Toward Decision Behavior Governance: Governance Existence, Invocation, and Decision Formation

Abstract

Current AI governance and compliance frameworks—including ISO/IEC 42001—largely conceptualize governance targets in terms of systems, models, or development processes. However, in the era of foundation models deployed via commercial APIs, organizations no longer possess control over model internals, versioning, or training dynamics. This loss of model sovereignty severs the traditional proxy relationship between system compliance and behavioral accountability.

This paper argues that AI governance must be fundamentally reoriented toward **Decision Behavior**—the dynamic process through which an AI agent synthesizes premises, applies constraints, and forms outcomes with organizational impact. We propose **Decision Behavior Governance** as a conceptual framework that distinguishes `governance existence` from `governance invocation`, and decomposes governed behavior into three institutional elements: decision premises, execution manner, and governance evidence. By clarifying the conditions under which governance exists, becomes valid, and remains auditable, this paper provides a principled path for organizations to reclaim **Evidence Sovereignty** in the absence of model control.

1. Introduction: The Disappearance of the Governance Target

Generative AI systems increasingly participate in organizational decision-making not as deterministic tools, but as probabilistic agents. Traditional software governance assumes that behavior can be indirectly controlled through system design, code review, and deployment oversight. This assumption holds only when organizations retain authority over the decision engine itself.

In foundation model deployments, this assumption collapses. Organizations retain business sovereignty but lose **Model Sovereignty**: they cannot inspect internal logic or enforce version stability. The resulting governance vacuum is not caused by insufficient monitoring, but by a misidentified governance target. When auditors ask whether an AI system is "compliant," they often examine documentation—while remaining unable to answer: How was this specific decision formed?

2. Sovereignty Loss and the Governance Vacuum

The loss of model sovereignty introduces a structural asymmetry. While organizations remain responsible for AI-assisted decisions, they lack control over the engine's internal operation. Existing frameworks respond by intensifying procedural controls, implicitly treating the "System" as a proxy for "Behavior."

We argue that this proxy relationship is no longer valid. Governance must be reconstituted around **Evidence Sovereignty**—the ability to establish and analyze authoritative traces of how decisions were formed, independent of model opacity. This shift reframes governance from controlling the model to governing the institutional conditions of decision formation.

3. Defining the Governance Target: Decision Behavior

We formally define **Decision Behavior** as the non-deterministic trajectory through which an AI agent synthesizes premises, adheres to constraints, and forms an outcome with organizational impact.

Decision behavior is distinct from system output. Outputs are artifacts; behaviors are processes. In black-box contexts, behavior is the only governable object. Any framework that does not explicitly target decision behavior governs appearances rather than accountability.

4. Governance as an Institutional Fact

A central contribution of this paper is the distinction between **Governance Existence** and **Governance Invocation**.

- **Governance Existence:** The institutional fact that a decision domain is bounded by defined premises and constraints.
- **Governance Invocation:** Whether those constraints are actively applied during a specific instance.

Governance exists once premises and constraints are formally defined—regardless of whether a particular decision triggers them. This distinction exposes a critical flaw: a decision may be procedurally compliant yet behaviorally ungoverned. We therefore introduce **Governance Nullity**: a decision that cannot be attributed to an identifiable governance regime, regardless of administrative compliance.

5. The Three Elements of Decision Behavior Governance

For decision behavior to be governable, three institutional elements must be present:

1. **Decision Premise:** The authorized knowledge and legal constraints defining the legitimacy boundary.
2. **Execution Manner:** Constraints governing how premises are prioritized during reasoning.
3. **Governance Evidence:** The traceable residue (constraints references, execution traces) produced by formation.

Evidence is not a condition for existence, but a condition for **validity and enforceability**.

6. Why Existing Frameworks Fall Short

Most current practices, including many ISO/IEC 42001 implementations, focus on whether policies were documented or roles assigned. They establish administrative order but fail to define behavioral boundaries. This reveals a systemic gap: decisions that are administratively compliant yet behaviorally "Nullities."

7. Implications for ISO/IEC 42001 Practice

This paper proposes a reframing of audit practices. Instead of asking "Was a policy documented?", auditors must ask: **"Was this decision formed within an attributable governance regime?"** By anchoring governance to decision formation, ISO/IEC 42001 can transition from static compliance to dynamic accountability.

8. Evidence Sovereignty and Institutional Learning

Evidence enables **Evolution**. Aggregated governance evidence allows organizations to detect **Behavioral Drift**—where the model's interpretation of constraints shifts despite no change in the premise. Evidence enables learning, while legitimacy arises from the institutional boundaries.

9. Conclusion

Accountability emerges not from the transparency of models, but from the traceability of decision formation. Decision Behavior Governance transforms AI governance from a documentation exercise into an institutional fact. Any regime that cannot attribute decisions to a bounded behavioral framework fails at governance itself.

Without an explicit governance target at the level of decision behavior, AI governance regimes remain administratively complete but institutionally void.+