Four Stages of AI Governance: A Stage-Based Framing of Contemporary Practices

Abstract

As AI systems transition from experimental tools to integral components of organizational decision-making, discussions of AI governance have expanded rapidly across technical, operational, and administrative domains. However, contemporary practices often address governance concerns at different points in the AI lifecycle without clearly distinguishing their scope or limitations.

This paper proposes a descriptive framework that identifies **Four Stages of AI Governance**: **Model**, **Development**, **Deployment**, and **Operation**. Rather than assessing maturity or effectiveness, this stage-based framing clarifies *where* governance mechanisms are applied and *what* aspects of AI systems they govern at each stage. In particular, the framework highlights the **Development Stage** as the phase in which **decision formation structures and constraints** are established—an area that has received comparatively less explicit attention amid increasing focus on deployment controls and runtime monitoring.

By articulating these stages, this paper provides a neutral reference model for situating existing governance practices and emerging proposals within the broader AI lifecycle, without prescribing specific governance methodologies or standards.

1. Introduction: Mapping Governance Across the AI Lifecycle

The adoption of foundation models and AI-enabled systems has accelerated governance initiatives across industries. Organizations now employ a range of practices, including model evaluation, policy documentation, deployment controls, and runtime monitoring. While these measures address legitimate concerns, they are often discussed without a shared vocabulary for distinguishing the stages at which governance is exercised.

This paper addresses this gap by presenting a stage-based classification of AI governance practices. The goal is not to prescribe a specific governance approach, but to clarify the *lifecycle position* and *governance object* associated with commonly observed practices.

2. Stage 1: Model Governance

Model governance concerns the properties and behavior of AI models themselves, independent of a specific organizational context.

- **Governance Object:** Model capabilities and general behavioral tendencies.
- **Common Focus Areas:** Safety characteristics such as bias, robustness, toxicity, and alignment with broad usage expectations.
- **Typical Practices:** Pre-training alignment, fine-tuning, red-teaming, and benchmark-based evaluation.
- **Scope Boundary:** Model governance shapes general behavior but does not directly account for organization-specific decision accountability or contextual constraints.

This stage is typically addressed by model providers and research organizations prior to downstream integration.

3. Stage 2: Development Governance

Development governance addresses how AI-assisted decisions are *structured* prior to execution.

- **Governance Object:** The structure of decision logic, including permitted decision types, applicable constraints, and validation conditions.
- **Temporal Characteristic: Ex-ante**, occurring before execution or deployment.
- **Typical Practices:** Defining decision rules, specifying constraints, documenting assumptions, and designing decision workflows.
- **Decision Formation Focus:** At this stage, governance is concerned with how decisions will be formed rather than how they are later deployed or observed.

In many organizations, this stage is embedded within software design, system integration, or policy translation activities.

4. Stage 3: Deployment Governance

Deployment governance governs which decision structures or system configurations are authorized to operate in specific environments.

- **Governance Object:** System configuration and version selection.
- **Primary Concerns:** Risk exposure, environment suitability, and rollback capability.
- **Typical Practices:** Approval workflows, configuration gating, environment separation, and release management.

- **Functional Role:** Deployment governance manages operational risk but generally does not alter the internal structure of decision logic.

This stage serves as a control point between development artifacts and live operation.

5. Stage 4: Operation (Runtime) Governance

Operation governance focuses on system behavior during execution.

- **Governance Object:** Executed behavior and observable outputs.
- **Primary Concerns:** Compliance with existing policies, detection of anomalies, and evidence collection.
- **Typical Practices:** Runtime guardrails, monitoring systems, logging, and post-hoc audits.
- **Descriptive Limitation:** Since decisions are already formed at execution time, governance at this stage is primarily observational or reactive.

Operational governance is commonly emphasized due to its visibility and measurability.

6. Comparative Summary of Governance Stages

Stage	Governance Object	Primary Focus	Temporal Nature
Model	Model behavior	General safety tendencies	Pre-integration
Development	Decision structure	Logic and constraints	**Ex-ante**
Deployment	Configuration & access	Risk exposure & control	Transitional
Operation	Executed behavior	Monitoring & evidence	Runtime

7. Conclusion: A Descriptive Reference for Governance Placement

By distinguishing AI governance practices according to lifecycle stage, this framework provides a reference for analyzing where governance claims are situated and what they can reasonably achieve. Governance mechanisms applied at different stages operate on distinct objects and exhibit inherently different limitations.

While many contemporary discussions emphasize deployment and runtime controls due to their visibility and measurability, this stage-based view highlights that governance objectives differ fundamentally depending on *when* and *where* they are applied. In particular, governance mechanisms introduced at later stages cannot fully substitute for decision structures and constraints that are established during development.

This paper does not advocate a specific governance methodology or evaluate governance effectiveness. Instead, it offers a descriptive structure for locating existing practices and for situating subsequent governance proposals within the broader context of AI system development and operation.