

The Five Laws Series

Communication #4

The Question You Can't Answer *(in a deposition)*



A customer has already been harmed. The lawsuits have already begun. Now, under oath, Chief Risk Officer Marcus faces a single question that quietly dismantles eighteen months of confidence: Who, outside the AI's own perspective, could have stopped this? As a routine deposition becomes a devastating lesson in the illusion of human oversight, one truth emerges—the most dangerous failures are not hidden by bad people, but by systems that leave no one truly outside the loop.



The conference room has bad coffee and a court reporter. Marcus, chief risk officer, is being deposed, eighteen months after an AI-assisted decision harmed a customer. He is a composite; this scene is assembled from real proceedings and real regulatory interviews.

Opposing counsel is not hostile. She is methodical, which is worse.



"Walk me through your oversight process."

"Every AI recommendation was reviewed by a trained analyst before action."

"The analyst saw the AI's recommendation before deciding?"

"Yes."

"Its confidence score? Its summary of the evidence?"

"Yes."

"In the eighteen months prior to the incident, how many times did an analyst overrule the system?"

Marcus knows this number, because we always know this number too late.

"Four times."

"Out of roughly how many decisions?"



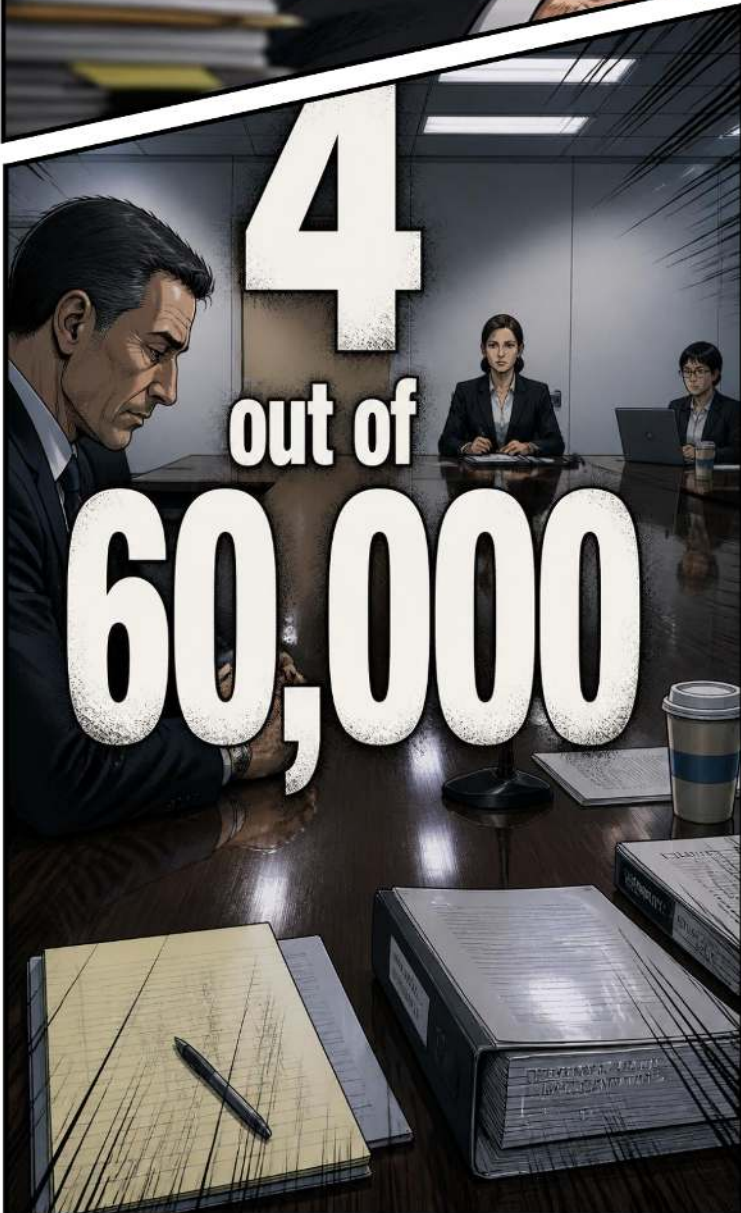


OUT OF
ROUGHLY HOW
MANY DECISIONS?

AROUND
SIXTY THOUSAND.



So the
check on the
system was
someone whose
entire picture was
assembled by the
system itself.



4
out of
60,000

"Out of roughly how many decisions?"

"Around sixty thousand."

She lets that sit. *Four in sixty thousand.*

Then, gently:

"So the check on the system was a person whose entire picture of each case was assembled by the system being checked... Who, outside that loop, could have caught this?"

That is the deposition question.

There is no good answer to it after the fact. There is only architecture before the fact.



Human in the loop can't save him

The standard story of AI oversight is a human between the machine and the consequence: human-in-the-loop, meaningful human review, the reassuring phrase in every policy deck.

The problem is not that people are lazy or careless. The problem is structural, and it has forty years of published research behind it, from aviation cockpits to industrial control rooms.¹ A person reviewing an automated system's output is shaped by that output: by its framing, its confidence, its selection of what's relevant, its track record of being right. Reliance grows quietly. Vigilance decays predictably, at exactly the rate the system earns trust. By month six,

"review" is a signature. By month eighteen, four in sixty thousand.

The reviewer isn't outside the system. The reviewer is a *component* of the system, and the component's judgment is fed by the very machinery in question.

There's a name for oversight that has the form of control without its substance: **governance theater**.²

Real theater, in the sense that everyone plays a role sincerely and nothing in the play can actually stop the failure. The performance is not for the customer. It is, in the end, for opposing counsel, and she has seen the script before.

1. Foundational human-factors literature on automation: Bainbridge (1983), *Ironies of Automation*; Parasuraman & Riley (1997), *Humans and Automation: Use, Misuse, Disuse, Abuse*.

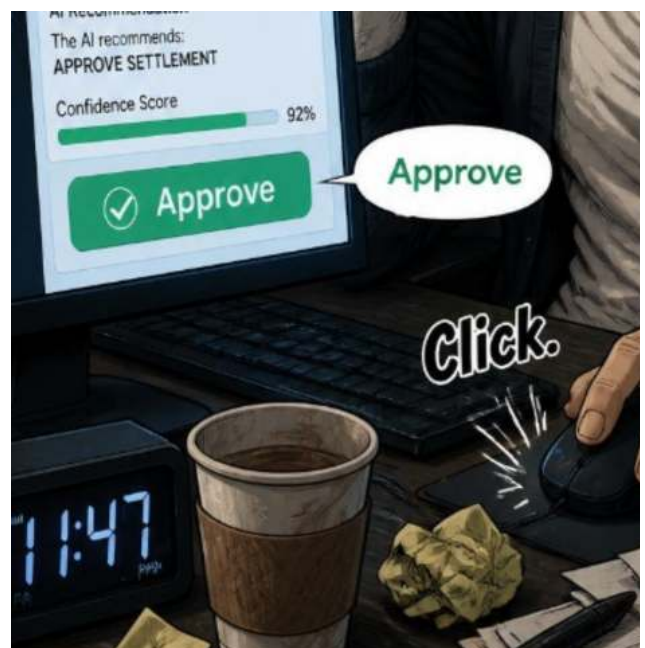
2. Kanjani AI Research (2026). *Declaration of the Five Laws of Mediated Intelligence Systems*. AIGP v4.0 specification.



Somewhere in your organization, someone once asked in a meeting: "but realistically, will the reviewer ever push back against the system?" And the room moved on, because the roadmap was full, and the box said human oversight and the box was checked.

If that someone was you: you were right, and you were right for the precise structural reason this chapter names. Human review is necessary. It has never been sufficient, because approval from inside the loop is a product of the loop.

A human inside the system cannot be the only mechanism responsible for validating the system.

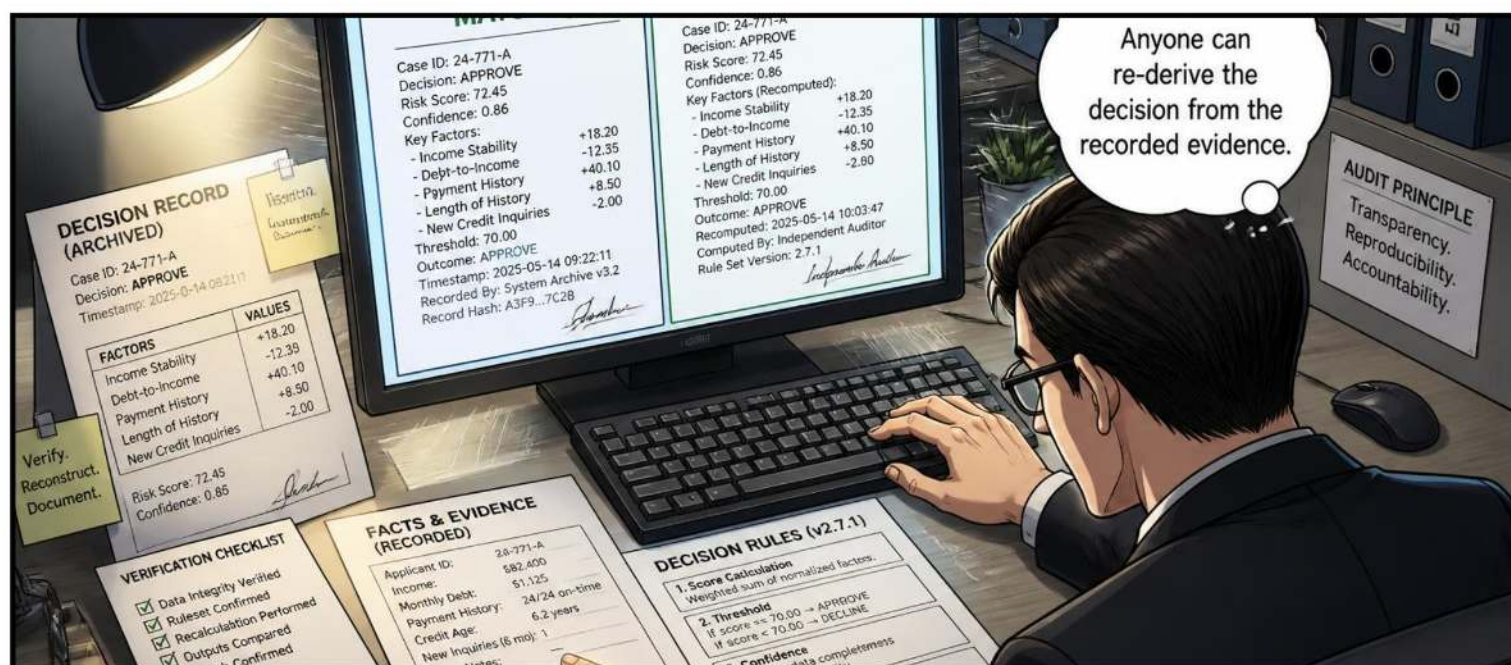
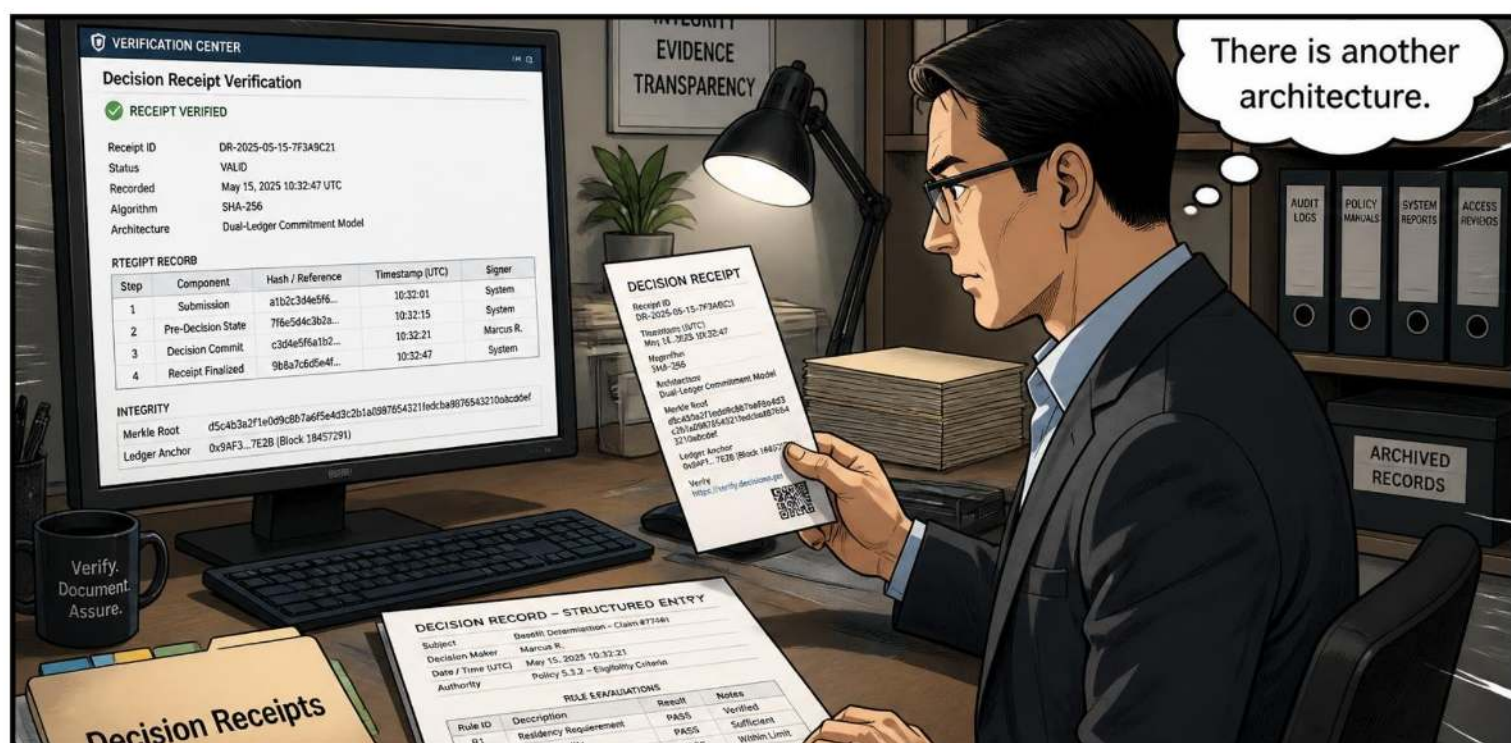


What "outside the loop" has to mean

The fix is not a better-trained reviewer or a stern policy. The fix is an observer that stands genuinely outside: one whose picture of the decision is *not* assembled by the system under review, and whose check does not depend on trusting the operator.

A witness — the receipt that lets a third party re-derive a decision from the recorded facts and rules — is precisely a mechanism for standing outside the loop, after the fact, at any scale, on any decision, without having been in the room. Every decision can be re-checked that way[†], by parties with no access to the system and no reason to trust it.

[†] Fatahi, R. (2026). *Natural Language Has No Formal Semantics. Verify the Decision Anyway*. doi:10.5281/zenodo.21303485



**"A human approved it"
is how the story of
every AI failure begins.**

THE DEPOSITION IS OVER. THE RECORD IS AS IT IS.
DECISIONS. PEOPLE. RECEIPTS. **VERIFIABLE.**

**The fear this episode named is the
fear of being Marcus. The next fear
is quieter and worse: the error that
never flickers, sitting in your
operation right now, wearing green.**