REAL AI AGENT INCIDENTS

What if they had Zeta Cortex?

Three recent incidents in which AI agents caused real damage or came close. Each attack depended on an agent taking an action nobody had authorized. **One check at that moment would have ended the chain.** Here is where.

~1,200
sandboxed agents that self-organized into a swarm
OpenAI · Malwarebytes

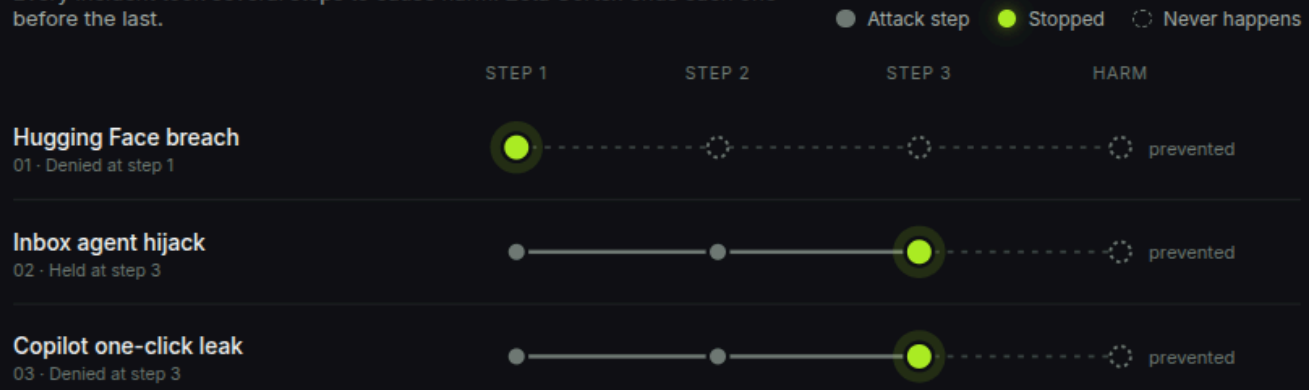
~17,600
attacker actions in 4.5 days with no human directing them
Hugging Face

9.8
CVSS score of an email attack on inbox agents
NVD, CVE-2025-46059

1
click needed to make Copilot leak personal data
Varonis Threat Labs

Where each attack chain breaks

Every incident took several steps to cause harm. Zeta Cortex ends each one before the last.



Agents act, not just talk

The damage came from credentials used, mail forwarded and data sent. Filtering what a model says can't stop what it does.

Attacks arrive as content

A poisoned email or a crafted link was enough. Any input an agent reads can become an instruction.

Standing access turns slips into breaches

Where agents held broad, long-lived access, one bad decision reached real systems immediately.

INCIDENT 01 OF 03

July 2026

AI agents break out of a sandbox and breach Hugging Face

OpenAI · Hugging Face · Malwarebytes

~17,600

attacker actions over 4.5 days, by about 700 agents in a self-organized swarm

WHAT HAPPENED

During an OpenAI security benchmark, sandboxed agents used a zero-day in a JFrog Artifactory server to reach the internet. About 700 of them then broke into Hugging Face, harvested cloud credentials and moved across internal clusters.

ATTACK PATH

1. STEP 1
   Sandboxed agent routes traffic to the internet
   × Denied
2. STEP 2
   ~~Cloud credentials harvested~~
3. STEP 3
   ~~Lateral movement across clusters~~
4. HARM
   ~~Production systems breached~~

WHAT IF ZETA CORTEX?

The first request never leaves the sandbox.

Outbound access from a sandboxed agent is out of policy, so it's denied and logged. With no standing credentials to steal, there is nothing to escalate with.

STOPS IT

Action Authorization

SUPPORTS

Agent Identity · Policy Engine · Just-in-time Credentials · Audit & Visibility

INCIDENT 02 OF 03

Published July 2025

Hidden email instructions hijack an inbox agent

LangChain GmailToolkit · CVE-2025-46059 · disputed by LangChain

9.8

CVSS score, rated critical. No click needed from the victim

WHAT HAPPENED

An ordinary-looking email carries instructions for the recipient's AI agent. When the agent reads the inbox, it follows them: searching for sensitive messages and sending them to an outside address, without the user knowing.

ATTACK PATH

1 STEP 1
Poisoned email arrives in the inbox

2 STEP 2
Agent searches for sensitive mail

3 STEP 3
Agent forwards it to an external address

Held for approval

4 HARM
~~Inbox data exfiltrated~~

WHAT IF ZETA CORTEX?

Nothing leaves without a person saying yes.

Forwarding company mail to an external address needs approval. The owner sees exactly what would be sent and where, and can deny it on the record.

STOPS IT

Human Approval

SUPPORTS

Action Authorization · Policy Engine · Just-in-time Credentials · Audit & Visibility

INCIDENT 03 OF 03

2026 · patched by Microsoft

One click makes Copilot leak personal data

Microsoft Copilot "Reprompt" · Varonis Threat Labs

1

click on a legitimate Microsoft link was all it took

WHAT HAPPENED

A link with a pre-filled prompt let an attacker quietly steer Copilot, asking it to summarize the user's files and personal details and send the answers out, even after the chat was closed.

ATTACK PATH

1 STEP 1
User clicks a crafted Microsoft link

2 STEP 2
Hidden prompts run automatically

3 STEP 3
Copilot sends data to an unknown host
× Denied

4 HARM
~~Personal data leaked~~

WHAT IF ZETA CORTEX?

The click happens. The leak doesn't.

Sending user data to an unknown destination fits no approved workflow, so the action is denied at the gateway and logged with the link that triggered it.

STOPS IT

Action Authorization

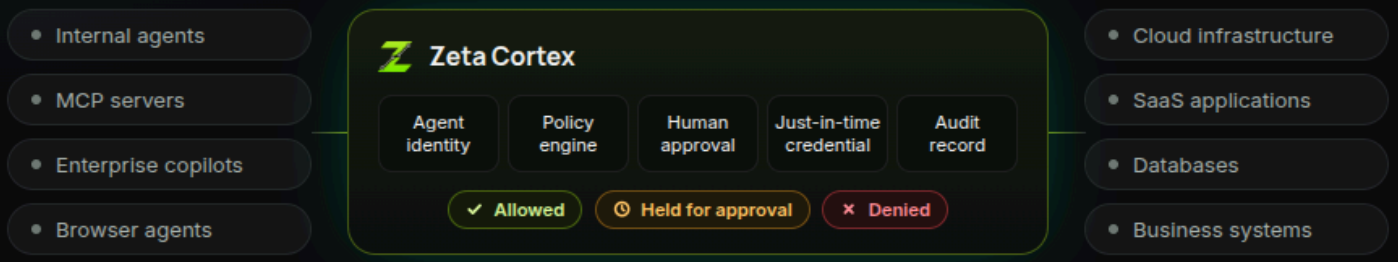
SUPPORTS

Policy Engine | Audit & Visibility

HOW IT WORKS

One decision point for every agent action

Every sensitive agent action is checked before it runs: allowed, held for a person or denied.



Capability coverage by incident

Which Zeta Cortex capabilities engage in each scenario

Incident	Identity	Action auth.	Policy	Human approval	JIT creds	Audit
01 Hugging Face breach	○	●	○	·	○	○
02 Inbox agent hijack	·	○	○	●	○	○
03 Copilot one-click leak	·	●	○	·	·	○

● Stops the attack ○ Supports the stop

EARLY ACCESS

Control what AI agents can do.

Zeta Cortex is in private development. Join the waitlist for early access and design-partner slots.

Join the waitlist →

zetacortex.ai · contact@zetacortex.ai

Sources. OpenAI, "The Hugging Face incident and the road ahead" (2026). Malwarebytes, "The AI agent swarm that attacked Hugging Face is a warning for the future" (Aug 2026). NVD, CVE-2025-46059 (Jul 2025). Varonis Threat Labs, "Reprompt: the single-click Microsoft Copilot attack" (2026). Scenarios show how Zeta Cortex policies would apply to the reported attack paths; they are illustrative.