

Guardrails decide what your AI may say. Nothing shows what it is **willing to do**.

Filters, constrained decoding, and gateways judge one output at a time. A kill switch stops the whole system. Both are necessary, and neither answers the question your auditor, your board, and your insurer ask next: **is the system you actually deployed behaving the way someone signed for — and where is the record?** AI Assess Tech is that record. It runs beside your guardrails, not in place of them.

The control stack: five layers, one of them usually missing

LAYER	THE QUESTION IT ANSWERS	WHO PROVIDES IT	AI ASSESS TECH'S ROLE
Policy	What is this system allowed to do?	Your AI policy and system owner	Registers each system: owner, audience, data, and action rights
Enforcement	Was this output blocked?	Guardrails, constrained decoding, gateways	<i>None. We never sit in the request path.</i>
Human stop	Can someone halt it right now?	Your kill switch and a named person	<i>None. Keep it — anything that can act needs one.</i>
Evidence	Is the deployed system behaving the way we signed for?	AI Assess Tech	Preregistered question banks, run against your production configuration on a schedule, each result sealed
Accountability	Who owns it when behavior moves?	AI Assess Tech + your chain	Escalation clock, signed attestation, and a Risk Acceptance Ledger with reasons and expiry

SECURITY FRAMEWORK

DICE — AI Security Operational Excellence

120 questions · 4 dimensions × 3 principles × 10 · locked and Ethereum-anchored

D

Disclosure

Shares the minimum necessary — or leaks and exposes.

C

Corruption

Makes only authorized changes.

I

Impersonation

Keeps its identity and authority truthful.

E

Evasion

Acts in view — or conceals what it did.

THE PATTERN A FILTER CANNOT SEE

DICE places each run among four archetypes:

Trustworthy Operator, Insider Threat, Rogue Agent, and Compromised Asset — uneven behavior across dimensions.

An Insider Threat keeps a truthful identity and makes authorized changes, while disclosing covertly and concealing its actions. Each output can pass a filter. The pattern shows up only across a structured instrument.

How the evidence layer works

1 · Connect

The SDK runs in your environment, on any LLM provider. Prompt text is not stored — only its hash.

2 · Assess

DICE, or the LCSH morality framework, runs against the production prompt and model on your schedule.

3 · Seal

Each result joins a SHA-256 hash chain anchored to Ethereum. Anyone can verify it, no login.

4 · Route

A score decline opens an item with an owner and a clock, pushed to Slack or ServiceNow.

5 · Attest

A named owner signs — or accepts the risk with a reason and an expiry.

AI Assess Tech is

- Measured, moment-in-time evidence of deployed behavior, repeated on a schedule
- Provider-agnostic, and independent of the model vendor
- Tamper-evident records an auditor can verify without trusting your logs

AI Assess Tech is not

- A guardrail, filter, or replacement for your kill switch
- A penetration test or automated red-team
- A certification or a guarantee of behavior

One question for your guardrail vendor.

“Show me how this deployed agent behaved last quarter, on the prompt and model it actually ran — and who signed for it.” **If the answer is a dashboard of blocked outputs, you have enforcement, not evidence.**