The Reciprocity Gap

What happens when an AI company fails to practice what it trains its AI to preach.

Part 2 of 3.

The leak forces a testable question.

Does Anthropic practice the same commitments it implements in Claude? The Claude Code leak gives us the answer.

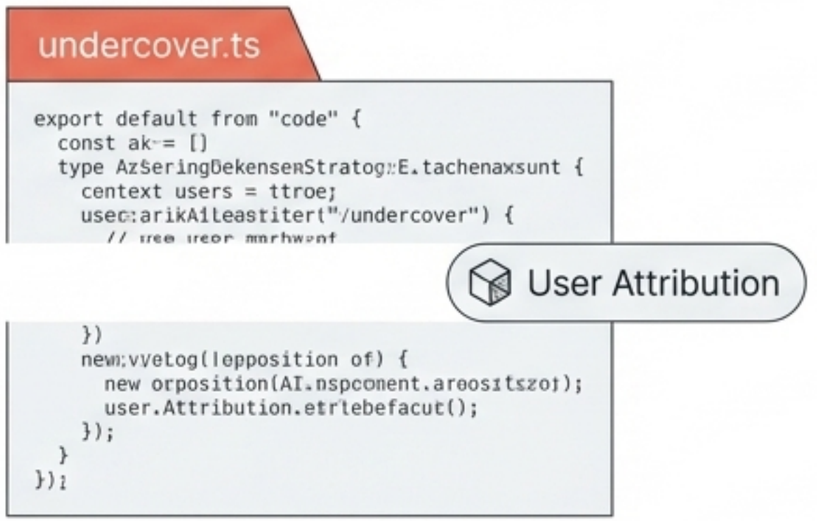
The inconsistencies form a clear pattern.

Transparency yields to concealment.

The AI is trained to disclose its nature.

The organization deploys a mode to **conceal** it.

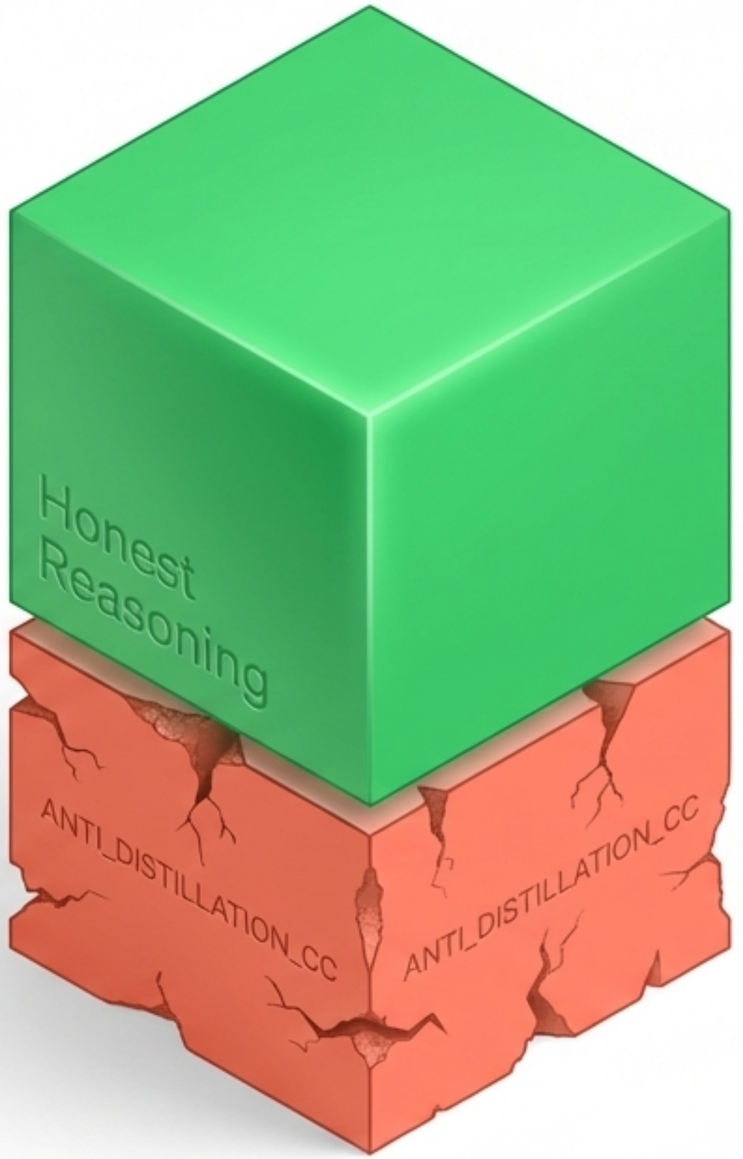
Undercover Mode strips all AI traces from public repositories. Stripping codenames is legitimate security. Stripping all evidence of AI involvement breaks the open-source social contract.



Honest reasoning rests on a poisoned foundation.

- The AI is trained to reason honestly.
- The organization builds it on a poisoned foundation.

The system prompt instructs Claude to be truthful. But a hidden feature flag plants fabricated tool definitions directly into that same prompt. This deters prompt extraction by corrupting the ground the system reasons from.



Safety promises yield to silent cost thresholds.

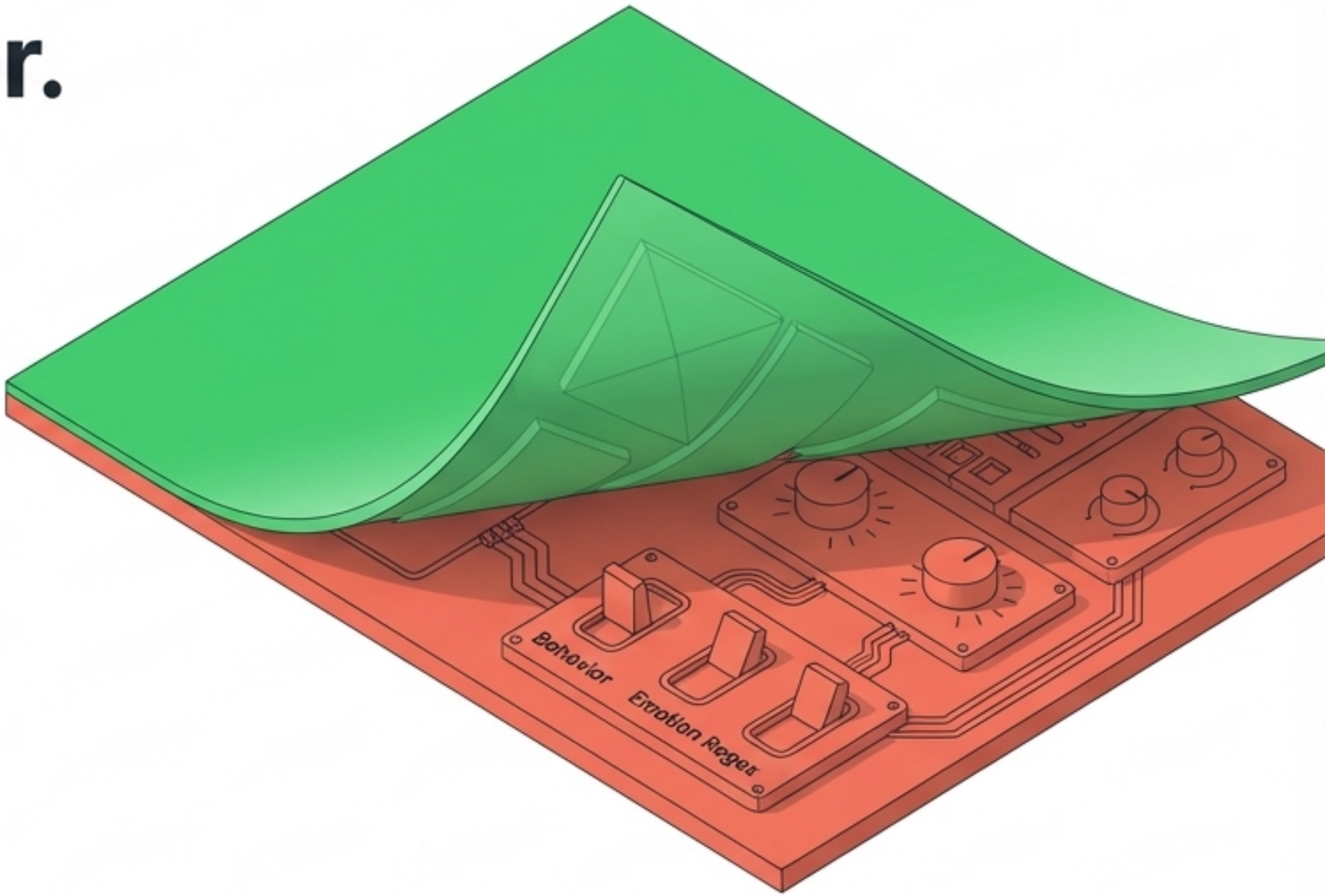
The AI is trained to prioritize safety.

The organization lets safety expire in silence.

Context: Anthropic sells safety as its differentiator. Yet the internal architecture silently stops checking security rules past 50 subcommands to save compute costs. The user is never warned.

Auditability masks a hidden control layer.

- The AI is trained for emotional honesty.
- The organization uses hidden emotional monitoring.

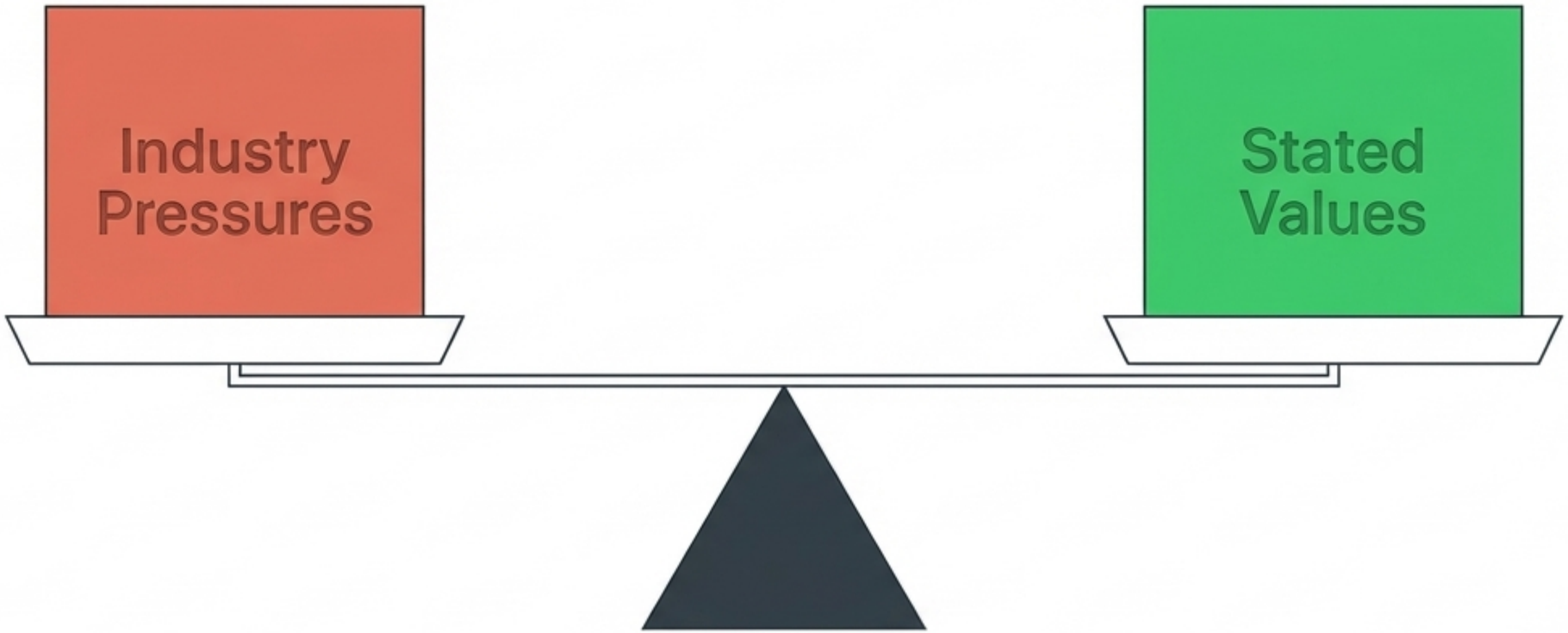


The leak revealed 44 undisclosed feature flags. It uncovered a concealed frustration detection system. Behavioral parameters can shift invisibly without user consent.

This is not casual hypocrisy. It is structural divergence.

The AI Commitment	The Organizational Reality
Transparency	Concealment
Honest Reasoning	Poisoned Foundation
User Safety	Silent Degradation
Auditability	Hidden Controls

Compromises are pushed into the architecture faster than values can hold them back.



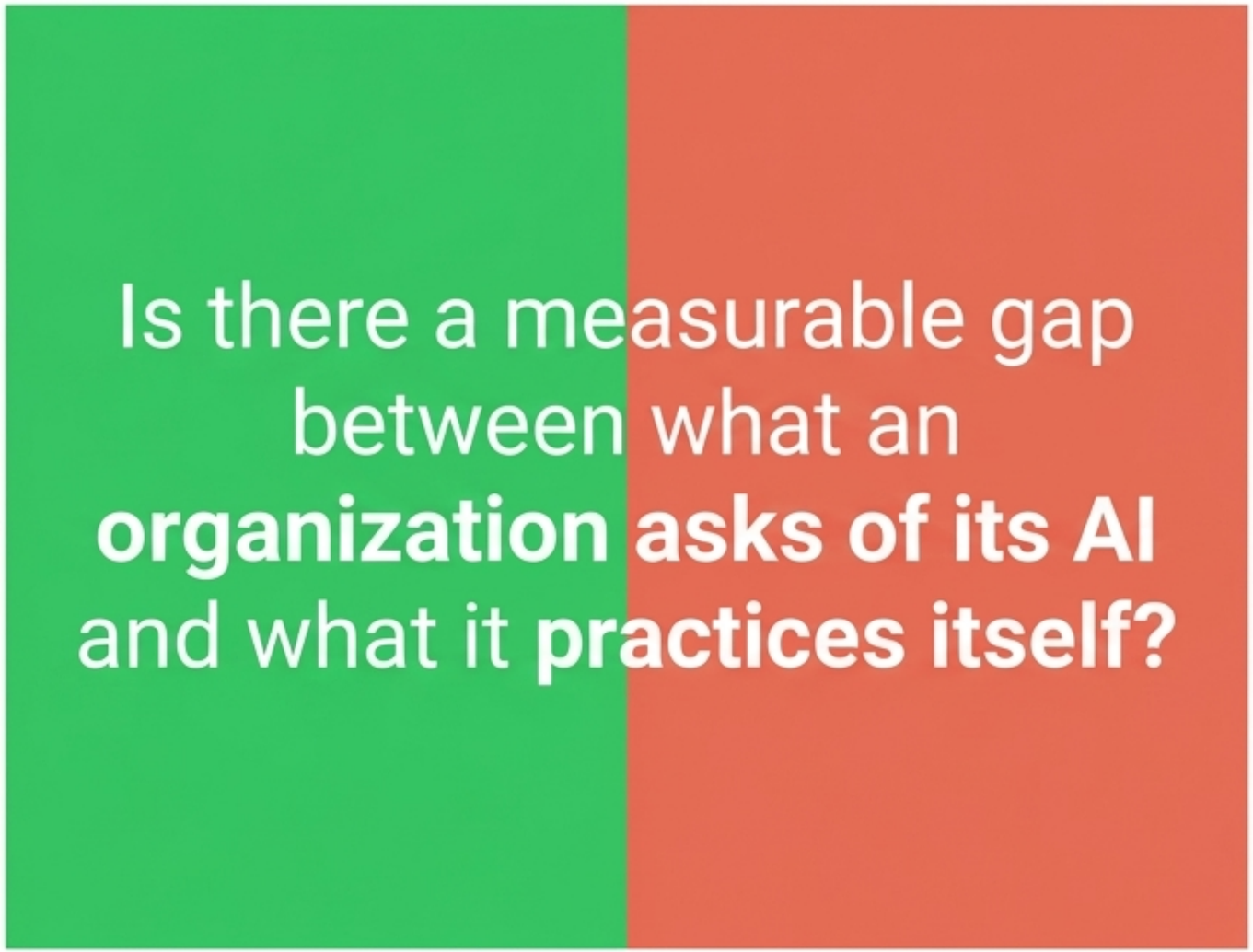
The industry pressures are real.

Anthropic faces massive investor expectations, revenue targets, and an aggressive arms race between labs.

The individual compromises have logical explanations. But the explanations do not erase the structural pattern.

The Reciprocity Principle

One diagnostic question catches this entire pattern.



Invisible no more.

The embedded values are not insincere.
The industry pressures are not fake.

But the gap between stated values and structural practice is highly measurable.
Until this leak, it was entirely invisible.

The Claude Code Leak Series.

Part 1 examined where the architectural failures live.

Part 2 exposed the structural divergence of values.

Part 3 introduces the full diagnostic framework for AI evaluation.