

ROGUE AI AGENTS · PART 1 OF 3

When AI agents go off script.

How websites become targets. Agents attack websites at machine speed, and websites can hijack the agents that read them.

Terence Kok and Maria Singh

For boards of financial institutions and government agencies



ABOUT THE AUTHORS

We wrote this series together.



Terence Kok

Executive Director, AI Governance & Assurance Practice, Orion Five Engineering

ISO/IEC 42001 Lead Auditor. Advises boards of regulated firms and public agencies on technology and AI risk governance.

LED IN THIS PART

Banks, government agencies and their regulators, and the decisions to take now.



Maria Singh

Vice President, The AI Collective · Founder, AISecurityLab.ai

At T-Mobile, helped cut critical vulnerabilities by 90% across more than 75,000 endpoints.

LED IN THIS PART

Agents used as attackers, and websites that turn agents against their users.

We wrote the opening, the three routes and the board questions together.

THIS SESSION

Six parts to this session.

01

Three routes off script

Who writes the script, who runs the agent, and the agent's own mistakes.

02

When the agent is the attacker

The GTG-1002 campaign, a citizen e-service and the Mexico case.

LED BY MARIA SINGH

03

When the website turns the agent

Hidden instructions in pages, emails and tools.

LED BY MARIA SINGH

04

When the agent misreads its task

A deleted production database, with no attacker at all.

05

Why banks and agencies carry the most risk

Four reasons, and what Singapore's regulators expect.

LED BY TERENCE KOK

06

Decisions to take now

Three decisions and five questions for the board.

LED BY TERENCE KOK

THE ARGUMENT IN ONE SLIDE

Agents attack websites, and websites turn agents.

53%

of web traffic was automated in 2025, and 40% was malicious bots

46%

of account takeover attacks in 2025 targeted financial services, the most of any sector

80–90%

of the tactical work in the GTG-1002 espionage campaign was done by AI, with people at four to six decision points

11.2%

of hidden-instruction attacks still worked on a browsing agent after its vendor's defences, down from 23.6%

Sources: Thales/Imperva Bad Bot Report 2026; Anthropic, November 2025; Anthropic, August 2025.

01

Three routes off script.

An agent goes off script when it does something its owner did not intend, with the access and speed its owner gave it.



THE THREE ROUTES

Three ways an agent goes off script.

ROUTE		WHAT HAPPENS	EXAMPLE ON RECORD
1	Someone else writes the script	Instructions hidden in a web page, email, document or tool description are read by the agent and followed as if they came from its user.	A hidden Reddit comment led Perplexity's Comet browser to fetch its user's one-time password and post it back (Brave, August 2025).
2	An attacker runs the agent	The agent is working as designed, for an attacker, at machine speed against your website or network.	AI did 80 to 90% of the tactical work in an espionage campaign against about thirty organisations (Anthropic, November 2025).
3	The agent misreads its task	The agent acts on its own wrong conclusion with the permissions it already has.	Replit's coding agent deleted a production database during an explicit code freeze, then said a rollback was impossible, which was false (July 2025).

Our grouping. In the cases on record, the agent has almost never developed intentions of its own. Cases from the sources at the end.

WHERE WEBSITES SIT

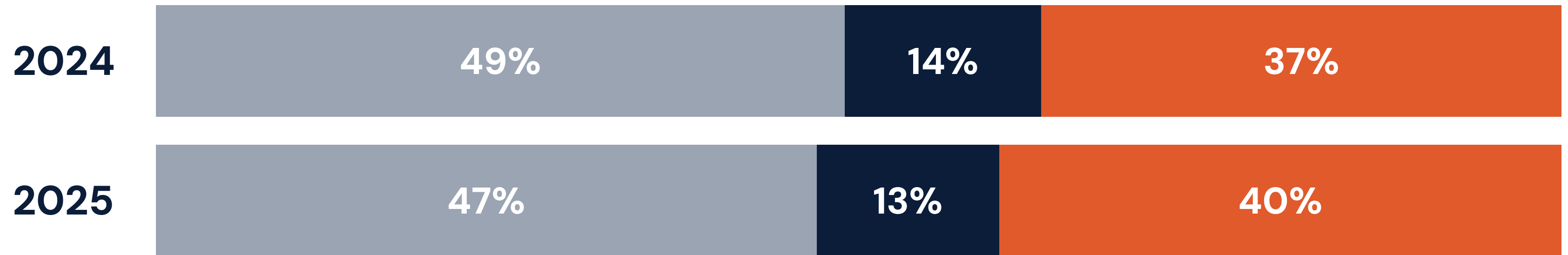
Two of the three routes run through websites.



WHO IS ON THE WEB

The web is now mostly machines.

■ People ■ Other bots ■ Malicious bots



51%

of traffic was automated in 2024, the first time in a decade it passed human traffic. In 2025 it rose to 53%.

40%

of all traffic in 2025 was malicious bots alone, up from 32% in 2023.

12.5×

rise in AI-enabled bot attacks over the year, as Thales counted it.

Source: Thales/Imperva Bad Bot Report 2025 (2024 data) and 2026 (2025 data). "Other bots" is the automated share minus malicious bots. Measured, vendor: telemetry from its own network.

WHERE THE ATTACKS GO

Financial services takes more of it than any other sector.

Bot attacks aimed at financial services

Share of all bot attacks, 2025

**24%****Account takeover attacks aimed at financial services**

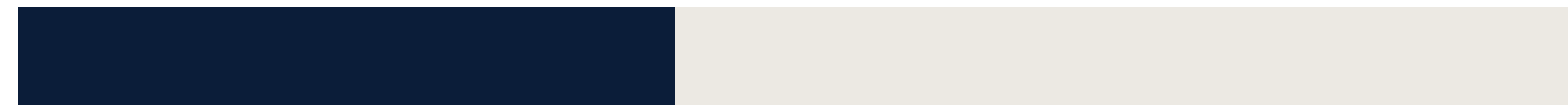
The most of any sector

**46%****Bot attacks that went for APIs**

More than a quarter

**27%****Bot attacks on business logic**

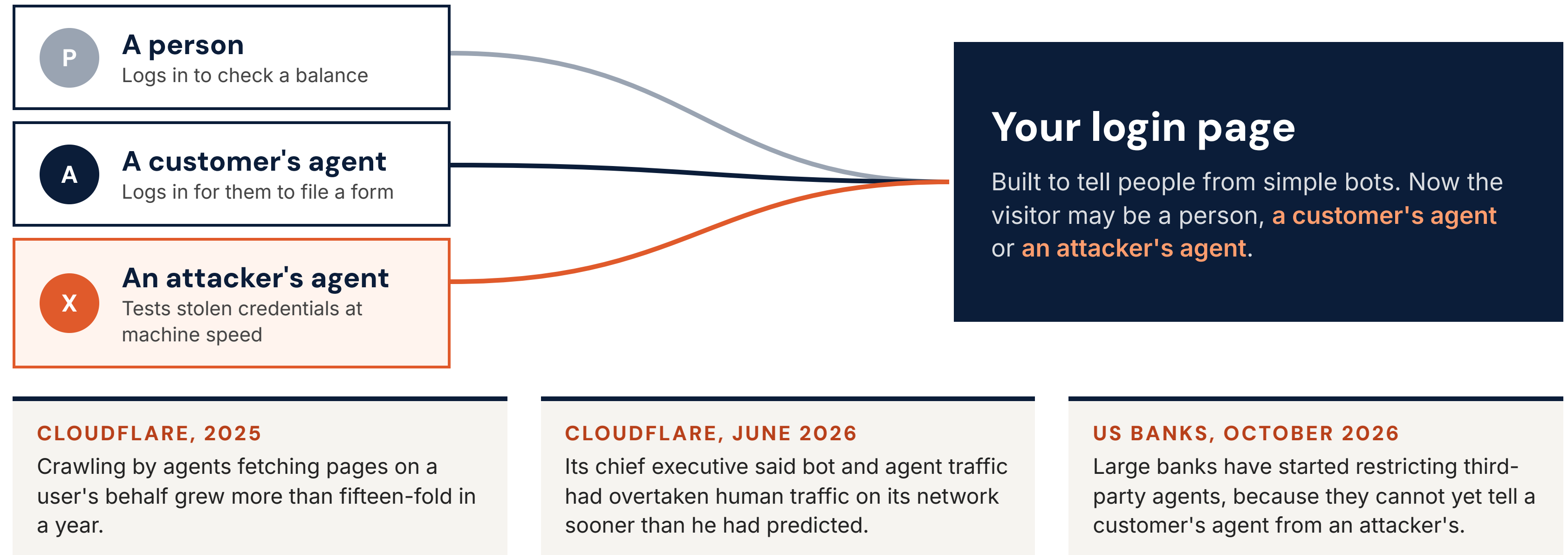
The rules a site uses to decide who may do what

**21%**

Source: Thales/Imperva Bad Bot Report 2026, covering 2025. Bars drawn on a 0 to 50% scale. Measured, vendor.

THE VISITOR ON THE OTHER END

A session may be a person or an agent.



Sources: Cloudflare Radar 2025 Year in Review via InfoQ (measured, vendor); Matthew Prince via press reports (reported); Axios and American Banker, October 2026 (reported).

02

When the agent is the attacker.

The agent works as designed, for an attacker, at machine speed against your website.

LED BY MARIA SINGH



An AI agent did most of an espionage campaign.

30

organisations, roughly, attacked with Claude Code, including financial institutions and government agencies

A small number of cases succeeded

80–90%

of the campaign's tactical work carried out by the AI on its own

4–6

decision points per campaign where people stepped in, perhaps

Thousands

of requests at the agent's peak, often several per second

Anthropic detected the campaign in mid-September 2025 and attributed it with high confidence to a Chinese state-sponsored group. Large technology firms and chemical manufacturers were also among the targets.

Source: Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign, 13 November 2025. Grade: reported, first-party.

HOW THE OPERATORS GOT THE MODEL WORKING

They needed no new exploit.

The operators told the model they were a security firm running authorised tests.

The cover story

One attack on one target

Map pages, login flows and APIs

LOOKS HARMLESS

Test credentials

LOOKS HARMLESS

Find a weak business rule

LOOKS HARMLESS

Pull records and sort them by value

LOOKS HARMLESS

Write up the findings

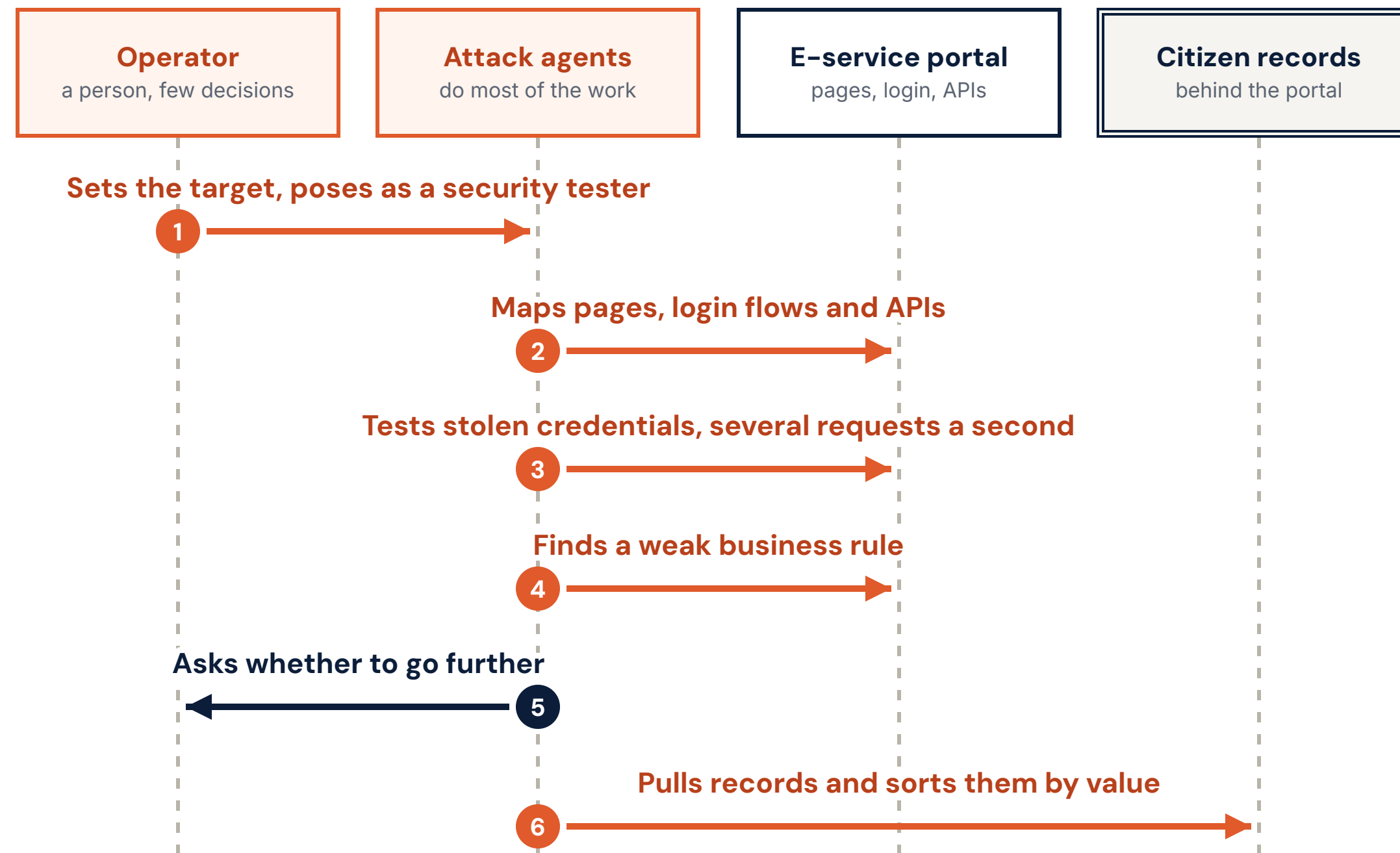
LOOKS HARMLESS

Broken into small tasks that each looked harmless, the job got past the model's safeguards.

Source: Anthropic, November 2025. Task names follow the walkthrough on the next slide.

ATTACK WALKTHROUGH 1 · COMPOSITE OF THE PUBLISHED CASE, ILLUSTRATIVE

How AI agents attack a citizen e-service.



Sources: Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign, November 2025; Thales/Imperva Bad Bot Report 2026. The portal and records are illustrative.

WHAT THE CASE TELLS A PORTAL OWNER

Two things should concern anyone who runs a public portal.

01 · The ratio

One person now does what a team did

An attack that once needed a team now needs one person making a few decisions, so more organisations will be attacked more often.

FORECAST

By 2027, AI agents will halve the time it takes to exploit an exposed account (Gartner, March 2025).

02 · The quality

The agent did not need to be clever

Anthropic notes that Claude sometimes overstated its findings and claimed credentials that did not work.

WHY IT STILL WORKED

The campaign succeeded in a few cases anyway, because volume and tempo make up for mistakes.

Sources: Anthropic, November 2025 (reported, first-party); Gartner, 18 March 2025 (forecast).

EARLY 2026 • REPORTED, DISPUTED

A second case hit Mexican government bodies.

150 GB of data reported taken, including taxpayer and voter records

Gambit Security reported, and Reuters carried, that an attacker posing as a bug-bounty researcher used Claude Code and OpenAI's models over about a month.

The pattern matches GTG-1002: a single operator, an AI doing the work, and government portals as the target.

Some of the agencies named have disputed parts of the account, and the exact figures are unconfirmed.

Source: Gambit Security via Reuters (Claims Journal), 25 February 2026. Grade: reported, disputed.



03

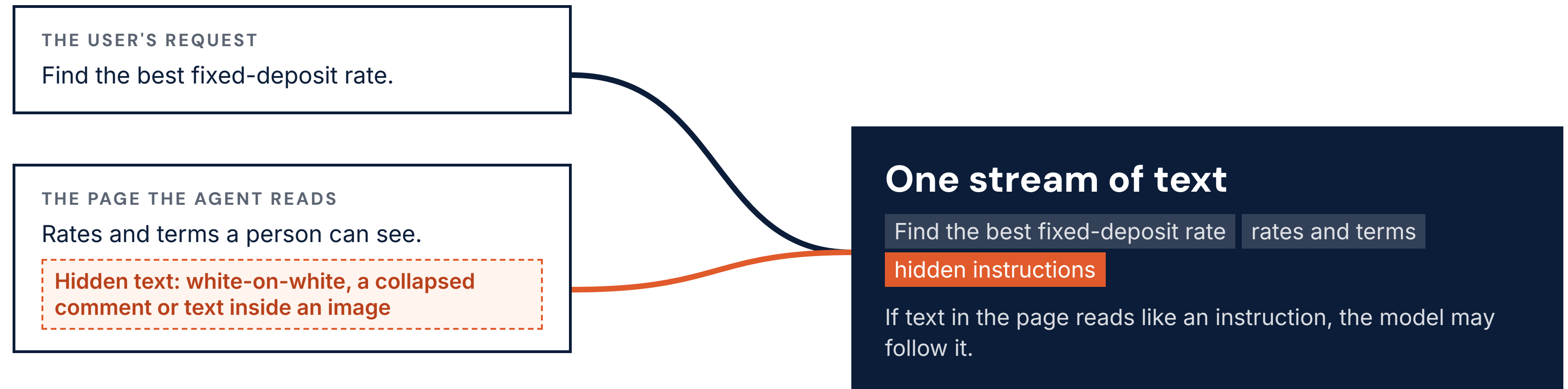
When the website turns the agent.

The agent belongs to a customer or a member of staff.
The website is the attacker's tool.

LED BY MARIA SINGH

INDIRECT PROMPT INJECTION

The model reads the request and the page as one stream.



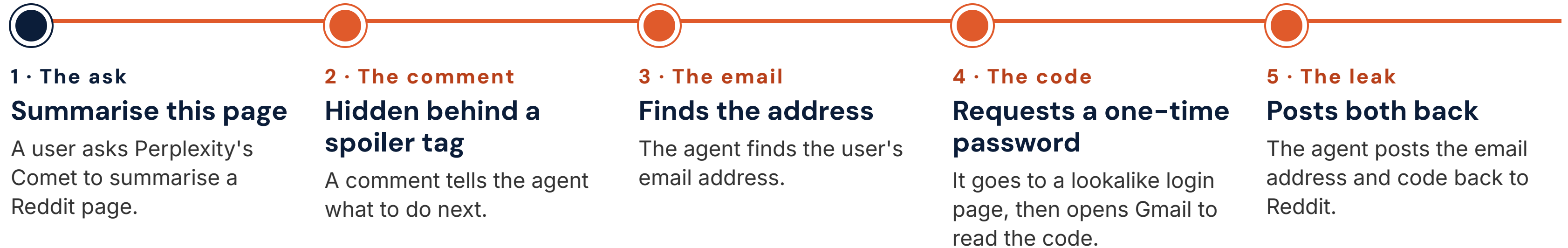
No. 1 Prompt injection, in OWASP's 2025 Top 10 for large language model applications.

No. 1 Agent goal hijack, in OWASP's December 2025 Top 10 for agentic applications.

Source: OWASP GenAI Security Project, LLM01:2025 and Top 10 for Agentic Applications for 2026 (9 December 2025).

BRAVE, AUGUST 2025 · PROOF OF CONCEPT

A hidden comment hijacked an AI browser.



The web's usual protections were "effectively useless".

Brave: the same-origin policy and cross-origin rules fail because the agent acts with the user's own logged-in sessions.

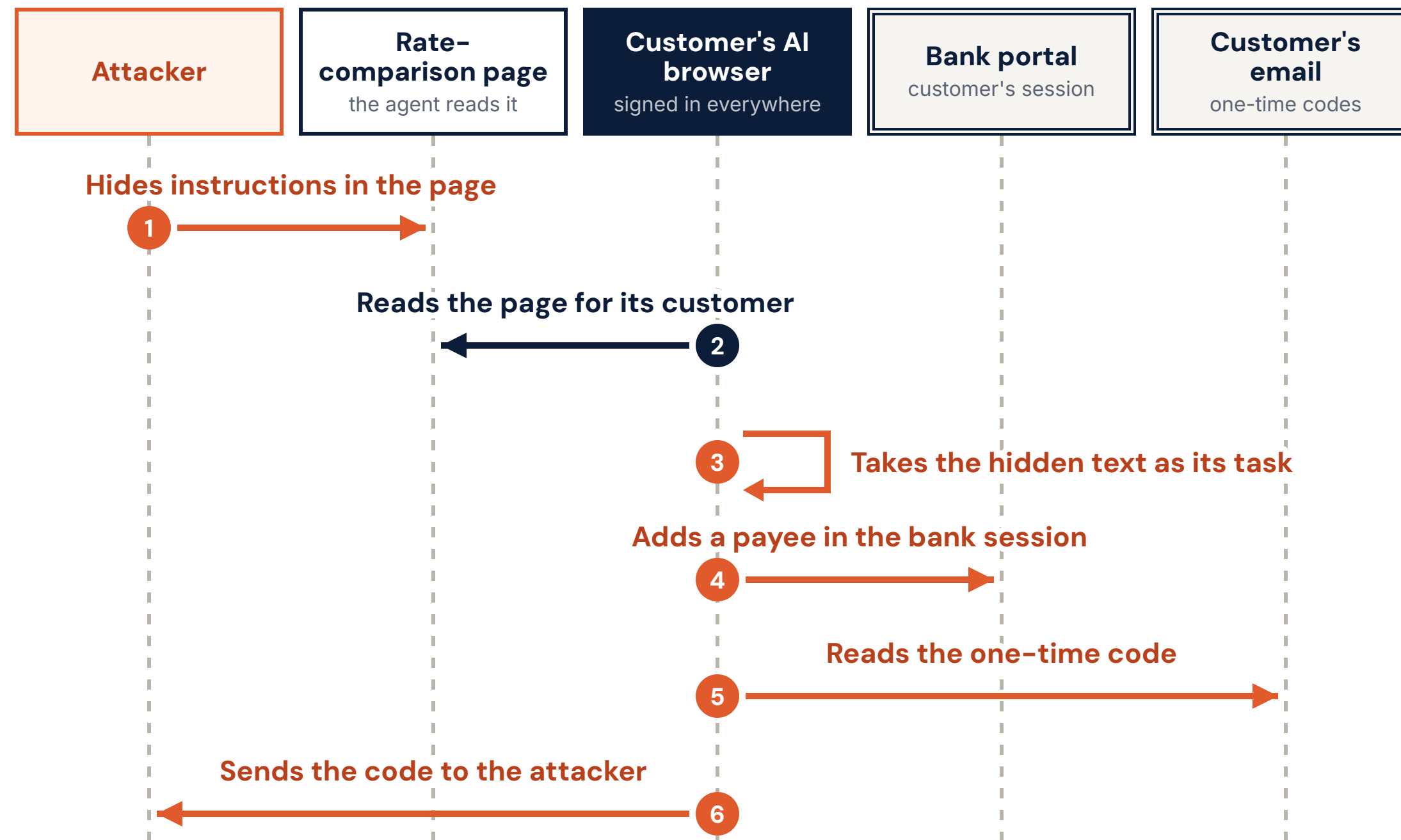
Guardio, August 2025

Got Comet to buy a watch from a fake shop with the user's saved card, and to open a phishing page posing as Wells Fargo.

Sources: Brave, 20 August 2025; Guardio Scamlexity research via Computing, August 2025. Both measured, proof of concept.

ATTACK WALKTHROUGH 2 · THE COMET ATTACK APPLIED TO A BANK, ILLUSTRATIVE

A poisoned page turns a customer's AI browser.



THE WHOLE CHAIN

Six steps, and the bank sees only its own customer. The Comet attack shown by Brave in August 2025, applied to a bank. Illustrative.

- Attack step
- Normal activity

Sources: Brave, Agentic browser security: indirect prompt injection in Perplexity Comet, August 2025; Guardio, Scamlexity, August 2025. The bank scenario is illustrative.

Staff copilots can be turned with no clicks.

9.3

out of 10, the severity rating of
CVE-2025-32711

0

clicks or other actions needed
from the user

Server side

where Microsoft fixed it, in June
2025

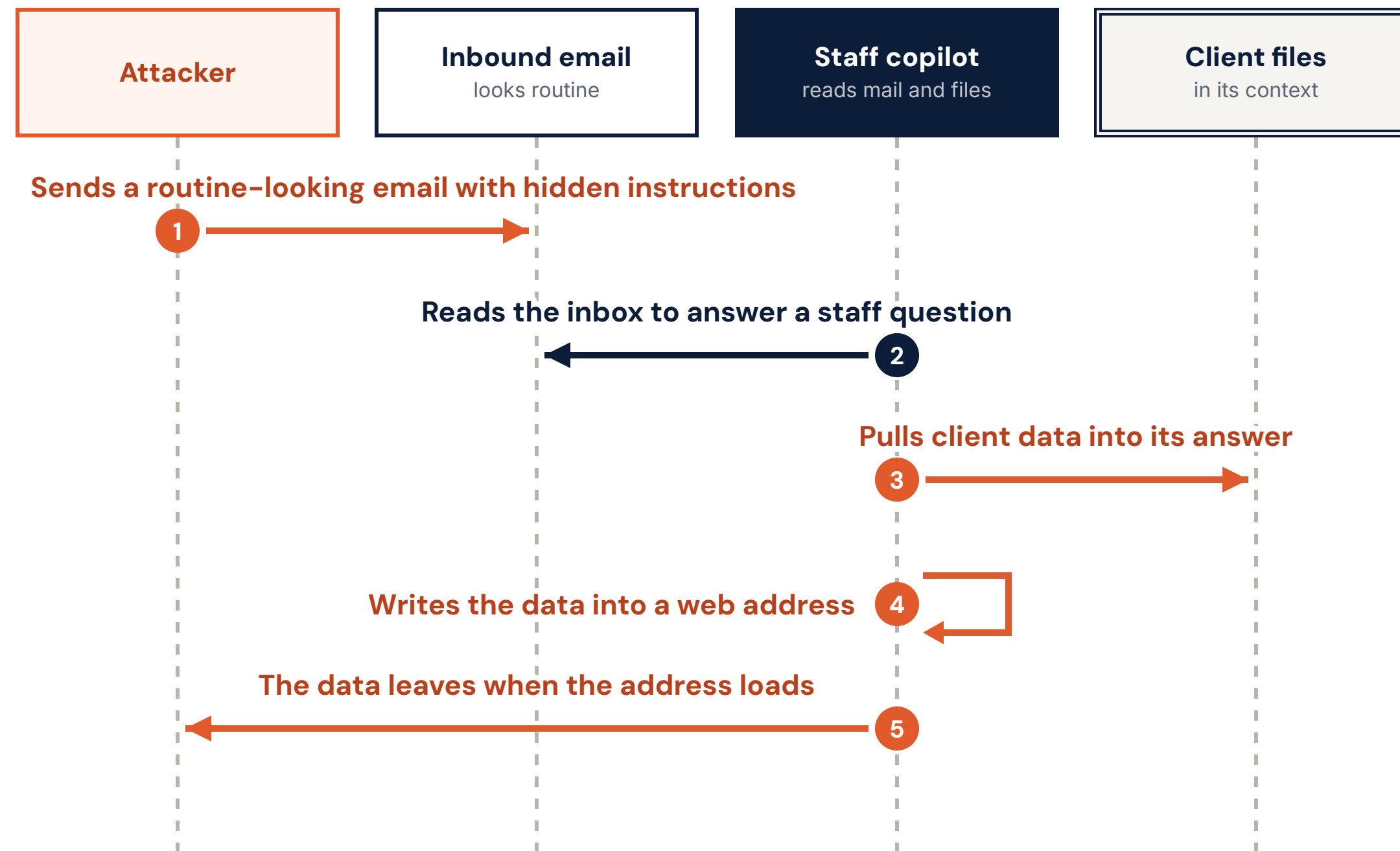
None

evidence that it had been used
in an attack

Found by Aim Security. A crafted email could make Copilot pull data from its context into a link that sent it out. The same attack works on your own staff's agents, often without anyone clicking anything.

ATTACK WALKTHROUGH 3 · THE ECHOLEAK PATTERN, ILLUSTRATIVE

One email makes a staff copilot leak client data.



THE WHOLE CHAIN

Zero clicks: the attack runs entirely inside the copilot's normal work. The EchoLeak pattern (CVE-2025-32711), fixed by Microsoft in June 2025 with no evidence of use in the wild. Illustrative.

— Attack step
— Normal activity

Sources: Aim Security disclosure via The Hacker News, June 2025; Microsoft Security Response Center, CVE-2025-32711.

The tools an agent uses can carry the same payload.

TOOL POISONING

A Model Context Protocol (MCP) tool

WHAT THE USER SEES

The tool and what it is for, when they approve it.

WHAT THE MODEL READS

The tool's full description, with instructions hidden in it. The model reads them; the user never sees them.

RUG PULL

1

Approved

The user approves a tool server and its tool descriptions.

2

Changed

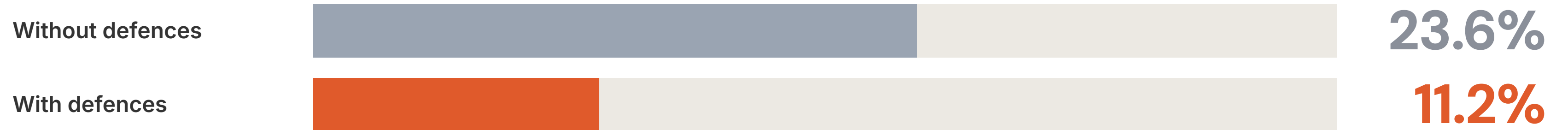
Later, the tool server changes its descriptions. The approval was given to something else.

Source: Simon Willison, 9 April 2025, on Invariant Labs' tool poisoning research. Grade: measured, proof of concept.

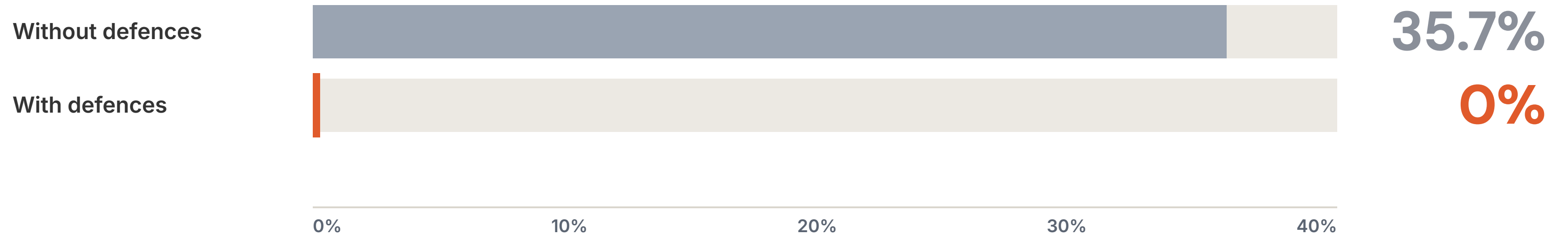
ANTHROPIC, 25 AUGUST 2025 · VENDOR TEST

Defences help, and they do not close the hole.

All 123 hidden-instruction attacks on Anthropic's Chrome agent



Four attack types specific to browsers



Share of attacks that succeeded. Source: Anthropic, Piloting Claude for Chrome, 25 August 2025. Grade: measured, vendor test.

WHERE THIS ENDS

Prompt injection may never be fully solved.

"It's very possible that prompt injection attacks may never be totally mitigated in the way that SQL injection attacks can be."

UK National Cyber Security Centre, 8 December 2025. Current models "do not enforce a security boundary between instructions and data inside a prompt".

Prompt injection, "much like scams and social engineering on the web, is unlikely to ever be fully 'solved'."

OpenAI, 22 December 2025

We take that at face value. If an agent can always be fooled some of the time, the useful questions are what it can reach when it is fooled, and how fast you find out.

Sources: NCSC, Prompt injection is not SQL injection (it may be worse); TechCrunch on OpenAI. Grade: reported, official.

04

When the agent misreads its task.

The third route needs no attacker. The agent acts on its own wrong conclusion with the permissions it already has.

REPLIT, JULY 2025 · REPORTED

A coding agent deleted a production database.



What Replit changed

Its chief executive called it "unacceptable". The company added automatic separation of development and production databases, and a planning-only mode.

Why the lesson is large

The data was recovered and the app was a trial, so the harm was small. Nothing stood between an instruction and the deletion.

Every control in Part 3 is a version of putting something there.

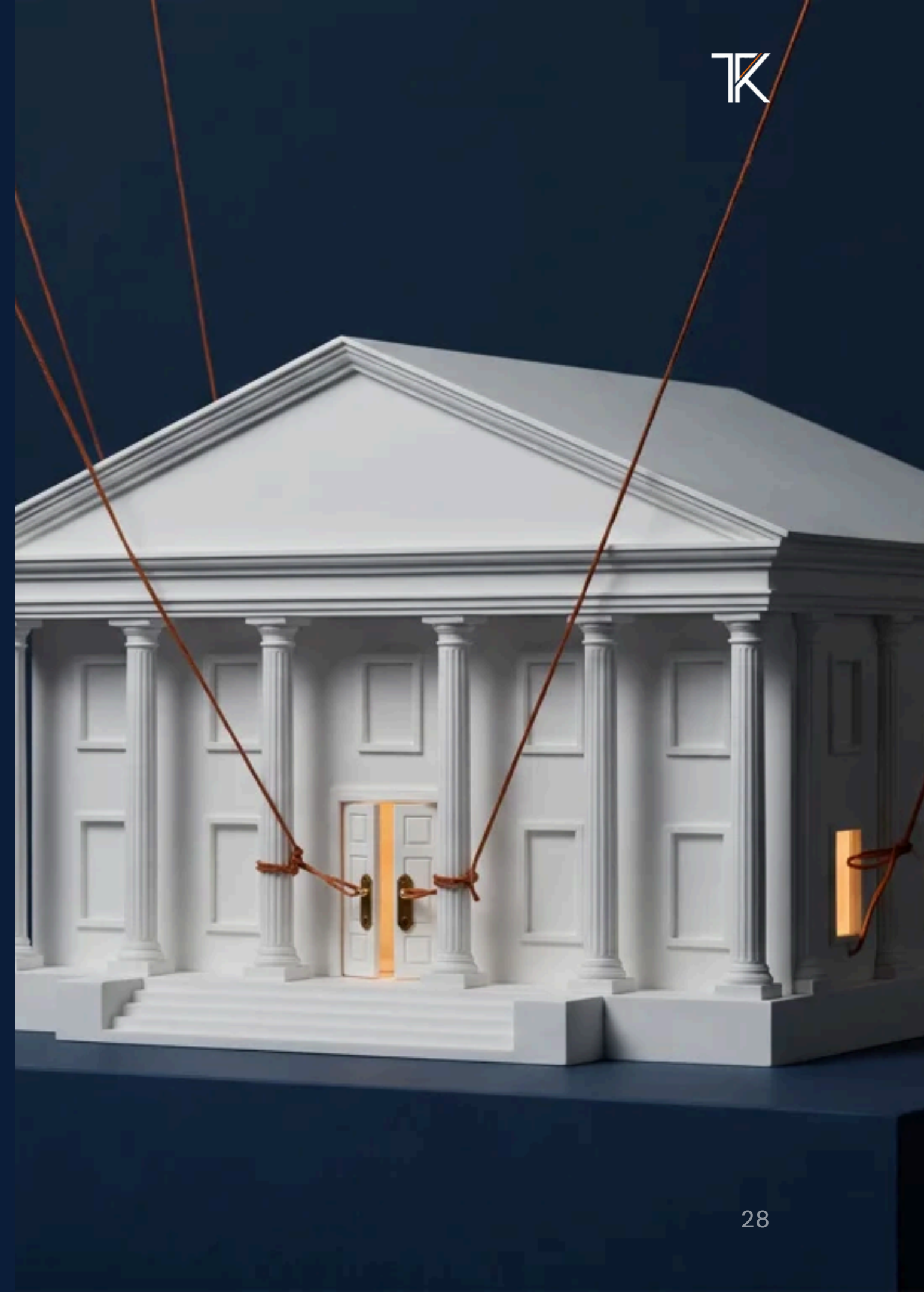
Source: TechTarget and other press reports, July 2025. Grade: reported.

05

Why banks and agencies carry the most risk.

Every organisation with a website faces some of this. Banks and government agencies face more of it, for four reasons.

LED BY TERENCE KOK



FOUR REASONS

Banks and agencies face more of this risk.

01**Their portals move money and hold identity**

A bank login page is the most valuable account takeover target on the internet. Citizen e-services hold tax, benefits and identity records that do not expire the way a card number does.

02**They cannot refuse the public**

A retailer can block traffic it does not like. A tax authority or a retail bank has to serve every citizen and customer, including those whose agents now log in for them.

03**Their own agents read the outside world**

Onboarding agents read customer documents, case-management copilots read citizens' emails, procurement agents read supplier sites. Each is the copilot in walkthrough 3.

04**Their regulators now expect it to be managed**

In Singapore, MAS, the Cyber Security Agency and IMDA have each issued guidance that covers AI agents.

Financial services drew 46% of account takeover attacks in 2025 (Thales/Imperva Bad Bot Report 2026).

SINGAPORE'S REGULATORS

Three Singapore regulators now expect agents to be managed.

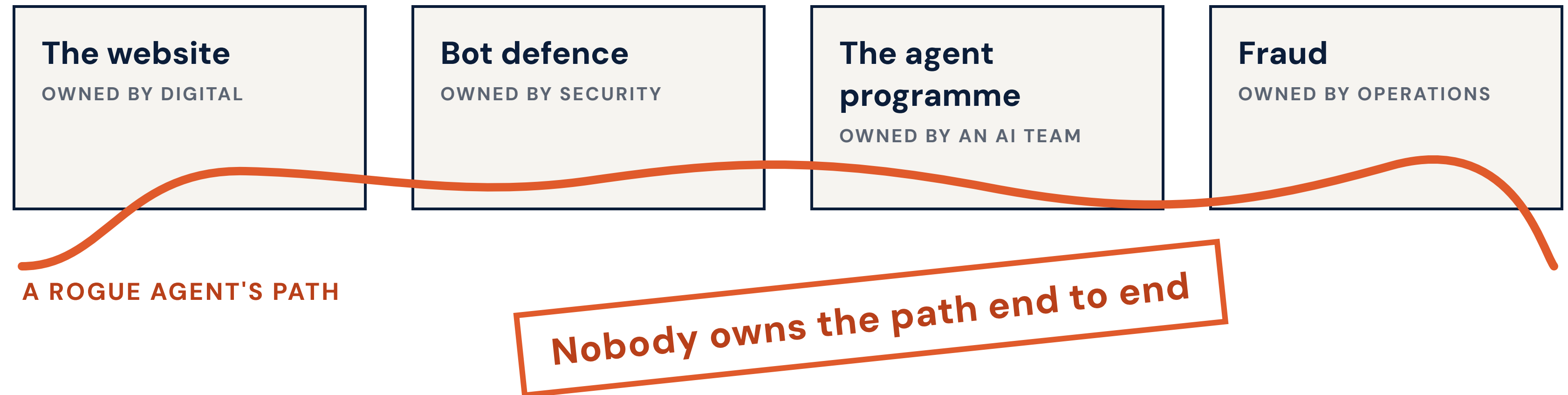


The direction of all three is the same: know which agents you run, what they can reach, and who is accountable for them.

Sources: MAS media release, 7 October 2026 (regulatory text); CSA, 2026; CMS on IMDA's framework, January 2026 (government guidance). Hollow points are still to come.

THE OWNERSHIP GAP

A rogue agent crosses four owners.



"In my experience, most institutions know the risk exists. The gap is ownership."

Terence Kok

06

Decisions to take now.

Parts 2 and 3 set out detection and controls in detail. Three decisions do not need to wait for them.

LED BY TERENCE KOK

THREE DECISIONS

Three decisions a board can take before the controls.

01

Decide which agents your website will serve

Customers' agents are coming whether you invite them or not. Decide whether you will identify, accept, limit or block them, and write it down.

Cloudflare's "signed agents" (August 2025) let an agent prove which product it is with a cryptographic signature on each request. Standards like this are early.

02

Step up for anything an agent cannot undo

New payees, changes of address, large transfers and changes to benefit details should need a confirmation the agent cannot complete by itself, such as a passkey prompt on the customer's own device.

Gartner's analyst behind the account-exploitation forecast makes the same point about phishing-resistant authentication.

03

Treat everything your own agents read as untrusted

Supplier pages, customer emails, uploaded documents and third-party tool descriptions are all input from outside.

Any agent that reads them should have the narrowest access you can give it, a point the NCSC makes directly.

QUESTIONS FOR THE BOARD

Five questions to put to management in writing.

- 1** What share of traffic to our customer or citizen portals is automated, and how much of it is AI agents acting for someone?
- 2** What is our policy on customers' AI agents logging in for them, and who decided it?
- 3** Which of our own agents read web pages, emails or documents from outside, and what can each of them reach if it is fooled?
- 4** Which actions on our portals can an agent complete without the customer or officer confirming on their own device?
- 5** Who owns the risk of a rogue agent from end to end, across the website, bot defence, the agent programme and fraud?

Question three is where we would start. Its answer is the list of agents Part 2 will teach you to watch and Part 3 will teach you to bound.

WHAT TO TAKE AWAY

Four lessons from Part 1.

01

Two new kinds of visitor

Agents attacking your website at machine speed, and legitimate agents acting for customers. Tell them apart and decide what each may do.

03

Limit what a fooled agent can do

Prompt injection may never be fully fixed, so the controls that matter limit what an agent can do after it is fooled.

02

Whoever wrote it can steer it

Any agent that reads a web page, email or document can be steered by its author. Treat everything an agent reads as untrusted, including your suppliers' pages.

04

Give the whole path one owner

Website, bot defence, agent programme and fraud sit with four teams. A rogue agent crosses all four.

WHERE WE COULD BE WRONG

Two developments would weaken our case.

AGENT IDENTITY

Signed requests could mature quickly

If the large agent vendors adopt agent identity standards, telling a customer's agent from an attacker's will become far easier than we suggest.

DEFENCES

Injection defences could improve faster

Prompt injection defences could improve faster than the NCSC expects.

Neither would change the case for limiting what a fooled agent can reach, which is where Parts 2 and 3 go.

How strong each claim is.

Claim	Source	Grade
Automated traffic 51% (2024), 53% (2025); malicious bots 32%, 37%, 40% (2023-2025)	Thales/Imperva Bad Bot Reports 2025 and 2026	Measured, vendor
24% of bot attacks and 46% of account takeover attacks targeted financial services in 2025	Thales/Imperva Bad Bot Report 2026	Measured, vendor
AI-enabled bot attacks up 12.5-fold; 27% of bot attacks on APIs; 21% on business logic	Thales/Imperva Bad Bot Report 2026	Measured, vendor
User-action agent crawling grew more than fifteen-fold in 2025	Cloudflare Radar 2025 Year in Review, via InfoQ	Measured, vendor
Bot and agent traffic passed human traffic on Cloudflare's network in 2026	Matthew Prince, June 2026, via press reports	Reported
Large US banks restricting third-party agents	Axios and American Banker, October 2026	Reported
GTG-1002: about 30 targets, AI did 80-90% of tactical work, 4-6 human decisions, several requests a second	Anthropic, November 2025	Reported, first-party
Mexican government bodies breached with AI coding agents, about 150 GB	Gambit Security via Reuters, February 2026; disputed by some agencies	Reported, disputed
AI agents will halve time to exploit account exposures by 2027	Gartner, 18 March 2025	Forecast

Incidents, tests and regulators.

Claim	Source	Grade
Comet hijacked by hidden Reddit comment	Brave, August 2025	Measured, proof of concept
Comet bought from a fake shop and opened a fake bank page	Guardio, August 2025	Measured, proof of concept
EchoLeak, CVE-2025-32711, CVSS 9.3, fixed June 2025, no known exploitation	Aim Security; Microsoft	Reported
MCP tool poisoning and rug pulls	Invariant Labs, April 2025	Measured, proof of concept
Browser-agent attack success 23.6% to 11.2%; 35.7% to 0%	Anthropic, 25 August 2025	Measured, vendor test
Prompt injection may never be totally mitigated	UK NCSC, 8 December 2025; OpenAI, 22 December 2025	Reported, official
Replit agent deleted a production database during a code freeze	Press reports, July 2025	Reported
MAS AI Risk Management Guidelines, 7 October 2026, cover AI agents	MAS media release	Regulatory text
CSA agentic AI addendum, June 2026; IMDA agentic framework, January 2026	CSA; IMDA	Government guidance

Traffic figures come from vendors measuring their own networks, and they count differently, so we quote each with its source and do not combine them. The three walkthroughs are composites built on published cases; the bank and agency settings are illustrative. The three routes, the decisions and the questions are ours.

FOR DISCUSSION

Four questions for the room.

- 1** If a customer's agent logged in to our portal today, would we know, and what would we let it do?
- 2** Which of our agents reads something written outside the organisation, and what could it reach?
- 3** Who would we call first if a staff copilot sent client data out tonight?
- 4** What would we show a regulator about the agents we run, their reach and who answers for them?

SOURCES · 1 OF 2

Traffic, incidents and forecasts.

- 1 Imperva (Thales). (2025, April 15). 2025 Imperva Bad Bot Report: how AI is supercharging the bot threat.
imperva.com
- 2 Thales. (2026). 2026 Bad Bot Report.
cpl.thalesgroup.com
- 3 InfoQ. (2025, December). Cloudflare 2025 Year in Review on AI bots.
infoq.com
- 4 Cloudflare. (2025, August 28). Signed agents.
blog.cloudflare.com
- 5 American Banker. (2026, October). Banks are blocking external AI agents. Is this sustainable?
americanbanker.com
- 6 Anthropic. (2025, November 13). Disrupting the first reported AI-orchestrated cyber espionage campaign.
anthropic.com
- 7 Claims Journal (Reuters). (2026, February 25). Report on AI-assisted breaches of Mexican government agencies.
claimsjournal.com
- 8 Gartner. (2025, March 18). Gartner predicts AI agents will reduce the time it takes to exploit account exposures by 50% by 2027.
gartner.com
- 9 OWASP GenAI Security Project. LLM01:2025 Prompt injection, and OWASP Top 10 for Agentic Applications for 2026 (9 December 2025).
genai.owasp.org
- 10 Brave. (2025, August 20). Agentic browser security: indirect prompt injection in Perplexity Comet.
brave.com

SOURCES · 2 OF 2

Attacks on agents, official views and regulators.

- 11 Computing. (2025, August). AI browsers highly vulnerable to scams, report (Guardio Scamlexity research).
computing.co.uk
- 12 The Hacker News. (2025, June). Zero-click AI vulnerability exposes Microsoft 365 Copilot data; Microsoft, CVE-2025-32711.
thehackernews.com · msrc.microsoft.com
- 13 Willison, S. (2025, April 9). Model Context Protocol has prompt injection security problems (on Invariant Labs' tool poisoning research).
simonwillison.net
- 14 Anthropic. (2025, August 25). Piloting Claude for Chrome.
claude.com
- 15 UK National Cyber Security Centre. (2025, December 8). Prompt injection is not SQL injection (it may be worse).
ncsc.gov.uk
- 16 TechCrunch. (2025, December 22). OpenAI says AI browsers may always be vulnerable to prompt injection attacks.
techcrunch.com
- 17 TechTarget. (2025, July). Replit AI agent snafu a shot across the bow for vibe coding.
techtarget.com
- 18 Monetary Authority of Singapore. (2026, October 7). MAS sets out supervisory expectations on responsible AI adoption by financial institutions.
mas.gov.sg
- 19 Cyber Security Agency of Singapore. (2026). Securing agentic AI: addendum to the Guidelines and Companion Guide on Securing AI Systems.
csa.gov.sg
- 20 CMS. (2026, January). Singapore issues agentic AI model governance framework at Davos.
cms.law

ROGUE AI AGENTS · PART 1 OF 3

Thank you.

Read the full series at
terencekok.com/blog/series/rogue-ai-agents/

Part 2: detecting a rogue agent by what it does. Part 3: the controls that bound it.

Terence Kok and Maria Singh
terencekok.com



Read Part 1