppression are considered valid system outcomes when the authorization criteria are not met.

FIGURE 3 — HARDWARE / SOFTWARE EMBODIMENT

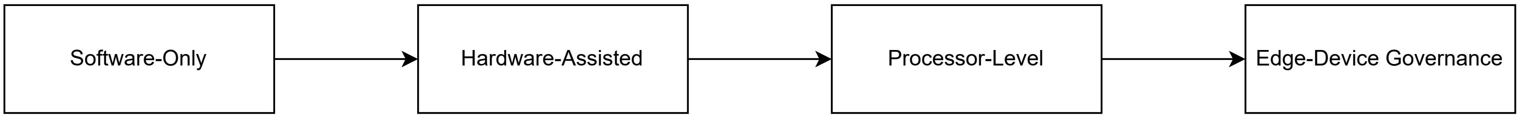

Figure 3: Hardware and Software Embodiments. The governance architecture may be implemented across software-only, hardware-assisted, processor-level, or edge-device configurations without altering its architectural role.

FIGURE 4—MULTI-MODEL GOVERNANCE

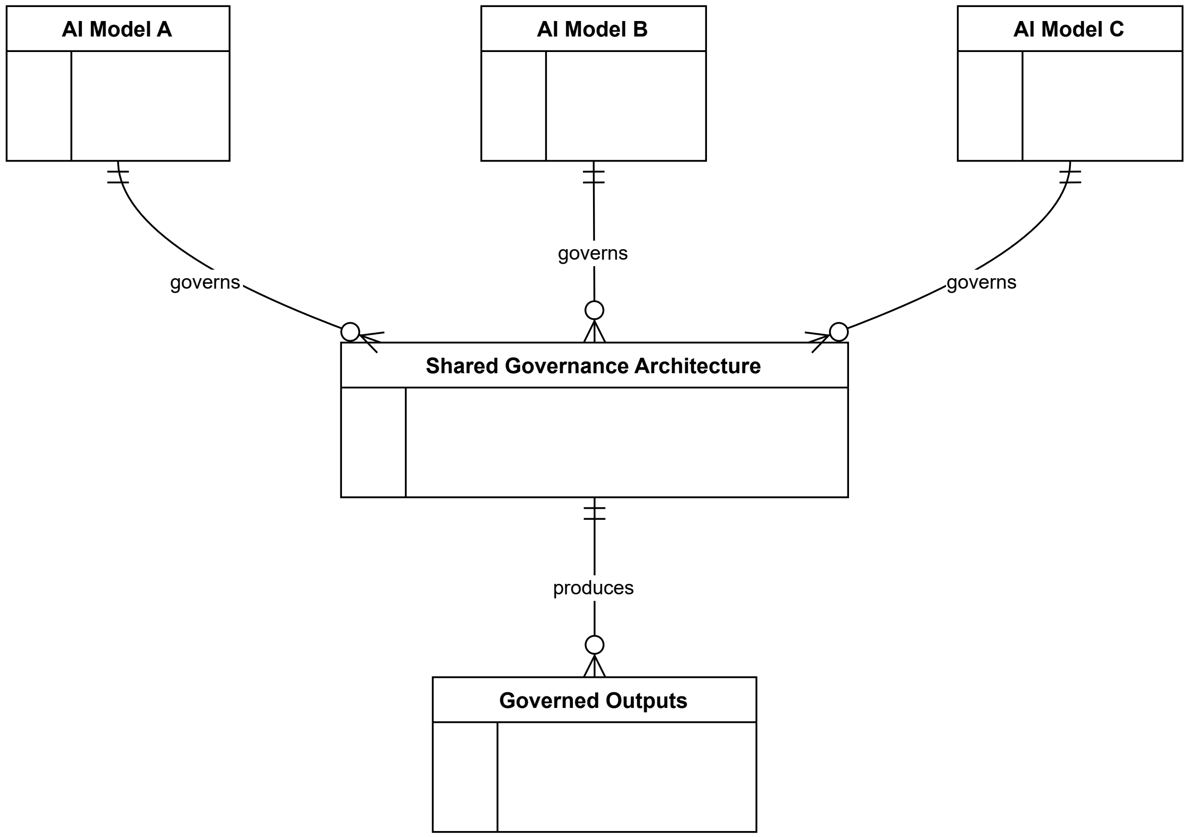

Figure 4: Multi-Model Governance. A shared governance architecture mediates outputs from multiple independent AI systems, ensuring consistent authorization across heterogeneous intelligence sources.