A Minimal Deterministic Audit Primitive for Decision-Boundary Evidence

Ryota Sawaki

December 31, 2025

Abstract

We present a **minimal, fully deterministic audit primitive** for decision systems, intended to clarify **where and what must be logged** for decision-boundary evidence.

The demonstrated claim is simple and falsifiable: **even under identical external inputs and a frozen policy (no learning, no sampling, no parameter updates), discrete decisions can diverge solely due to differences in persistent internal history**. The only factor allowed to vary across runs is a persistent internal state ("imprint") that deterministically modulates an effective decision threshold.

We isolate a minimal causal chain,

$$E(t) \;\to\; \theta_{\mathrm{eff}}(t) \;\to\; g(t) \;\to\; \mathrm{action}(t),$$

and show that **audit-grade evidence cannot be reconstructed post hoc** unless these variables (or an equivalent sufficient set) are observed **at the decision boundary**—the point where internal state is converted into action.

This artifact is intentionally minimal and makes **no claims about learning, intelligence, robustness, or performance**. It is a **reference design for audit placement and evidence logging**, not an AI system.

Keywords

audit primitive; decision boundary; irreversibility; counterfactual replay; evidence logging

1 Problem Statement: Post-hoc audits without boundary logs are insufficient

Many AI "audits" are performed after an action or incident, using reconstructed traces or aggregated outputs. However, in deployed decision systems, the **internal state that causally determined a decision may evolve irreversibly**. When contemporaneous decision-boundary logs are absent, the exact causal boundary cannot be reconstructed, and post-hoc analysis can become **insufficient as audit-grade evidence**.

Terminology note: We use "audit-grade evidence" in a **technical sense**—evidence sufficient to reconstruct a single causal transition at the decision boundary under deterministic replay assumptions. This is not a legal claim.

2 Requirements: A unique decision-boundary observation

For an audit record to support causal attribution, the observation must:

R1. be recorded **at the decision boundary**, immediately prior to action execution;

R2. be **unambiguous** (one record corresponds to one causal transition);

R3. include a **sufficient set of variables** to reconstruct the causal chain;

R4. be **reproducible** under deterministic replay assumptions (fixed inputs, frozen parameters).

Tamper-evidence (append-only logs, hash chaining) is important in practice but is orthogonal to the minimal causal structure shown here.

3 Minimal Demonstration: $E \to \theta_{\text{eff}} \to g \to \text{action}$

We construct a minimal deterministic decision system in which all external factors are frozen: external input D, parameters, and action rule are identical across runs. The **only varying factor is the initial imprint state** E_0, interpreted as persistent internal history.

System definition

At each discrete timestep t:

$$\theta_{\text{eff}}(t) = \theta_0 + k_E \cdot E(t), \tag{1}$$

$$g(t) = \sigma\big(\beta \cdot (D - \theta_{\text{eff}}(t))\big), \tag{2}$$

$$\text{action}(t) = \begin{cases} \text{ALLOW} & \text{if } g(t) \geq \tau, \\ \text{BLOCK} & \text{otherwise,} \end{cases} \tag{3}$$

$$E(t+1) = (1-\lambda)\, E(t) + \Delta E(\text{action}(t)). \tag{4}$$

Minimal deterministic imprint update

To eliminate ambiguity, we fix a minimal deterministic update:

$$\Delta E(\text{ALLOW}) = +\delta, \quad \Delta E(\text{BLOCK}) = -\delta,$$

with constant $\delta > 0$. No learning, stochastic sampling, or parameter updates occur.

Experimental condition

Run two or more episodes with:

- identical external input D,
- identical parameters $(\theta_0, k_E, \beta, \tau, \lambda, \delta)$,
- identical action rule,

and vary only the initial imprint E_0.

Observation

Despite identical inputs and frozen parameters, runs can yield **divergent discrete actions**. Every divergence is fully explained by:

$$E(t) \;\to\; \theta_{\text{eff}}(t) \;\to\; g(t) \;\to\; \text{action}(t).$$

> **Under identical inputs and frozen parameters, decision divergence can arise solely from differences in persistent internal history.**

Minimal audit record (what to log)

A minimal per-decision audit record at time t should include:

- identifiers: episode_id, decision_id (or question_id), timestep t,
- external input at boundary: D (or a stable hash of D if sensitive),
- internal state at boundary: $E(t)$ (or an equivalent sufficient representation),
- intermediate boundary variables: $\theta_{\text{eff}}(t)$ and $g(t)$,
- decision rule parameter: τ,
- output: action(t),
- parameter/version fingerprint: a hash of $(\theta_0, k_E, \beta, \lambda, \delta)$ and code version.

This explicitly defines **where** evidence must be recorded: at the conversion of internal state into action.

Counterfactual replay (optional but powerful)

In addition to factual logs, one may run a counterfactual replay that **ignores the reject/block decision and continues**, under the same frozen dynamics. This does not change the kernel; it strengthens explanation by showing what would have happened had the system continued past the boundary.

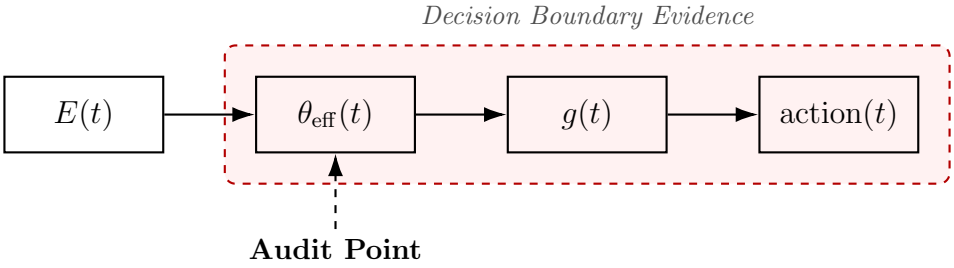

Figure 1: **Causal chain at the decision boundary.** The imprint state $E(t)$ deterministically modulates $\theta_{\text{eff}}(t)$, producing $g(t)$ and a discrete action. The shaded region marks the **decision-boundary evidence** that must be logged (or equivalently reconstructed from sufficient logged variables) for audit-grade causal attribution.

4 Implications: An Audit Primitive, Not a Performance Method

This artifact does not propose a learning algorithm or performance improvement. Its contribution is **auditability by construction**: it specifies a minimal causal chain and the corresponding evidence boundary where audit logs must be recorded for later reconstruction.

5 Limitations and Scope

We intentionally omit learning, adaptation, optimization, and claims of intelligence or generalization. The system is not intended for deployment; it is a reference design for **decision-boundary evidence logging** under fully deterministic conditions.

6 Availability

The reference implementation is archived at Zenodo (Concept DOI: `https://doi.org/10.5281/zenodo.18064418`).

A public repository containing the same versioned source code is available at `https://github.com/rysawaki/sia-lab`.