Tavri Inc. — tavri.com

Inference Authority: Controlled Governance of Authoritative Output

A governance vocabulary for safety-critical AI: designate which model or compute pathway may produce authoritative output, keep candidate models non-authoritative during shadow-mode validation, evaluate divergence validation against explicit alignment conditions, and transfer authority only within a controlled transition window governed by failover governance.

Published: January 18, 2026

Core Technical Terminology

The following terms are used throughout this framework to describe architectural roles and governance mechanisms in safety-critical AI systems.

Inference Authority
The governance designation, within a safety-critical AI system, of which model, compute unit, or inference pathway is permitted to produce authoritative output that may influence operational behavior.

Authority Gating
A governance control layer that enforces the authoritative/non-authoritative boundary by allowing, suppressing, constraining, or redirecting outputs based on policy triggers and alignment conditions.

Divergence Validation
A structured validation process that compares the authoritative pathway against a candidate model under functionally equivalent inputs to determine whether observed output differences remain within defined limits over a validation interval.

Candidate Model
A model executed in a non-authoritative state for evaluation relative to the authoritative pathway, with authority withheld by authority gating.

Alignment Conditions
Explicit, testable criteria that determine eligibility for authority transfer based on divergence limits, stability requirements, and policy constraints evaluated across a validation interval.

Controlled Transition
A managed authority handover process that sequences authority assignment, revocation, and optional output constraints to reduce discontinuity and enforce safety policy.

Failover Governance
The policy and control logic that governs response to failure, drift, or anomaly including authority reassignment and optional output gating to suppress unsafe outputs.

Notes: Definitions clarify governance expectations and do not prescribe a single implementation. Timing, thresholds, and handover mechanics are environment dependent.

Scope

This Conceptual Framework defines terminology for controlled governance of inference authority in regulated and safety-critical environments. It is intended to support architecture planning, policy discussions, and auditability requirements. Implementation details such as latency budgets, thresholds, and safety policies are environment dependent.

Executive Summary

In safety-critical AI, the question is not only whether a model is accurate, but which pathway is allowed to control downstream behavior at any given time. Many systems run multiple models or compute pathways in parallel but treat authority as an implicit engineering detail rather than an explicit, auditable state.

Inference Authority names and structures that missing layer: a system-level designation of which model or pathway may produce authoritative output, enforced by authority gating. Candidate pathways execute concurrently in a non-authoritative state for divergence validation, and authority is transferred only when explicit alignment conditions are satisfied within a controlled transition.

1. Why Inference Authority is a Distinct Governance Concept

Traditional control systems assume deterministic control paths. Safety-critical AI introduces probabilistic outputs, dynamic model updates, and multi-path inference. When outputs can influence physical systems, it becomes necessary to govern which output is allowed to act.

2. Definition: Inference Authority

Inference Authority is the system designation of which model, compute unit, or inference pathway is permitted to produce authoritative output that may influence operational processes such as control, routing, dispatch, navigation, or decision-making.

Authoritative Pathway

The inference source currently permitted to issue operationally binding outputs.

Non-Authoritative Pathway

A parallel inference pathway that may execute under equivalent inputs for validation but is prevented from issuing authoritative outputs by authority gating.

3. Roles in a Controlled Governance Framework

- An authoritative inference pathway
- Non-authoritative candidate pathways executing in shadow mode
- A divergence validation mechanism
- Explicit alignment conditions and gating policies
- Failover governance for authority assignment and revocation
- Output gating to suppress unsafe outputs

4. Shadow Mode Validation and Alignment

Shadow mode validation runs candidate models in parallel using equivalent inputs. Divergence validation compares outputs over a validation interval and determines eligibility for authority transfer based on alignment conditions.

5. Controlled Transition Window

Transferring inference authority is a controlled process executed within a transition window designed to reduce discontinuity and instability. Authority gating may sequence handover or constrain outputs until policy criteria are satisfied.

6. Auditability and Model Provenance

Authority transitions should be explainable. Governance frameworks should log why authority changed including policy triggers, divergence metrics, alignment outcomes, and system state, supported by tamper-evident logs.

7. Relevance Across Domains

The vocabulary applies wherever AI outputs influence operational processes including infrastructure, industrial automation, robotics, transportation, large-scale computing, and defense systems.

Conclusion

Inference Authority provides a concise vocabulary for controlled governance of authoritative output in safety-critical AI systems. By requiring explicit authority states, divergence validation, controlled transitions, and governance-grade auditability, it supports engineering clarity and regulator-ready safety framing.