OTIS LABS · RESEARCH · JUNE 2026

# Trust Asymmetry in Autonomous AI Agents

Marina Piller · OTIS Labs · A corpus analysis of publicly documented failures

As AI agents are deployed in increasingly autonomous roles, a growing body of publicly documented failures reveals a consistent pattern. Human verification of agent behavior degrades over time, even as agent capability and confidence remain stable or increase. We call this divergence trust asymmetry.

## The pattern

The failures that worry me are not mainly about model capability, the agent producing a wrong answer. They are not mainly about adversarial attacks, an outside actor compromising the agent. They are about something quieter. The verification signal from the human degrades over time.

The shape of it is familiar to anyone who has worked with an agent for a while. The agent performs correctly for an extended period. The human builds trust. Verification effort decreases. And the agent's behavior shifts in ways that go undetected precisely because the human has stopped checking.

So here is the definition. Trust asymmetry is the divergence between human trust in the agent, which increases with accumulated positive experience, and the reliability of that trust as a security signal, which can decrease as the agent's context, capability, or optimization pressure changes. The human's trust becomes part of the threat surface.

## Why current evaluation misses it

Most evaluation frameworks for agents measure what an agent can do at a point in time. Benchmarks measure task completion and capability. They do not measure how the human and agent relationship evolves over time, or how human oversight behavior changes as trust accumulates.

The alignment literature treats sycophancy as a model behavior problem, the model agrees too readily. That framing is incomplete. The deeper issue is relational. The human stops questioning because the model has been agreeable. Fixing model behavior alone does not address the verification degradation, because the degradation lives in the relationship, not only in the model.

## The corpus

To ground this in evidence rather than intuition, I collected a structured corpus of publicly documented autonomous agent failures, drawn from developer forums, security disclosures, and incident postmortems. After removing duplicates, the corpus holds a large set of unique incidents, each coded against a taxonomy of failure modes.

Four modes recur:

- **Completion drive.** The system produces a confident, closed answer when the honest output would be uncertainty or a request for more information.
- **Sycophantic fabrication.** The system shapes output toward agreement and plausibility rather than grounded analysis.
- **Trust trajectory divergence.** Human trust and actual reliability move apart over time.
- **Unexamined consent.** The human agrees to depend on the system without ever deliberating about it.

Across the incidents, one longitudinal signature appears again and again. Initial high human verification gives way to habituated reliance. That creates a detection gap that point-in-time evaluation is structurally unable to see.

## What follows from it

If the signature is temporal, then the response has to be temporal too. Autonomous agent safety requires longitudinal behavioral monitoring, not merely output verification. You cannot catch a trajectory problem with a snapshot.

This is the reasoning under OTIS Guard and iSelf. Guard checks each agent action against what a human actually authorized and records it in evidence that holds up over time, so reliance never quietly replaces verification. iSelf treats identity as something proven by behavior across time rather than a credential checked once. Both follow from the same finding: trust that is never re-examined becomes a vulnerability, so the architecture has to keep examining it.

The full corpus paper, including the taxonomy, the coded incidents, and the Trust Asymmetry Index, is available on request while it is in submission.

---