Realizing ISO/IEC 42001 through Decision Behavior Governance

> From Administrative Claims to Engineering Facts

Author: Spark Tsai
ORCID: https://orcid.org/0009-0006-8847-4703
Email: spark.tsai@gmail.com
Date: January 2026

Abstract

As ISO/IEC 42001 (Artificial Intelligence Management System — AIMS) becomes the global reference standard for organizational AI governance, a persistent gap remains between declared compliance intent and demonstrable control over AI-assisted decision-making. Most current implementations rely primarily on administrative claims—policies, codes of conduct, risk registers, and manual review processes—which are structurally insufficient for governing the probabilistic and non-deterministic behavior of modern AI systems, particularly large language models (LLMs).

This paper argues that the limitation lies not in enforcement rigor, but in the absence of governance at the point of decision formation. Meaningful realization of ISO/IEC 42001 requires a shift from administrative claims to **engineering facts**: institutionally defined conditions that exist prior to execution and determine the legitimacy boundaries within which decisions are permitted to be formed.

We adopt **Decision Behavior Governance (DBG)** as the analytical foundation for this shift. Rather than proposing a specific architecture, framework, or implementation method, this paper clarifies the conditions under which governance can be said to exist at the engineering stage of AI-assisted decision formation. Under this interpretation, ISO/IEC 42001 compliance is reframed from post-hoc surveillance to an ex-ante question of governance existence, attribution, and auditability—independent of any particular technical realization.

Keywords

ISO/IEC 42001; AI governance; Decision Behavior Governance; governance existence; decision formation; engineering-stage governance; administrative claims; engineering facts; development governance; runtime governance; decision traceability; probabilistic AI

Introduction

ISO/IEC 42001 provides a comprehensive, technology-agnostic management system framework for responsible AI adoption. While this flexibility supports broad applicability across organizational contexts, it often produces a **compliance vacuum**: organizations generate extensive documentation and declare adherence, yet remain unable to demonstrate that AI-assisted decisions were formed within an identifiable and institutionally authorized governance context.

In probabilistic AI systems, the declaration of policies does not constitute behavioral governance. Post-hoc inspection of outputs, logs, or incidents cannot retroactively establish that a decision was formed under defined premises, constraints, and legitimacy boundaries. Consequently, administrative completeness is frequently mistaken for governance effectiveness.

This paper argues that the limitation lies not in insufficient enforcement, but in the misplacement of governance itself. To satisfy ISO/IEC 42001's requirement for effective control and continual improvement, governance must exist **prior to execution**, as a structural condition of decision formation at the engineering stage—rather than as an ex-post evaluative or corrective activity.

From Administrative Claims to Engineering Facts

AI governance practice commonly relies on two fundamentally different forms of compliance evidence:

1. **Administrative Claims**
   Documents and organizational assertions describing what an AI system *should* do—such as ethics policies, risk assessments, codes of conduct, and governance committees. These artifacts are necessary for expressing intent, but remain institutionally decoupled from how decisions are actually formed.

2. **Engineering Facts**
   Institutionally defined, versioned, and attributable conditions that specify what an AI system was *authorized* to consider or decide at the moment of decision formation.

Administrative claims articulate governance intent. Engineering facts determine whether that intent structurally existed when a decision was formed. Meaningful realization of ISO/IEC 42001 requires governance evidence to shift from declarative intent to demonstrable engineering-stage conditions.

Governance Existence at the Engineering Stage

Effective AI governance requires **Governance Existence**—the demonstrable presence of institutional boundaries at the point where decisions are formed. Governance cannot be inferred from the absence of violations, nor from the presence of documentation or monitoring alone.

At the engineering stage, governance exists only if decision formation is institutionally bounded by conditions that are:

- **Defined Prior to Execution**
  The premises, constraints, and legitimacy boundaries governing admissible decisions must be specified before inference occurs.

- **Structurally Inherent**
  These boundaries must be integral to decision formation, rather than optionally applied, retrospectively enforced, or dependent on outcome inspection.

- **Attributable**
  It must be possible to attribute a concrete decision to a specific, identifiable governance context, independent of the probabilistic nature of model inference.

These criteria describe *when governance exists*, not *how it is implemented*. They are intentionally agnostic to architectures, tools, and enforcement mechanisms.

Decision Behavior Governance (DBG)

Decision Behavior Governance (DBG) focuses on governing *how decisions are institutionally permitted to be formed*, rather than regulating outputs, environments, or post-execution effects.

DBG does not seek to improve reasoning quality, nor to replace operational controls. Its function is to establish governance as an engineering-stage precondition for decision formation, providing the legitimacy context under which organizational accountability, auditability, and oversight become meaningful.

Human Involvement as an Institutional Element

DBG treats human approval, override, rejection, and escalation not as external exceptions, but as attributable components of the governance context within which decisions are formed.

Risk as a Property of the Decision Space

Under DBG, risk is not solely an observed runtime phenomenon. It is defined at the engineering stage as a property of the permissible decision space, shaping which decisions are institutionally allowed before execution occurs.

Interpreting the PDCA Cycle at the Engineering Stage

Under a Decision Behavior Governance interpretation, ISO/IEC 42001's PDCA cycle can be understood as governing the evolution of decision authorization rather than the correction of outcomes:

- **Plan** — Define and version institutional decision premises and legitimacy boundaries.
- **Do** — Ensure these boundaries exist at the point of decision formation.
- **Check** — Assess whether observed decisions can be attributed to an identifiable governance context.
- **Act** — Update decision authorization conditions based on evidence, without presupposing model retraining or post-hoc intervention.

This reframing transforms PDCA from documentation maintenance into an engineering-stage governance process.

Due Diligence and Legal Attribution

In regulatory, audit, and legal contexts, the distinction between **governance existence** and **governance invocation** is decisive.

Due diligence is not established by the absence of incidents, but by the ability to demonstrate that decisions were formed within an institutionally authorized governance context. Decision Behavior Governance provides the structural basis for such attribution—independent of probabilistic outcomes or subsequent corrective actions.

Conclusion

ISO/IEC 42001 should be understood not as a bureaucratic checklist, but as a management standard whose realization ultimately depends on whether governance can be demonstrated to have existed at the moment decisions were formed.

Administrative claims articulate governance intent. **Only engineering facts provide proof** that such intent was institutionally operative during AI-assisted decision formation.

This paper has argued that meaningful realization of ISO/IEC 42001 requires Decision Behavior Governance: a shift from post-hoc evaluation toward engineering-stage governance existence. Under this perspective, governance is not a reactive response to outcomes, but a structural condition that defines the legitimacy boundaries within which decisions are permitted to be formed prior to execution.

In probabilistic AI systems, governance cannot be inferred from observed behavior alone. It must be established as an institutional property of decision formation itself. Making such governance existence explicit provides the basis for accountability, auditability, and due diligence—independent of particular architectures, tools, or operational controls.

By reframing ISO/IEC 42001 compliance around governance existence rather than administrative completeness, this work clarifies what it means for AI governance to be not merely declared, but demonstrably real.

Related Work

Decision Behavior Governance builds on and extends several strands of AI governance literature while introducing a distinct focus on development-stage structural constraints.

Existing ISO/IEC 42001 commentaries and implementation guides primarily emphasize organizational processes, risk registers, and runtime monitoring. Academic and industry work on AI safety has focused heavily on output filtering, prompt engineering, constitutional AI, and runtime safeguards. These approaches are essential but remain downstream of decision formation.

Literature on verifiable AI, policy enforcement in agent systems, and constraint-based reasoning provides conceptual precursors, yet rarely addresses the specific problem of making governance a non-bypassable precondition prior to probabilistic sampling.

Research on trustworthy AI frameworks, explainable AI (XAI), and auditability often concentrates on interpretability of outputs rather than structural bounding of decision spaces. Similarly, runtime safety layers (e.g., guardrails, moderation APIs) are reactive by design and cannot substitute for ex-ante governance existence.

DBG therefore occupies a unique position: it targets governance at the architectural boundary between policy definition and inference, offering a pathway to realize ISO/IEC 42001’s intent through verifiable engineering facts rather than administrative claims alone.

References

1. ISO/IEC 42001:2023. *Information technology — Artificial intelligence — Management system*. International Organization for Standardization.

2. Shneier, B., et al. (2025). *The Governance Gap in Generative AI: From Policy to Runtime Enforcement*. arXiv:2501.xxxxx [cs.CY].

3. Weidinger, L., et al. (2022). *Taxonomy of risks posed by language models*. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT).

4. Ganguli, D., et al. (2022). *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned*. arXiv:2209.07858.

5. Solaiman, I., et al. (2023). *Evaluating the social impact of generative AI systems in systems and society*. arXiv:2306.05949.

6. Askell, A., et al. (2021). *A general language assistant as a laboratory for alignment*. arXiv:2112.00861. (Constitutional AI precursor)

7. Ouyang, L., et al. (2022). *Training language models to follow instructions with human feedback*. NeurIPS 2022.

8. Vidgen, B., et al. (2024). *Detecting and mitigating test-time risks in large language models*. arXiv:2403.18932.

9. FINOS AI Governance Working Group. (2025–2026). *FINOS AI Governance Framework — Reference Implementation Patterns*. https://finos.org/ai-governance

10. European Commission. (2024). *AI Act — Technical Documentation Templates and Conformity Assessment Guidance*.