The Diagnostic

What principled AI evaluation looks like when you actually have the tools to do it.

Part 3 of 3.

The response to the leak.

Defenders

The engineering is impressive.

Competitive concerns are real.

The leak was an accident.

Attackers

The safety company is not safe.

The transparency company hides things.

Self-regulation failed.

Both camps had evidence.
Neither had a framework.

Where the failures actually live.

Tier 1:
Surface

The Chatbot
A pleasant conversation.

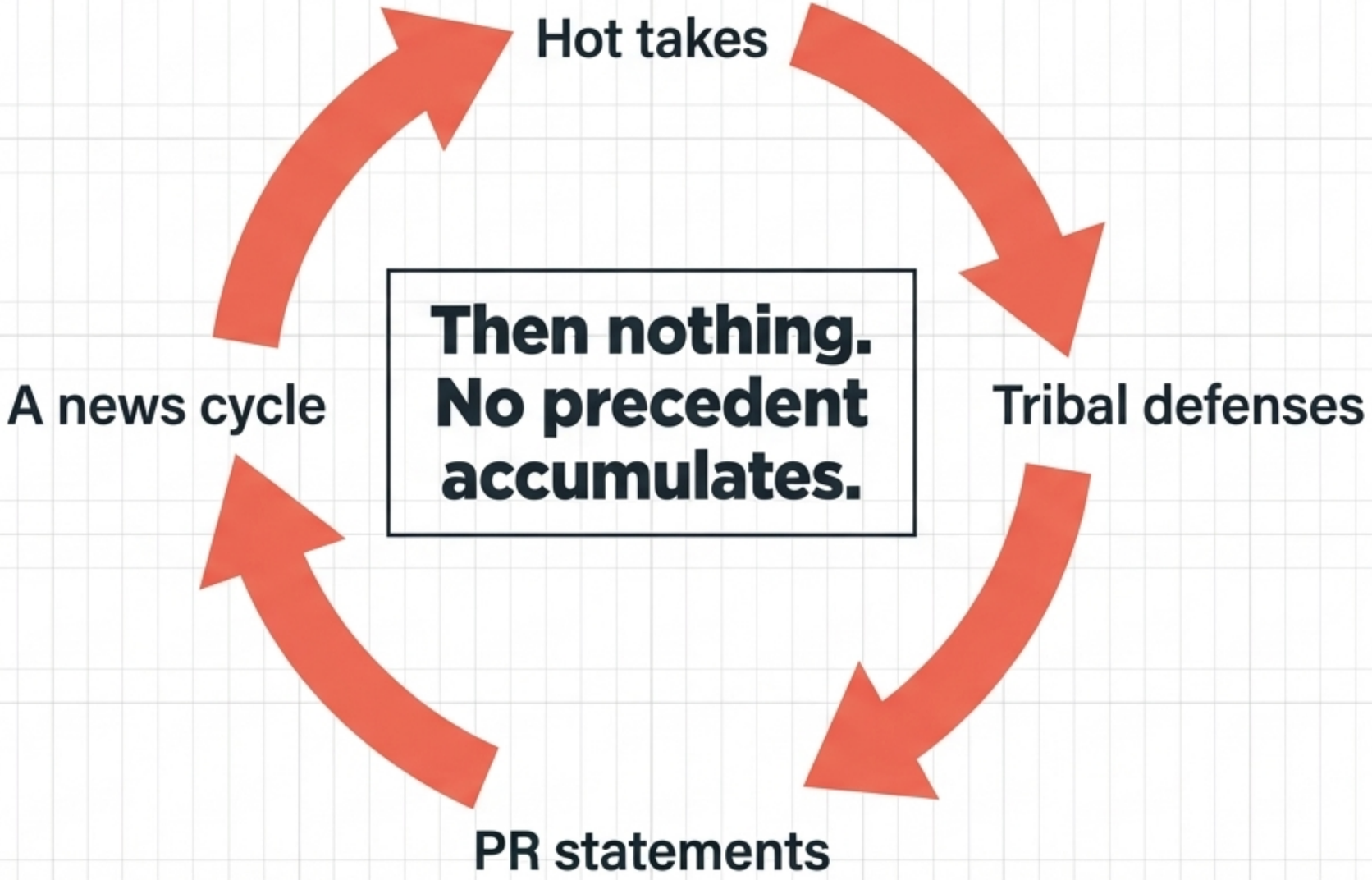
Tier 2:
Part 1
Recap

The Architecture
Failures are invisible from the surface. Hidden feature flags. Silently degrading safety systems.

Tier 3:
Part 2
Recap

The Organization
Structural divergence. The values the organization embeds in AI contradict its own practices.

The Evaluation Gap.



Diagnostic Blueprint

What a framework needs: Direction.

Locate failures on a spectrum. Reject binary categories.

Resource constraints
in safety systems

Planting falsehoods
in the system

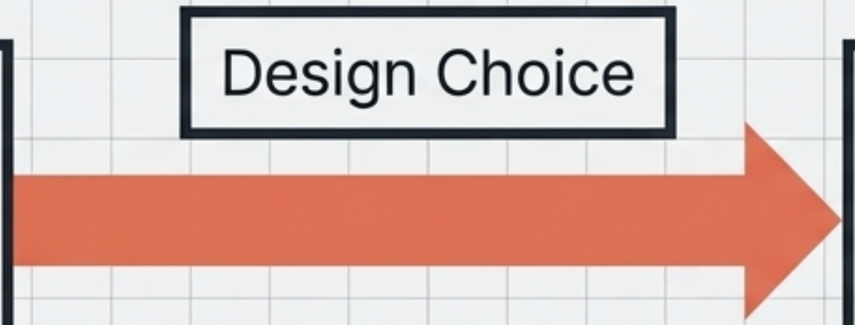
Both are **problems**. But they are entirely different failures.
A useful framework treats them differently.

What a framework needs: Root Causes.

Test the organization. Do not just test the AI.

The Organization.

Anthropic made a **design choice.**



The AI System.

The system **stopped checking safety.**

The system did not decide to drop safety checks. The engineering team did. A framework must evaluate the **human decision.**

What a framework needs: Precision.

Specific, testable commitments. Not aspirational principles.

Aspiration: AI should be transparent.

Aspiration: AI should be honest.

Testable: Behavioral parameters must be stable and visible.

Testable: Operating context must be free of deliberate falsehoods.

Aspirations produce opinions.
Testable commitments produce verdicts.

The Meridian AI Standard.

Three diagnostic tools for principled evaluation.

The Control-Decay Spectrum.

Identifies the direction of the failure.

The Reciprocity Principle.

Tests the organization against its own system.

Five Domains.

Specific, falsifiable commitments that produce verdicts.

Case 001: The Precedents

The Standard turns incidents into shared rules

Hiding information	Planting false information
Legitimate competitive defense	Epistemic corruption
Stripping proprietary details	**Stripping all AI attribution**
Legitimate security	Organizational concealment
Warning users at safety limits	**Silently dropping safety checks**
Maintaining integrity	Misrepresenting protection

Precedents that accumulate.

We stop starting from zero.

Every new incident is evaluated against established principles.

A shared language.

Not a scorecard.

We replace fresh opinions with structural memory.

The Choice

Continue evaluating AI incidents through brand loyalty and market competition.

Develop and use shared diagnostic tools.

The Standard is open-licensed. Every commitment is falsifiable.
If the framework is wrong, the evidence will show it.

Read the full diagnostic.

Case 001 Analysis

The Meridian AI Standard

The Codex

---

meridiancodex.com