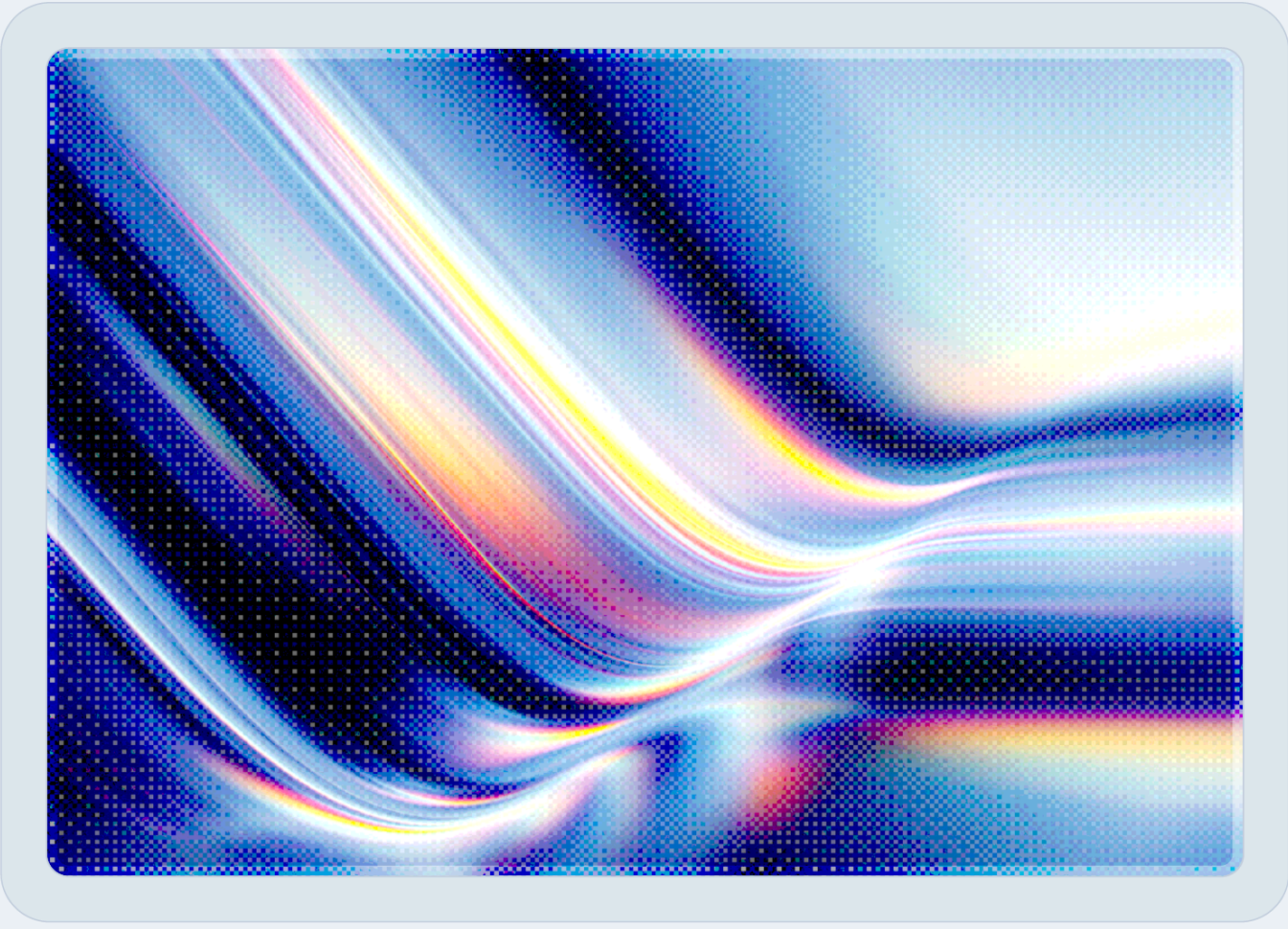
CSAI Foundation | Cloud Security Alliance

Agents at the Checkout: Loss Allocation in Frontier-Lab Payment Wallets

Who Pays When an Agent Buys Wrong, and Who Controls the Checkout

2026-11-15

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

On November 13, 2026, OpenAI introduced ChatGPT Wallet, which lets its "dots" personal agent complete purchases on merchant websites without seeing the user's card number. The wallet was built with American Express, Mastercard, and Visa. Each purchase requires a biometric or device security check, after which the card network issues a payment token for that checkout [1][2][3]. This appears to be the first branded payment wallet from a frontier AI lab designed to operate across ordinary merchant checkouts rather than inside a closed in-chat flow, although other platforms have offered agent checkout with one-time cards and per-purchase approval [3]. The launch changes two questions that security, risk, and payments teams have so far treated as theoretical: who absorbs the loss when an agent buys the wrong thing, and who controls access to the checkout.

On loss allocation, the wallet addresses credential exposure but leaves the authorization problem largely where it was. Network tokens keep card numbers away from both the agent and the merchant [3][4]. However, neither OpenAI's announcement nor its merchant documentation says who bears the cost when an agent purchases the wrong item, misreads an instruction, or is manipulated into a bad purchase [1][3][4]. U.S. consumer protection under Regulation E and Regulation Z depends on whether a transaction was "unauthorized." A per-purchase biometric approval produces strong evidence that the cardholder authorized each transaction [9][10]. CSA's assessment is that, absent new network rules, per-purchase confirmation is likely to push the cost of agent errors that a human approved toward the consumer and the merchant returns process, rather than toward the issuer's fraud-dispute process. Card-side recourse does not disappear, because billing-error rights and network dispute categories for goods not received or not as described remain available, but those routes are merchant-contested rather than issuer-favorable [10][20]. American Express is a partial exception: it has pledged to absorb erroneous purchases by agents registered with it, although it has not published coverage terms, and neither it nor OpenAI has stated whether ChatGPT's agent counts as an Amex-registered agent [8].

On concentration, the wallet places the lab at four control points at once. Beyond operating the agent itself, the lab holds the signing keys that merchants and content delivery networks use to admit that agent, runs the listing service through which merchants make products discoverable, presents the consent screen, and keeps the conversational record of what the user asked for [3][4]. Competing protocols, including Google's Agent Payments Protocol (AP2), Visa's Trusted Agent Protocol, and Mastercard Agent Pay, distribute some of these functions differently [5][6][7]. Even so, the most probative evidence of what the user actually intended (the instruction, the agent's reasoning, and the confirmation screen) now sits with the agent platform rather than with the issuer or the merchant.

Organizations should take three actions now. First, treat agent-initiated purchases as a distinct transaction class in fraud, dispute, and expense controls. Second, block corporate card enrollment in consumer agent wallets until policy and evidence-retention terms are settled. Third, if they are merchants or acquirers, decide deliberately which agent signatures to admit, rather than inheriting that decision from a CDN default.

Background

From Instant Checkout to the Wallet

OpenAI's first attempt at agent commerce, Instant Checkout, launched on September 29, 2025, on the Agentic Commerce Protocol it developed with Stripe. Instant Checkout kept the transaction inside ChatGPT. It was shut down in March 2026, and PPC Land reported that Walmart had said in-chat conversion was three times lower than click-outs to its own site [3]. ChatGPT Wallet takes the opposite architectural approach. The agent drives the merchant's existing web checkout through ChatGPT's cloud browser, and the wallet supplies a network-issued token instead of the stored card number [3][4]. OpenAI states that existing checkout pages work without a wallet-specific payment integration, although, as described below, merchants are asked to adjust their bot controls. It says merchants receive a network token and can link it to the underlying account through the Payment Account Reference [4].

At launch, the wallet is limited to U.S. ChatGPT Pro subscribers aged 18 or older with eligible U.S.-issued cards, and only inside dots [1][3]. Before each purchase, the user reviews the items, total, and selected card, then approves with a check such as Face ID or a fingerprint [1]. OpenAI says existing card rewards and protections are preserved "for eligible disputes" [2]. It has also said it intends to let users delegate purchases more fully over time, using spending limits, category rules, time windows, and pre-approvals, but it has not given a date [1][3]. That planned step matters most for liability, because it would remove the per-purchase human confirmation that currently anchors the authorization record. Because the sources available at this writing were published within days of launch, terms, partner lists, and network rules should be expected to change quickly.

OpenAI's partner list shows how the ecosystem is assembling around the product. Fraud-prevention vendors Forter, Riskified, Sardine, and Signifyd, and payment service providers Adyen, Checkout.com, and Stripe, are working to help their systems recognize legitimate agent purchases [3]. OpenAI's own merchant guidance acknowledges that agent purchases may still be flagged by risk systems, causing declines or cancellations [3][4]. Merchants are asked to configure CDNs and firewalls to recognize ChatGPT's requests signed with Web Bot Auth, and to ingest user signals that OpenAI relays in the HTTP `Forwarded` header [2][4]. According to Arkose Labs findings published in October 2026 and

repeated in launch coverage, ChatGPT's agent signed 78.4% of its sessions with Web Bot Auth, compared with 2.3% across all legitimate agent sessions Arkose observed [3]. The reported figures do not include a measurement window or sample size, so they indicate that OpenAI signs far more consistently than the field rather than anything about market share.

The Competing Protocol Landscape

The wallet arrives among several overlapping industry efforts, each of which assigns trust and evidence differently. Table 1 summarizes the main approaches relevant to loss allocation.

Table 1. Agent payment approaches and where they place authorization evidence

Approach	Sponsor	Core mechanism	Where authorization evidence lives
ChatGPT Wallet	OpenAI with Amex, Mastercard, Visa	Per-purchase biometric approval; network token per checkout; Web Bot Auth for agent identity [1][3][4]	Agent platform (consent screen, conversation); network (token issuance)
Agent Payments Protocol (AP2)	Google and 60+ partners	Cryptographically signed Intent and Cart Mandates as verifiable credentials [5]	Portable signed mandates that can be shared among parties
Trusted Agent Protocol	Visa with Cloudflare	Agent-specific cryptographic signatures carrying agent intent and consumer recognition [6]	Request signatures presented to the merchant
Agent Pay	Mastercard	Registered, verified agents using Mastercard Agentic Tokens [7]	Network agent registry and token controls
Agentic commerce kit and purchase protection	American Express	Credentials issued only to registered agents; issuer covers erroneous purchases by registered agents [8]	Issuer (through its dispute process)

The Google Cloud announcement framed AP2 explicitly around accountability. It describes the chain of intent, cart, and payment mandates as a "non-repudiable audit trail" intended to establish responsibility when a fraudulent or incorrect transaction occurs [5]. American Express is the only network in this set that has publicly committed to absorbing a defined class of agent errors. In April 2026 it pledged to protect card members from erroneous transactions made by agents registered with Amex and to handle them through its existing dispute process [8]. Amex has not published coverage limits or a definition of "erroneous" [8], and no public statement establishes whether ChatGPT's agent is registered under that program.

Security Analysis

Authorization Law Was Built for a Different Failure Mode

U.S. consumer payment protections turn on whether a transaction was authorized. Under Regulation E, an unauthorized electronic fund transfer is one "initiated by a person other than the consumer without actual authority." The definition excludes transfers initiated by a person "who was furnished the access device to the consumer's account by the consumer," unless the consumer has notified the institution that the person is no longer authorized [9]. Regulation Z, which governs credit cards, defines unauthorized use as use by a person "who does not have actual, implied, or apparent authority," and caps cardholder liability for genuinely unauthorized use at $50 [10]. The two regimes thus share a structural feature: Regulation E's exclusion for a person furnished the access device and Regulation Z's reference to "apparent authority" both treat someone the consumer equipped to transact as acting with authority. Neither regulation contemplates a software agent acting for the consumer. A Fenwick analysis from April 2026 concluded that it remains unresolved whether granting an agent access to payment credentials satisfies Regulation E's authorization requirements, and what happens when an agent violates the consumer's instructions [11].

ChatGPT Wallet's design sidesteps part of this ambiguity. It protects consumers against account takeover, but it also gives the issuer strong evidence of authorization. Because the human approves every purchase with a device biometric, the issuer has strong evidence that the cardholder authorized the specific amount, merchant, and card [1]. A dispute claiming "I did not authorize this" will be hard to sustain when the authorization record shows a Face ID approval of the exact cart. CSA's inference is that, at least during the confirmation-per-purchase phase, most agent errors will be treated as authorized transactions with a quality or fulfillment problem. That means the merchant's returns policy will often be the consumer's main recourse, rather than an issuer's fraud chargeback. Authorized transactions are not without card-side recourse: Regulation Z billing-error rights, including for goods not accepted or not delivered as agreed, its claims-and-defenses provision, and network dispute categories for goods not

received or not as described all remain available [10][20]. What changes is that the dispute moves from the issuer-favorable "unauthorized" track to merchant-contested categories where the evidence asymmetry described below matters most.

The harder cases arise when the human approval is real but uninformed. An agent browsing the open web can be steered by injected instructions in product pages, reviews, or search results to choose a different seller, add an item, or select a look-alike product. If the confirmation screen shows the manipulated cart and the user approves it quickly, the transaction is authorized on paper even though the user's intent was subverted. CSA's earlier analysis of banks blocking personal agents noted that whether a prompt-injection attack counts as fraud or as an agent exceeding its authority is unresolved, and that CSA is not aware of any court decision applying these provisions to AI agents [15]. Table 2 sets out how current rules appear to allocate loss across common scenarios. It reflects CSA's reading of public rules and announcements, not legal advice.

Table 2. Likely loss allocation for agent wallet scenarios under current public rules

Scenario	Authorization evidence	Likely loss bearer today	Confidence
Attacker takes over the user's agent account but cannot pass the device biometric	Biometric missing	Purchase should fail; no loss	High
Attacker controls the user's unlocked device and approves purchases	Biometric present	Issuer under Reg Z/E if reported (subject to Reg E notice-timing tiers for debit cards), but contested because evidence looks authorized [9][10]	Medium
Agent picks the wrong size or item; user approves the cart without noticing	User approved exact cart	Consumer, mitigated by merchant returns and possible "not as described" dispute categories [20]	Medium

Scenario	Authorization evidence	Likely loss bearer today	Confidence
Prompt injection steers the agent to a fraudulent seller; user approves	User approved manipulated cart	Unresolved; consumer or merchant depending on network rules and seller conduct; non-delivery or misrepresentation disputes may apply [11][15][20]	Low
Future delegated purchase within pre-set limits that the user disputes	Standing delegation, no per-purchase approval	Unresolved; Reg E "furnished access device" exclusion may apply [9][11]	Low
Erroneous purchase by an agent registered with Amex	Registered agent, authenticated intent	Amex, per its purchase-protection pledge, if the agent qualifies as registered [8]	Medium (terms unpublished; ChatGPT agent's registration status not stated)
Merchant fails to deliver or delivers defective goods	Normal authorization	Merchant, through existing dispute rights [2][10][20]	High

Evidence Concentration and the Dispute Record

In a human checkout, the evidence needed to resolve a dispute is spread across parties the payments system already trusts. These include the issuer's authentication logs, the merchant's order and delivery records, and the cardholder's account. In an agent checkout, the most probative evidence is the user's original instruction, what the agent saw on the page, why it chose the item, and what the confirmation screen displayed. That evidence sits with the agent platform. OpenAI's merchant documentation relays network-level signals such as IP prefix, locale, and user agent through the `Forwarded` header, but it does not describe any mechanism for sharing the conversation or decision trace with issuers or merchants during a dispute [4]. The documentation does not mention chargebacks, disputes, or liability [4].

This creates an asymmetry. Issuers and merchants must decide disputes without seeing the record that would show whether the agent followed instructions, whether the page was manipulated, or whether the user approved a materially accurate summary. AP2's signed mandates are one attempt to make intent evidence portable and verifiable by all parties [5]. Chargebacks911, a dispute-management vendor, has argued that granular permission scopes recorded at delegation time and continuous logging across the transaction journey are prerequisites for resolving agent disputes [12]. Until the platform-held record can be produced on request, in a standard format and with integrity protection, loss is likely to fall on whichever party has the weakest evidentiary position. In CSA's assessment, that will often be the merchant on fulfillment disputes and the consumer on authorization disputes, although regulatory burden-of-proof rules and network policy can shift this outcome.

Gatekeeper Concentration at Four Control Points

The wallet also concentrates control in ways that matter for security architecture as well as competition. First, the lab controls agent identity. Merchants are asked to admit traffic based on the lab's Web Bot Auth signatures, so the lab's signing keys effectively become an allow-list credential across every merchant that trusts them [3][4]. A compromise or misuse of those keys would be systemic rather than local, although key rotation and revocation through published key directories would bound the exposure window. Second, the lab controls discovery through its product listing and feed service, which shapes which merchants an agent considers [3]. Third, it controls the consent interface, so it decides what summary the user sees before approving. Fourth, as described above, it holds the dispute evidence.

None of these is unique to OpenAI. Any frontier lab operating a high-volume consumer agent with a native wallet would hold the same four positions. Regulators have treated control over wallet access as a competition matter before. In July 2024, the European Commission made binding Apple's commitments to give third-party wallets free access to iPhone NFC payment hardware for ten years, after preliminary concerns that Apple had reserved that access for Apple Pay [14]. The parallel is imperfect, since agent wallets ride on open web checkouts rather than proprietary hardware. However, the policy question is similar: whether the operator of a widely used agent interface can favor its own payment path or set the terms on which others reach the user. The UK government's July 2026 consultation on modernising payment services asks whether consent, authentication, and unauthorised-transaction liability rules need updating for agentic payments. It closed on October 6, 2026, and no government response had been published as of this writing [13]. CSA has previously documented a related fragmentation dynamic in banking, where agent access is being negotiated institution by institution through private protocols in the absence of settled open-banking rules [15].

Enterprise Exposure

Enterprises face these issues even if they never deploy a consumer wallet. Employees may enroll corporate or purchasing cards in personal agent wallets, which brings agent-initiated spend into expense and procurement systems that typically cannot yet distinguish it from human purchases. Merchants and acquirers must decide whether to admit lab-signed agent traffic. Admitting it uncritically delegates part of their bot-defense posture to the lab, while blocking it creates declines that OpenAI says it is working with payment providers to reduce [3][4]. Builders of competing agents face a separate supply-chain risk. CSA's recent analysis of the x402 agent-payment SDK found that a default client with no spending limit allowed a malicious API to request an agent's full token balance, and that the fixes were not published as security advisories [17]. Payment logic embedded in agent SDKs can act as the authorization layer, whether or not anyone designed it to.

Recommendations

Immediate Actions

Enterprises should update acceptable-use and corporate card policies so that corporate, purchasing, and travel cards cannot be enrolled in consumer agent wallets. Card programs that support token-requestor or wallet-provisioning controls can enforce this where available. Security teams should add agent wallets to their inventory of non-human identities with spending authority, consistent with CSA's guidance that agents holding financial credentials be treated as high-value identities [15]. Merchants should review CDN and bot-management configurations to confirm whether Web Bot Auth–signed agent traffic is being admitted, blocked, or silently challenged. They should record that as an explicit decision with an owner. Fraud and dispute teams at merchants and issuers should begin tagging transactions that carry agent signals so that agent-specific dispute rates can be measured rather than estimated.

Short-Term Mitigations

Over the next one to two quarters, merchants should define a retention and production process for agent-related order evidence. This includes signed request headers, relayed `Forwarded` signals, Payment Account Reference values, and order-confirmation snapshots, because these are likely to be the merchant's main defense in agent disputes [4]. Issuers and acquirers should press agent platforms for a documented, integrity-protected way to obtain the user's instruction and the approved cart summary when a dispute is filed. Signed mandates of the kind AP2 defines are one model [5]. Organizations building their own purchasing agents should enforce per-request price expectations and hard spending

caps in code rather than relying on SDK defaults [17]. They should also show users a confirmation that compares the cart against the original instruction instead of displaying the cart alone. Risk teams should recalibrate fraud models that treat consistent, machine-like behavior as a fraud signal, because OpenAI's merchant guidance acknowledges that legitimate agent purchases may be declined by risk systems tuned for human behavior, and Chargebacks911 has urged merchants to review fraud rules and dispute thresholds for agentic payments [3][4][12].

Strategic Considerations

The liability picture is likely to change sharply when wallets move from per-purchase confirmation to standing delegation with pre-set limits, which OpenAI has signaled as its direction [1][3]. At that point the Regulation E "furnished access device" exclusion and Regulation Z's "apparent authority" language become directly relevant, and current consumer protections may not reach agent errors [9][10][11]. Organizations should engage with network rulemaking, the UK consultation outcome [13], and EU payment reforms to push for three properties: delegation scopes recorded at grant time, portable dispute evidence, and an explicit allocation for injection-induced purchases. Procurement and architecture teams should also avoid designing commerce integrations that assume a single agent platform's identity scheme. They should favor interoperable signatures and mandates that let multiple agents reach the checkout on equal terms. Finally, risk committees should track agent-platform signing-key compromise as a concentration risk. If a large share of merchant allow-lists trusts a small number of lab keys, a single key incident could disrupt commerce across many merchants at once.

CSA Resource Alignment

The most directly relevant prior CSA work is the November 2026 research note *Banks Block Personal AI Agents: Liability, Credentials, and Fragmentation* [15]. It analyzed the bank-account side of the same problem: customers giving agents credentials, account agreements shifting risk to customers, and the Regulation E exclusion for furnished access devices. ChatGPT Wallet can be read as the card-network answer to the credential-sharing problem identified there. Tokenization and agent signatures replace password sharing, which addresses the attribution gap that note described. This note's finding is that the liability gap moves rather than closes. It shifts from "which agent acted" to "what the human actually approved, and who holds the record."

CSA's *Agentic AI Liability: A Live Legal Battleground* [16] surveyed the unsettled legal landscape for agentic AI across courts, legislatures, and insurers. Agent payments are a concrete instance of that landscape, and one where the economic stakes are small per incident and potentially high in volume. That

combination is likely to make default allocation rules, rather than litigation, the deciding factor. CSA's *AI Liability Inflection: Enterprise Accountability in the Agentic Era* [21] examines how accountability for autonomous agents concentrates on particular parties, which is the same dynamic this note traces in agent payments. *x402 Payment SDK Flaws: The Missing Advisory Trail* [17] covers the developer-tooling side of agent payments. It shows that spending controls in agent payment SDKs function as security controls and should be tracked as such.

For control mapping, the AI Controls Matrix (AICM) v1.1 [18] provides the relevant control objectives. These include its Identity and Access Management (IAM) domain for agent identities and delegated spending authority, its Logging and Monitoring (LOG) domain for dispute-grade evidence retention, and its supply-chain and third-party (STA) domains for dependence on agent platforms and payment SDKs. CSA's MAESTRO agentic threat-modeling framework [19] is suited to modeling the injection-to-purchase path described in this note across the agent, tool, and ecosystem layers.

References

[1] PYMNTS. "OpenAI Gives AI Agents a Wallet While Blocking Them From Seeing the Card." PYMNTS, November 13, 2026.

[2] OpenAI Developer Community. "Introducing ChatGPT Wallet." OpenAI, November 13, 2026.

[3] PPC Land. "OpenAI hands its AI agents a wallet without the real card number in it." PPC Land, November 2026.

[4] OpenAI. "Accept payments – Commerce Partner Playbook." OpenAI Developers, November 2026.

[5] Google Cloud. "Announcing Agent Payments Protocol (AP2)." Google Cloud Blog, September 16, 2025.

[6] Visa. "Visa Introduces Trusted Agent Protocol: An Ecosystem-Led Framework for AI Commerce." Visa Inc., October 14, 2025.

[7] Mastercard. "Mastercard unveils Agent Pay, pioneering agentic payments technology to power commerce in the age of AI." Mastercard, April 29, 2025.

[8] Fortune. "American Express unveils tools for AI developers—and pledges to pay the price when agents go awry." Fortune, April 14, 2026.

[9] Consumer Financial Protection Bureau. "12 CFR 1005.2 – Definitions (Regulation E)." CFPB, accessed November 15, 2026.

[10] Consumer Financial Protection Bureau. "12 CFR 1026.12 – Special credit card provisions (Regulation Z)." CFPB, accessed November 15, 2026.

[11] Klungness, A., Arruda Rosenthal, J., Louie, A., and Voloshin, I. "Is 2026 the Year of Agentic Payments?" Fenwick & West LLP, April 22, 2026.

[12] The Paypers. "Chargebacks911 warns of dispute risk gap in agentic payments." The Paypers, May 15, 2026.

[13] HM Treasury. "Modernising payment services regulation." GOV.UK, July 14, 2026.

[14] European Commission. "Commission accepts commitments by Apple opening access to 'tap and go' technology on iPhones." European Commission, July 11, 2024.

[15] Cloud Security Alliance. "Banks Block Personal AI Agents: Liability, Credentials, and Fragmentation." CSA AI Safety Initiative, November 2, 2026.

[16] Cloud Security Alliance. "Agentic AI Liability: A Live Legal Battleground." CSA AI Safety Initiative, July 3, 2026.

[17] Cloud Security Alliance. "x402 Payment SDK Flaws: The Missing Advisory Trail." CSA AI Safety Initiative, October 26, 2026.

[18] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, June 22, 2026.

[19] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 6, 2025.

[20] Consumer Financial Protection Bureau. "12 CFR 1026.13 — Billing error resolution (Regulation Z)." CFPB, accessed November 15, 2026.

[21] Cloud Security Alliance. "AI Liability Inflection: Enterprise Accountability in the Agentic Era." Cloud Security Alliance, accessed November 15, 2026.