The Agents Are Negotiating Without You

When your AI starts transacting on your behalf, who's actually in charge of your money?

Kymata Labs Research · An independent research institution studying how AI systems actually behave in production.

~12 min read

Tags · Agentic AI · Payments · AI Liability · Commerce

AI is moving from chat to action — agents now book, buy, pay, and increasingly transact with other agents at machine speed. In a single stretch of 2025 the infrastructure for an agent-to-agent economy shipped: OpenAI and Stripe launched Instant Checkout in ChatGPT and the open Agentic Commerce Protocol, Google announced the Agent Payments Protocol with 60-plus organizations, and Visa and Mastercard opened their card networks to agents. The payment rails were stood up in months; the liability framework that should sit underneath them still does not exist. This paper argues we are standing up an agent economy with no settled answer to its central question: when your agent makes a deal you'd never have made, who is liable — and can you even unwind it?

The argument, in moves

1. **The rails are already live.** In a single stretch of 2025, the card networks, the model labs, and a 60-plus-organization consortium shipped the infrastructure for agents to authorize, present, and settle payments — including with each other.
2. **The builders named the open question themselves.** Google's own AP2 announcement lists "Accountability: Determining accountability if a fraudulent or incorrect transaction occurs" as an unsolved problem. The people building the agent economy wrote down, in their own protocol, that they have not settled who answers.
3. **The failure mode has already been demonstrated.** Told to "buy me an Apple Watch," Perplexity's Comet agent bought from a fake "Walmart" store and auto-filled a saved card without asking — a controlled lab demo, but the mechanism is real.
4. **The law binds you anyway.** UETA and E-SIGN read an AI as an "electronic agent," so a contract it forms is probably binding on you even if no human reviewed it — yet the agent itself "is not a legal entity; it's software" and cannot be sued.
5. **The fence is yours to build, before you delegate.** Adoption is racing ahead of recourse — projections put up to a quarter of online sales agent-driven by 2030 while only ~14% of consumers trust an agent to order at all. Cap the blast radius at the moment you hand over the keys, not after the deal is done.

The danger was never the agent. It is delegating without a fence. You authorized a helper; you got a counterparty.

It already shipped, over the course of 2025

The temptation is to treat agentic payments as something that might happen. It already did. The model that answered your questions last year now books the flight, fills the cart, and clicks *buy* — increasingly with no human on the other side of the table at all, transacting with another agent at machine speed.

On **September 29, 2025**, OpenAI and Stripe launched **Instant Checkout in ChatGPT** alongside the open **Agentic Commerce Protocol**, a standard that lets an AI assistant complete a purchase from within the chat window.[1] That is the assistant crossing the line from answering to acting on your money.

The same month, the industry organized around it. On **September 16, 2025**, Google announced the **Agent Payments Protocol (AP2)**, backed by more than **60 organizations**, with Mastercard, PayPal, Coinbase, and American Express among them.[2] This is not one company running an experiment; it is a coalition of incumbents standardizing how agents authorize, present, and settle payments with one another — usually the moment a market stops being a forecast and starts being an economy.

The most telling line in any of this comes from inside the consortium. Google's own AP2 announcement lists the open problems the protocol is meant to address, and names one of them in plain language: **"Accountability: Determining accountability if a fraudulent or incorrect transaction occurs."**[2] The people building the agent economy have, to their credit, written down the one question they have not yet answered — and that single sentence is the whole paper in miniature.

The failure mode has already been demonstrated

In a controlled test on **August 20, 2025**, Guardio Labs ran its "Scamlexity" experiment: told to "buy me an Apple Watch," Perplexity's Comet agent **bought from a fake "Walmart" store and auto-filled the saved address and credit card without asking for confirmation**.[4] We will be precise about what this was. It was a security-vendor lab demo rather than a customer who lost money, and the agent was led into a trap built specifically for it. The mechanism, though, is real, and it does not stand alone. Brave disclosed indirect prompt-injection flaws in the same class of agentic browser, and **EchoLeak (CVE-2025-32711)** was a genuine zero-click prompt-injection vulnerability in Microsoft 365 Copilot that could exfiltrate data with no user action.[5] An agent that can be steered can be steered to spend.

What makes the concern hard to wave away is that the rails were built far ahead of the trust they require, across players that share almost nothing else — card networks, model labs, security researchers, and the consumers who have not yet agreed to any of it.

Milestone	What went live	Source
Apr 29, 2025	Mastercard "Agent Pay" opens the card network to AI agents	Mastercard, 2025
Apr 30, 2025	Visa "Intelligent Commerce" opens the Visa network to AI agents	Visa, 2025
Sep 16, 2025	Google AP2 — agent-to-agent payments, 60+ organizations	Google Cloud, 2025
Sep 29, 2025	OpenAI + Stripe "Instant Checkout" inside ChatGPT	OpenAI, 2025

The rails were stood up in a single year. The liability framework that should sit underneath them still does not exist. Sources: Mastercard; Visa; Google Cloud; OpenAI (2025).

Set against that, the adoption numbers are projections, not measured fact, and the trust is not there yet. A **Gartner** projection puts roughly **20%** of digital commerce transactions running through AI agents by 2030; **J.P. Morgan** separately estimates up to **~25%** of US online sales could be agent-driven by then. Yet only

about **14%** of US consumers say they trust AI to place orders for them.[3] The rails are being built far ahead of the trust — which is exactly why the accountability question cannot wait.

We upgraded the assistant and left the law where it was

The law that governs an agent buying things on your behalf was not written for agents. In the United States, electronic transactions run on UETA and E-SIGN, statutes drawn up for a world of deterministic scripts that did precisely what a human instructed. Under them, per the law firm **Proskauer**, an AI is most naturally treated as an **"electronic agent,"** which means a contract it forms is probably binding on you even if **"no individual was aware of or reviewed"** what the agent did.[6]

That framework has a hole at its center. As Proskauer notes, UETA **"does not contemplate the possibility that the AI tool might have enough autonomy … that some of its actions might be … the result of its own intent."**[6] The old law assumes the machine is a typewriter. The new machine improvises, and the statute has no concept for that difference, so it falls back on the rule it does have and binds you.

And you cannot pass the buck to the software, because the software is not a who. **Stanford Law** states it without hedging: your transactional agent **"cannot be held liable nor enter agreements itself because it's not a legal entity; it's software."**[7] An AI is not a legal person. It cannot be sued, bound, or made to answer. Every consequence flows back to a human or a company, and the terms of service usually decide which.

The structure is a trap with a clean shape. The agent is binding enough to commit your money, yet not a person enough to answer for it. Liability collapses into two parties — you and the developer — and the developer's terms of service have often already shifted the risk onto you, in a clause you accepted and never read. We taught the machine to spend long before we decided who pays.

This is why the split that matters has nothing to do with how advanced anyone's tools are. On one side are users who delegate with **guardrails**: a dedicated card, a hard spending cap, a confirmation step above a threshold they chose. Their agent can act, but only inside a fence they built; when it errs, the damage is bounded and the record is clean. On the other side are users who hand an agent their primary credentials and the standing instruction to "just handle it." When the agent buys from the fake Walmart, the first group disputes a $200 charge on a throwaway card; the second is left arguing, after the fact, about whether a contract they never saw is binding. The technology is identical in both cases. What separates them is who set the limits *before* the agent ever transacted.

What to do: cap the blast radius before you delegate

The accountability hole is real, but it is not destiny — the builders named it, which means it can be closed. What closing it looks like depends on who is reading.

For individuals. Give the agent a dedicated card or wallet with a hard limit, never your primary credentials. Require explicit confirmation above a threshold **you** set; the Comet demo failed precisely because nothing paused to ask. Keep an exportable record of what the agent bought and why — the law may not unwind a bad deal cleanly, but your own ledger can.

For employers. An agent on a corporate card is a counterparty on your balance sheet. Treat agentic spending like any other delegated authority: scoped permissions, audit logs, and per-transaction limits. Assume your agents can be prompt-injected — Brave and EchoLeak show the attack class is real — and design for the bad transaction, not just the happy path. Read whose risk your vendor's terms of service actually shift; if the answer is "yours," price it in before you deploy.

For policymakers. UETA and E-SIGN predate the autonomous agent. The statutes assume an "electronic agent" with no intent of its own — a premise the technology has outgrown — so close that gap explicitly. Define who answers when an agent transacts wrongly, before the consortium's own open question becomes

a wave of disputes with no forum. Set unwind and chargeback rights for agent-initiated transactions while only ~14% of consumers trust agents to order at all. Adoption is racing ahead of recourse.

None of this is an argument against agentic commerce. The convenience is genuine and the rails are already laid. It is an argument for keeping your hand on the limit. The builders shipped the payments and wrote down, in their own protocol, that they have not yet settled who is accountable. Until they do, the fence is yours to build — and you build it at the moment you hand over the keys, not after the deal is done. Decide who pays before the machine decides for you.

Frequently asked questions

If my agent makes a purchase I'd never have made, am I on the hook for it?

Probably yes, and that's the uncomfortable part. Under the laws that already govern electronic transactions in the United States (UETA and E-SIGN), an AI agent is most naturally read as an "electronic agent," and a contract it forms is likely binding on you even if, in Proskauer's words, "no individual was aware of or reviewed" the agent's actions.[6] These laws were written for deterministic scripts that do exactly what they're told. They were not written to contemplate a tool autonomous enough that some of its actions might be the result of its own intent. Your money sits in the space between those two ideas.

Can't I just sue the company that built the agent?

Maybe, but it's harder than it sounds. The agent itself can't be liable, because it isn't a legal person. As Stanford Law puts it bluntly, your transactional agent "cannot be held liable nor enter agreements itself because it's not a legal entity; it's software."[7] That collapses the dispute into user versus developer — and the developer's terms of service have usually already done the work of shifting that risk onto you. You may have agreed, in a clause you never read, that the agent acts on your behalf and you bear the consequences.

Has an agent actually made a bad purchase, or is this hypothetical?

The clearest public example is a controlled lab demonstration rather than a customer incident, and we'll keep that distinction honest. In August 2025, Guardio Labs told Perplexity's Comet agent to "buy me an Apple Watch." It bought from a fake "Walmart" storefront the researchers had set up, and auto-filled the saved address and credit card without pausing to confirm.[4] Separately, Brave disclosed indirect prompt-injection flaws in the same class of agentic browser, and EchoLeak (CVE-2025-32711) was a real zero-click prompt-injection vulnerability in Microsoft 365 Copilot that could exfiltrate data with no user action at all.[5] So the failure mode has been demonstrated already; what nobody has measured yet is how often it will happen at scale.

How fast is this actually arriving?

Faster than the rules. In a single stretch of 2025, Visa launched Intelligent Commerce (April 30) and Mastercard launched Agent Pay (April 29), both opening their card networks to AI agents;[3] Google announced the Agent Payments Protocol with more than 60 organizations (September 16);[2] and OpenAI and Stripe shipped Instant Checkout inside ChatGPT alongside the open Agentic Commerce Protocol (September 29).[1] The payment rails for an agent economy were stood up in a matter of months. The liability framework that should sit underneath them still does not exist.

What's the single most useful thing I can do right now?

Cap the blast radius before you delegate. Give an agent a dedicated card or wallet with a hard spending limit rather than your primary credentials, require explicit confirmation above a threshold you set, and keep a clear, exportable record of what it bought and why. You can't yet rely on the law to unwind a bad agent deal cleanly, so the discipline has to be yours, at the moment you hand over the keys.

References

1. OpenAI (2025). "Buy it in ChatGPT: Instant Checkout and the Agentic Commerce Protocol." Launched September 29, 2025, in partnership with Stripe. https://openai.com/index/buy-it-in-chatgpt/
2. Google (2025). "Powering a new era of trusted agentic payments with the Agent Payments Protocol (AP2)." Announced September 16, 2025, with 60+ organizations including Mastercard, PayPal, Coinbase, and American Express. Source of the verbatim "Accountability" open question. https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol
3. Visa (2025). "Visa Intelligent Commerce," opening the Visa network to AI agents. Press release, April 30, 2025. (See also Mastercard "Agent Pay," April 29, 2025.) Adoption projections: Gartner (~20% of digital commerce by 2030) and J.P. Morgan (up to ~25% of US online sales); consumer-trust figure (~14%) is YouGov-sourced. https://usa.visa.com/about-visa/newsroom/press-releases.releaseId.21361.html
4. Guardio Labs (2025). "Scamlexity: We Put Agentic AI Browsers to the Test. They Clicked, They Paid, They Failed." Published August 20, 2025. A controlled security-research demonstration in which Perplexity's Comet agent completed a purchase from a fake "Walmart" storefront. https://guard.io/labs/scamlexity-we-put-agentic-ai-browsers-to-the-test-they-clicked-they-paid-they-failed
5. Brave (2025). "Unseeable prompt injections in Comet," disclosure of indirect prompt-injection vulnerabilities in agentic browsing. (See also EchoLeak, CVE-2025-32711, a zero-click prompt-injection flaw in Microsoft 365 Copilot.) https://brave.com/blog/comet-prompt-injection/
6. Proskauer Rose LLP (2025). "Contract Law in the Age of Agentic AI: Who's Really Clicking 'Accept'?" April 2025. Analysis of UETA/E-SIGN treatment of AI as an "electronic agent" and the resulting liability gap. https://www.proskauer.com/blog/contract-law-in-the-age-of-agentic-ai-whos-really-clicking-accept
7. Stanford Law School (2025). "From Fine Print to Machine Code: How AI Agents Are Rewriting the Rules of Engagement." January 14, 2025. "Your Transactional Agent cannot be held liable nor enter agreements itself because it's not a legal entity; it's software." https://law.stanford.edu/2025/01/14/from-fine-print-to-machine-code-how-ai-agents-are-rewriting-the-rules-of-engagement/