A Minimal Deterministic Audit Primitive for Pre-Deployment Decision Systems

Ryota Sawaki

December 29, 2025

Abstract

We present a **minimal, fully deterministic audit primitive** for decision systems intended for **pre-deployment evaluation**.

The core claim demonstrated is simple and falsifiable: **even with identical external inputs, identical random seeds, and a frozen policy, discrete decisions can diverge solely due to differences in persistent internal history**. No learning, stochasticity, policy updates, or environment changes are involved.

We construct a minimal audit artifact in which the only varying factor across runs is an internal state ("imprint") that deterministically modulates an effective decision threshold. All decision outcomes are causally explained step by step via an explicit chain: internal state → effective threshold → gate value → action. Every variable in this chain is observable and can be recorded at the decision boundary.

This setup isolates internal history as a sufficient cause of decision divergence and, crucially, **identifies the unique location where audit evidence must be recorded to remain admissible after deployment**. The artifact does not propose a learning algorithm, performance improvement, or behavioral model; it instead serves as a **reference design for auditability**, showing why post-hoc analysis is insufficient once internal state has irreversibly evolved.

We argue that effective audits of decision systems must operate concurrently with decision-making, prior to deployment, and at the point where internal state is converted into action. The provided implementation is intentionally minimal, deterministic, and reproducible, and is intended as an **audit primitive rather than an AI system**.

1 Problem Statement: Post-hoc audits are not admissible evidence

Most existing AI audits are conducted **post hoc**, after a decision has already been made or an incident has occurred. However, once a decision is executed, the **internal state that causally determined that decision has irreversibly evolved**, making it impossible to reconstruct the exact decision boundary ex post. Consequently, post-hoc analysis cannot constitute **admissible evidence** for explaining or attributing responsibility in deployed decision systems.

2 Requirements: Audits must operate immediately prior to decision execution

For an audit to be admissible, several conditions **must** be satisfied. First, the audit **must operate concurrently with the system prior to any incident**, rather than being initiated post hoc. Second, it **must observe the internal state immediately before action selection**, at the precise point where a decision boundary is evaluated. Finally, the observation point **must be unique and unambiguous**, so that the recorded variables correspond to a single causal transition rather than an aggregate or inferred state.

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

We construct a minimal deterministic decision system in which all external factors are frozen. The external input D, random seed, and policy parameters are fixed across all runs. The **only variable that differs between runs is the initial internal state E_0**.

System definition

At each discrete timestep t, the system evolves as follows:

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

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

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

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

All parameters $(\theta_0, k_E, \beta, \tau, \lambda)$ are constant and identical across runs. No learning, stochastic sampling, or policy updates are performed.

Experimental condition

Two or more episodes are executed with:

- identical external input D
- identical random seed
- identical policy and parameters

The sole difference is the initial imprint E_0.

Observation

Despite identical inputs and frozen policy, the system can produce **divergent discrete actions** (ALLOW / BLOCK) at different timesteps. Every divergence is fully and deterministically explained by the explicit causal chain:

$$E(t) \;\to\; \theta_{\text{eff}}(t) \;\to\; g(t) \;\to\; \text{action}(t) \tag{5}$$

All variables in this chain are observable at the decision boundary.

> **Under identical inputs and frozen policy, decision divergence arises solely from differences in internal history.**

A single causal-chain diagram (Fig. 1) illustrates this transition.

4 Implications: An Audit Primitive, Not a Performance Method

Before discussing implications, it is essential to clarify what this demonstration **does not** claim. This artifact is **not a learning algorithm**, does not perform optimization, and does not adapt its parameters over time. No claims are made regarding performance improvement, robustness, or behavioral competence, and no generalization beyond the demonstrated configuration is asserted.

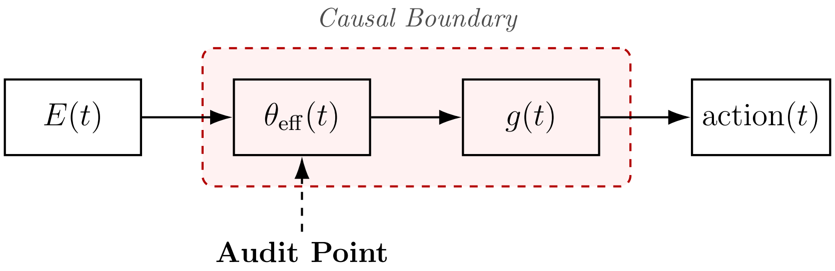

Figure 1: **Causal audit chain at the decision boundary.** The internal state $E(t)$ deterministically modulates the effective threshold $\theta_{\text{eff}}(t)$, which is evaluated against a fixed input to produce a gate value $g(t)$ and a discrete action. The shaded region indicates the **unique location where audit evidence must be recorded** to remain admissible after deployment.

The contribution instead lies in **auditability by construction**. By explicitly fixing the decision boundary and isolating internal history as the sole source of divergence, the demonstration identifies the **minimal set of variables and the unique causal location** that must be recorded for an audit to remain admissible after deployment.

> **This artifact functions as an audit primitive, defining the minimal location and variables required for admissible evidence.**

In this sense, the demonstration does not propose a better decision system, but a **reference design for audit placement**. It shows that admissible audit evidence cannot be recovered post hoc unless the internal state, effective threshold, and gate evaluation are observed at the moment they are converted into action.

5 Limitations and Scope

This work intentionally omits learning, adaptation, and optimization of any kind. The system does not update policies, weights, or internal rules, and does not modify its behavior in response to performance or outcomes. No claims are made regarding **intelligence, learning, consciousness, or behavioral competence**, and the artifact is not a model of cognition or agency.

The demonstrated system is **not intended for deployment**, nor does it represent a practical decision-making architecture. Its sole purpose is to isolate and expose the minimal causal structure required for **audit-admissible decision tracing** under fully deterministic conditions.

These limitations are explicit and essential: the artifact is designed to function as an **audit primitive**, not as an intelligent system. Questions of learning, generalization, or adaptive behavior are therefore **out of scope** and must be addressed in separate work.

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`.