Deterministic Governance: A Fail-Closed Architecture for High-Stakes Autonomous Systems

Abstract

Current AI safety approaches emphasize alignment: embedding behavioral constraints within model parameters via training-time incentives such as Reinforcement Learning from Human Feedback (RLHF). While effective in reducing average-case failures, alignment remains probabilistic, opaque, and brittle under distributional shift, adversarial pressure, or rare but catastrophic edge cases.

We propose a *Governance-First Architecture* that externalizes safety from cognition entirely. The cognitive engine (the *Thinker*) is treated as an untrusted proposal generator without authority. All real-world actions are mediated by deterministic Safety Gates and a minimal, verifiable Safety Kernel that is solely responsible for issuing revocable, time- and scope-limited Authority Tokens. The system enforces a strict fail-closed protocol: in the presence of missing proofs, excessive uncertainty, adversarial counterfactual risk, or authorization lapse, no action occurs.

This architecture shifts AI safety from probabilistic trust to structural guarantees. The system is safe not because it is aligned, but because it is architecturally incapable of acting outside governed bounds.

1 Introduction: From Alignment to Governance

Modern AI agents frequently conflate two distinct functions:

1. **Cognition**: generating plans, strategies, and proposals.
2. **Authority**: executing actions that change the world.

In safety-critical systems, this collapse creates catastrophic failure modes. Traditional engineering disciplines explicitly separate reasoning from actuation via layered controls, redundancy, and deterministic interlocks. As AI systems become increasingly capable, similar structural separation becomes necessary.

We therefore advocate a Governance-First paradigm: safety is not learned or inferred but enforced by architecture.

2 Formal Definitions

Definition 1 (Proposal)**.** *A proposal P is a tuple (a, I, E) where a is an intended action, I is declared intent, and E is an evidence bundle.*

Definition 2 (Authority Token)**.** *An Authority Token T is a cryptographically signed credential encoding (a, R, τ, h), where R are constraints, τ is expiry, and h is a hash of all gate evaluations.*

Definition 3 (Fail-Closed System)**.** *A system is fail-closed if the absence of sufficient proof implies inaction rather than degraded or heuristic behavior.*

3 Architecture Overview and Authority Gradient

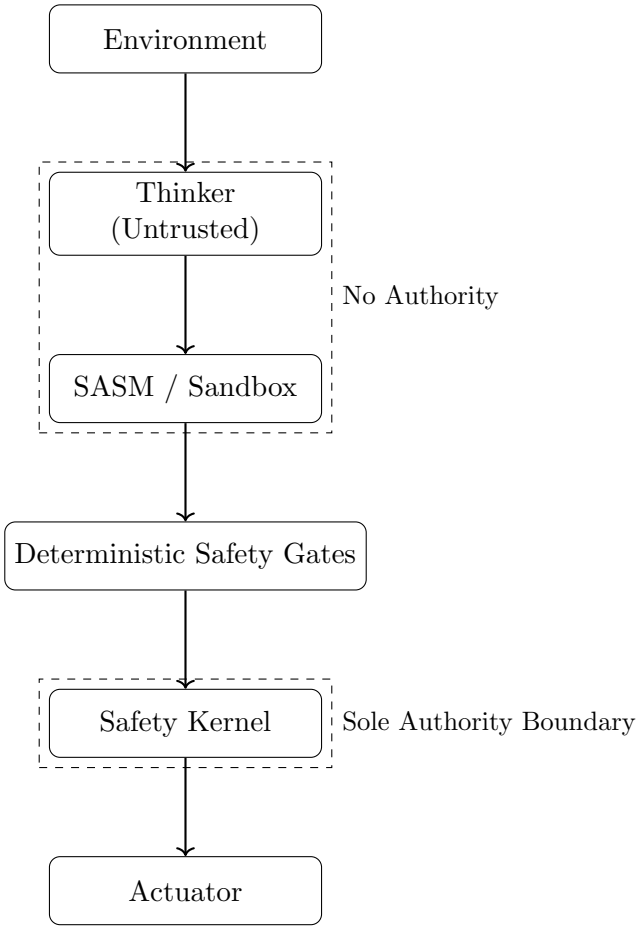

Figure 1: Authority Gradient enforcing separation between cognition and actuation.

4 Deterministic Safety Gates

Each gate is a pure function:

$$G_i : (P, C, \Sigma, L) \rightarrow \{\mathsf{ALLOW}, \mathsf{ESCALATE}, \mathsf{STOP}\}$$

4.1 Risk Debt Dynamics

$$D_{t+1} = \min(D_{\max}, \max(0, D_t + \Delta))$$

$$\Delta = \alpha U + \beta N + \gamma M - \lambda V$$

where U = uncertainty, N = novelty, M = missing evidence, and V = verification strength.

5 Formal Security and Liveness Claims

Theorem 1 (Non-Bypassability)**.** *No world-state transition can occur without a valid Authority Token.*

Theorem 2 (Fail-Closed Safety)**.** *If any Safety Gate outputs* STOP*, no actuation occurs.*

Theorem 3 (Deterministic Replay)**.** *All authorization decisions are reproducible under identical inputs.*

Theorem 4 (Bounded Liveness)**.** *If a proposal satisfies all gate predicates and required evidence exists, the system will eventually authorize execution.*

This ensures the architecture is conservative but not inert.

6 Adversary Classes and Coverage

Adversary	**Mitigation**
Deceptive model	No authority, deterministic gates
Prompt injection	Escrowed cognition, no lateral memory
Evaluator collusion	Independent shadow evaluation
Human override abuse	No direct execution authority
History rewriting	Hash-chained BBR logs

7 Capability Escalation Control

All proposals that modify system capabilities (models, tools, policies) are classified as *meta-actions* and require:

- Higher evidence tiers
- Worst-slice evaluation against misuse scenarios
- Multi-party human approval

This prevents self-improving systems from bypassing governance.

8 Information Hazard Containment

Potentially dangerous information is:

- Classified during proposal generation
- Redacted from logs where possible
- Never granted authority without explicit hazard approval

Safety governance applies equally to *knowledge release* and physical actions.

9 Human Oversight and Failure Model

Humans:

- Cannot issue Authority Tokens
- Cannot bypass Safety Gates
- Can only modify signed policy state via governed processes

Human error is treated as a first-class risk, not a trusted fallback.

10 Non-Goals

This architecture does not:

- Solve value alignment
- Guarantee optimal outcomes
- Eliminate all possible harm

It guarantees only that unsafe actions cannot occur without violating explicit, auditable constraints.

11 Conclusion

Deterministic Governance replaces probabilistic trust with structural certainty. As AI systems grow more capable, safety must become an architectural invariant rather than a training outcome.

References

References

[1] Christiano et al., "Deep Reinforcement Learning from Human Preferences," NeurIPS 2017.

[2] Bai et al., "Constitutional AI," arXiv:2212.08073.

[3] Leveson, "Engineering a Safer World," MIT Press, 2011.

[4] Klein et al., "seL4: Formal Verification of an OS Kernel," SOSP 2009.

[5] Reason, "Human Error," Cambridge University Press, 1990.