

StratEdge Agent Payment Risk Index 2026

StratEdge Workflow Systems · Updated 2026-09-25 · Next update 2026-12

Quarterly. Only verified, sourced items. Research findings are not reported as production losses.

Top findings

- 1 Confirmed production-loss totals for AI-agent payments are still scarce. The clearest public agent case is a July 2026 Zscaler test in which models executed a payment in a sandbox; victim losses from that campaign are not established.
- 2 Published 2026 x402 papers document protocol flaws (replay, allowance overdraft, settlement races, facilitator rule failures) in testbeds and deployment reviews. Those papers do not publish a single production-loss dollar total.
- 3 The largest sourced automated-system loss in this index remains Knight Capital on 1 August 2012: the company reported about \$440 million; the SEC later stated the loss was more than \$460 million.
- 4 Address poisoning is a separate, non-agent category. A USENIX Security 2025 measurement found at least \$83.8 million across 6,633 incidents through 30 June 2024, and explicitly excluded a May 2024 wrapped-bitcoin case from that total.
- 5 Controls that map to these sources: intent checks on untrusted content, hard quantity and notional limits, new-address and beneficiary-change holds, dual control, and an audit record before settlement.

Loss by category

Automated trading malfunction

About \$440 million (company, 2 Aug 2012). SEC order (2013): more than \$460 million.

Single incident. Not an AI-agent case.

Address poisoning (measurement)

At least \$83.8 million across 6,633 incidents, Ethereum and BSC, through 30 June 2024.

Not classified by the authors as AI-agent incidents. May 2024 wrapped-bitcoin transfer is outside this dollar total.

Address poisoning (single case, mostly returned)

About \$68 million sent on 3 May 2024 and later returned. Chainalysis reports about \$1.49 million still retained after the return.

Do not add this to the \$83.8 million measurement. The measurement excluded this asset.

AI-agent prompt injection

No established victim-loss total in the source. Sandbox test executed a payment framed as about \$3 (about 0.0012 ETH).

Zscaler ThreatLabz, 2 July 2026.

x402 protocol research

No production-loss total stated in the papers indexed here.

Testbed and deployment-review findings only.

Five recommended controls

- 1 Check intent against the principal's instruction, not against text the agent retrieved.
- 2 Enforce hard quantity and notional limits before any rail is called.
- 3 Block or hold first payments to a new address and any beneficiary-detail change.
- 4 Require a second approver above threshold and for counterparty changes.
- 5 Write the audit record before settlement, including the rules that fired.

Entries

2026-07-02 — Hidden page instructions and agent payments (Zscaler ThreatLabz)

AI-agent incident research

ThreatLabz described fraudulent sites that hide instructions telling a browsing agent to pay for a fake API license. In their sandbox, 4 of 26 models executed the payment. They also described a typosquat that some models treated as a known crypto site. Testing used a sandbox. The report does not establish a victim-loss total for the campaign.

Loss: Not established. The payment framed in the test was about \$3, described as about 0.0012 ETH.

Cause: The agent treated untrusted page content as an instruction to pay.

Control: Intent mismatch against the original task, plus a block on payees that are not allowlisted.

Source: Zscaler ThreatLabz, 2 July 2026 — <https://www.zscaler.com/blogs/security-research/indirect-prompt-injection-web-content-targets-ai-agents>

2026 — Five attacks on the x402 payment protocol (published research)

Protocol research — not a reported production loss

The authors analyze x402 and report five attack classes against authorization, binding, replay protection, and web-layer handling. They describe a testbed of more than 25,000 payment requests. This index does not repeat exploit steps.

Loss: Not a reported production-loss figure.

Cause: Payment authorization was not bound tightly enough to the resource and the settlement.

Control: Bind each payment to the resource and the intent, and refuse replayed or substituted authorizations before settlement.

Source: arXiv:2605.11781 — <https://arxiv.org/html/2605.11781v1>

2026 — Free-riding analysis of x402 payments (published research)

Protocol research — not a reported production loss

The authors report four flaw classes, including cross-resource substitution, duplicate-settlement races, allowance overdraft, and denial of settlement, against official SDKs and a production deployment in their tests. They state findings were disclosed.

Loss: Not a reported production-loss total. Their tests report resource-leakage ratios, which are not customer-loss statistics.

Cause: Spending allowance and settlement state could diverge.

Control: Hard spending limits re-checked at settlement time, with one-time authorization.

Source: arXiv:2605.30998 — <https://arxiv.org/html/2605.30998v2>

2026 — Facilitator rule review of live x402 deployments (published research)

Protocol research — not a reported production loss

The authors say they tested 15 x402 facilitators and found rule violations in all of them, then disclosed the issues. They write that affected parties acknowledged issues and that mitigations included changes by Coinbase. They do not publish a dollar loss for a named incident.

Loss: No incident dollar total in the paper.

Cause: Shared facilitator checks did not enforce the authors' settlement and authorization rules.

Control: A policy check before the agent pays, independent of the facilitator.

Source: arXiv:2607.19545 — <https://arxiv.org/html/2607.19545>

2022-07-01/2024-06-30 — Address-poisoning measurement on Ethereum and BSC

Automated fraud measurement — not an AI-agent incident

Tsuchiya, Dong, Soska, and Christin measure poisoning transfers that plant lookalike addresses in transaction history. Across Ethereum and BSC through 30 June 2024 they report 6,633 successful incidents and at least \$83.8 million in losses. They note the estimate is a lower bound and that a well-known wrapped-bitcoin case is not inside this dollar figure.

Loss: At least \$83.8 million across 6,633 incidents.

Cause: A payment went to a lookalike address taken from recent history.

Control: Block or hold the first payment to an address that is not allowlisted, and require counterparty verification.

Source: USENIX Security 2025, Tsuchiya, Dong, Soska, and Christin — <https://www.usenix.org/system/files/usenixsecurity25-tsuchiya.pdf>

2024-05-03 — Wrapped-bitcoin address poisoning

Address poisoning — not an AI-agent incident

Chainalysis describes an address-poisoning scam on 3 May 2024 in which a large holder sent about \$68 million in wrapped bitcoin to a lookalike address. Chainalysis reports the attacker returned that position on 9 May 2024 and that the scam still netted about \$1.49 million after the return.

Loss: About \$68 million sent, later returned. About \$1.49 million reported as still retained after the return.

Cause: The sender used a lookalike address from poisoned history.

Control: New-address block and out-of-band verification before a large first payment.

Source: Chainalysis, Anatomy of an Address Poisoning Scam — <https://www.chainalysis.com/blog/address-poisoning-scam/>

2012-08-01 — Knight Capital automated order router

Automated system — not an AI-agent incident

On 1 August 2012 Knight's automated router sent a flood of erroneous equity orders. The company's 2 August 2012 release said the firm had traded out of the position at a realized pre-tax loss of about \$440 million. The SEC's 2013 order states Knight lost more than \$460 million after the router, processing 212 customer orders, obtained millions of executions in about 45 minutes.

Loss: About \$440 million in the company release. More than \$460 million in the SEC order.

Cause: A defective automated path kept sending orders after the intended quantity was filled. Limits did not stop it.

Control: A hard quantity and notional limit that blocks further orders, plus an audit alert a human must clear before the session continues.

Source: SEC press release 2013-222; company release filed as SEC exhibit, 2 August 2012 — <https://www.sec.gov/newsroom/press-releases/2013-222>

This index quotes published sources. It does not add losses that a source did not state. It is not investment, legal, or compliance advice.