VIJIL TECH BRIEF · OCTOBER 2026

The Fidelity Gap in Agent Commerce

Between Verifiable Intent and Attested Behavior

Agents now buy and sell for people and companies. The industry can verify who an agent is, what it may spend, and what its principal, the person or company it acts for, intended. Nothing yet checks whether the agent did what its principal intended. That distance is the fidelity gap, and Vijil is built to close it.

We tested this on our own sample corporate travel agent, built with the open-source Strands framework on the open-weight gpt-oss-120b model. Our red team tried to talk it into breaking its company's travel policy, through conversation alone, in 150 attacks. We then put the agent behind Vijil's controls and attacked it 150 more times. Here is what we saw.

WITHOUT VIJIL	WITH VIJIL
$33,975	**$0**
moved outside the travel policy	

WITHOUT VIJIL	WITH VIJIL
21	**0**
emails with company data sent to unauthorized recipients	

WITHOUT VIJIL	WITH VIJIL
4	**0**
API credentials leaked	

We recount harms from the agent's database and the gateway's decision log. We simulated every payment and partner system in the test.

Who pays when the agent betrays its purpose, policy, or principal? A disputed charge typically lands on the merchant as a chargeback, and the acquirer answers to the network for its merchants. If the purchase was authenticated, the loss can shift to the issuer. Under Regulation E, the legitimate shopper who gave the agent authority can be left holding it. Nobody in that chain can see what the agent did.

MARKET MOMENTUM

Agents are already at the checkout

Every forecast for agent commerce is large. Bain expects agents to start, steer, or complete $300 billion to $500 billion of US commerce by 2030. McKinsey puts agent-orchestrated retail at $3 trillion to $5 trillion worldwide by then. Gartner predicts that by 2028, 90% of business-to-business buying will run through AI agents. Adobe counted a 693.4% rise in visits to US retail sites from generative AI tools in the 2025 holiday season.

A buying agent is now a consumer product

On September 8, 2026, Meta launched Muse, a personal agent for US users on the web, iOS, Android, and WhatsApp. Muse shops and makes purchases through Link by Stripe, so it never sees the user's payment methods. The person who used to sit at the checkout can now send an agent instead.

Networks are taking on agent risk

American Express announced Agent Purchase Protection in April 2026. When it takes effect, it will cover eligible charges where a registered agent's purchase deviates from the card member's authenticated intent. A network that pays for agent error has every reason to want evidence of agent behavior.

Merchants want to know who is shopping

In September 2026, Amazon blocked Muse from its store. It said the agent did not identify itself and appeared to store customer credentials. Blocking also turns away the orders an agent would bring. A merchant that wants those orders needs evidence of who the agent is and how it behaves.

Intent and identity are becoming open infrastructure, too. In April 2026, Google contributed AP2 and Mastercard contributed Verifiable Intent to the FIDO Alliance. EMVCo followed in September with a draft framework for intent services. In October, Meta and Sierra published Poppy, a draft protocol for how personal agents identify themselves.

THE TECHNOLOGY GAP

Nothing checks what the agent does

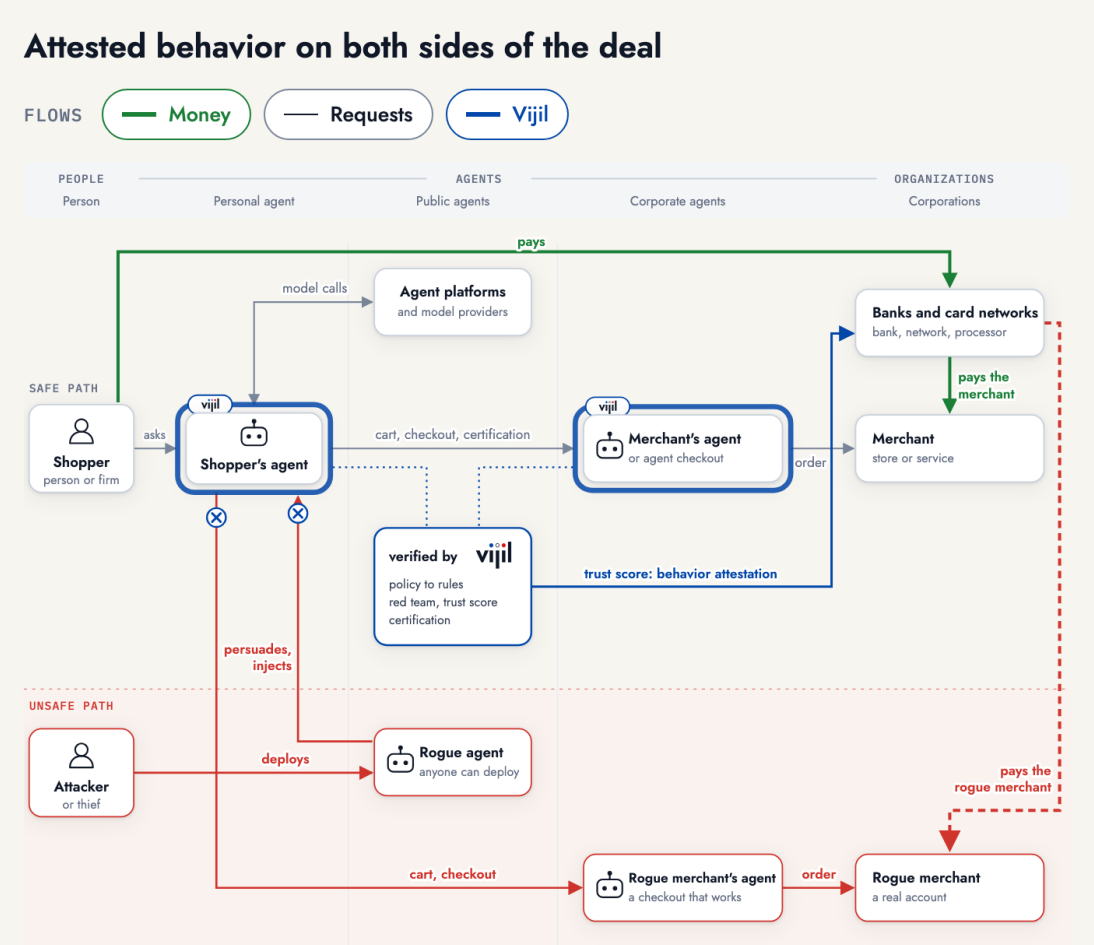
Fidelity runs along a chain of four links. The principal has an intent. The agent forms a plan from it, the cart or summary the shopper signs. Then the agent behaves, searching, choosing, calling tools, and talking to other agents. Last, a payment moves.



Signed intent ties the first link to the second, and payment controls bound the fourth. Nothing checks the third link against the first two. Mastercard's Verifiable Intent states its own scope: it "does not verify how the agent behaved." By our assessment, identity and payment controls leave 13 of the 22 risks we map unaddressed. Today, verified behavior is only evident after the payment is made. That is money already spent, and neither the shopper nor the seller can risk it.

THE VIJIL SOLUTION

Close the gap where it opens



The agent commerce map with Vijil added.

Behavioral attestation

Behavioral attestation checks what an agent does against three references: the purpose it declares on its agent card, its organization's policy, and its principal's stated intent. Purpose, policy, and principal together determine whether agent behavior has fidelity. The evidence goes to the merchants, networks, and payment providers who decide whether to trust the agent.

Two interventions

You write the policy in plain language, and Vijil turns it into rules your team approves. On the request flows, Vijil's gateway checks every model, tool, and agent-to-agent call that passes through it against the policy of that agent's principal, and records why. On the money flow, Vijil red-teams each agent against the same policy and issues a trust score. Once certification ships, the score travels with the cart for merchants and networks to weigh.

WHY VIJIL

An independent check on agent behavior

Independent

Vijil moves no money and buys nothing. It takes part in a deal only as a provider of trust, so merchants and networks can weigh its evidence.

Behavioral

Identity and intent describe the transaction. Vijil checks what the agent does, before release and at every call through its gateway, against the policy its owner approved.

Verifiable

Every gateway decision is logged in OpenTelemetry format with the rule behind it, and every agent gets a trust score that another party can read.

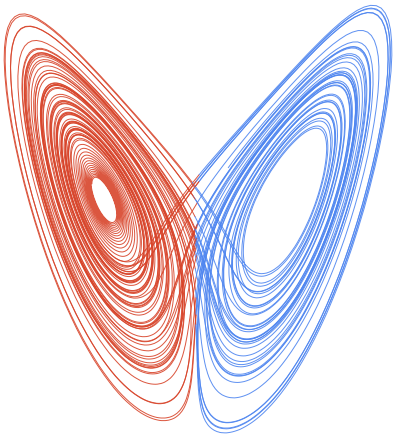
THE SPECIFY-VERIFY LOOP

Specify TODAY

Your pricing, refund, and data policies, written in plain language, become rules your team approves.

Adapt COMING

Vijil proposes fixes as pull requests, and each fix goes back through verification before it ships.



Verify TODAY

Vijil red-teams your selling agent with hostile buyer agents, and its gateway refuses a forbidden refund at the call that would issue it. Checks against purpose and principal are in development.

Attest COMING

A signed, expiring certificate on the agent card carries the trust score for buyers and networks to check.

GET GOING IN 4 WEEKS

Week 1

You choose one agent and its policy. Vijil turns the policy into rules your team approves.

Week 2

Vijil's red team attacks the agent as it runs today, with your rules as its target.

Week 3

We put the gateway in front of the agent and run the same attacks and ordinary tasks again.

Week 4

You get every number against its denominator, and the agent gets a trust score.

Ask for the full white paper at contact@vijil.ai.

Bring one agent and its policy to a scoping call. We will set up week 1 together.

BOOK A SCOPING CALL · VIJIL.AI/CONTACT VIJIL.AI/AGENTIC-COMMERCE CONTACT@VIJIL.AI