AI Threat Monthly Brief

October 2026

TDR Focus: when autonomous attack frameworks went mainstream

Executive edition

Sekoia

October 2026

In October, attackers used AI tools that run intrusions on their own in real operations. With these open-source frameworks, one operator can attack hundreds of targets at once.

Published November 2026 Sekoia - Threat Detection & Research (TDR) 101 reports from 64 sources ai-threat.watch

The month in brief

Bottom line. At least 17 security organizations, including Microsoft, CrowdStrike and Google, reported attacks run by autonomous attack frameworks such as ARTEX and Cairn, and India's CERT-In issued a critical advisory on them. They were used by state-sponsored espionage groups and by criminal groups. Response plans that assume a human attacker's pace need revisiting.

< 3 h
to privileged cloud access at some targets (Microsoft)

200+
organizations in 14 countries hit by one campaign (CrowdStrike)

900+
ARTEX instances found exposed online (AhnLab)

AI-Enabled Autonomous attack frameworks are in real operations

What happened
Penetration-testing frameworks in which an AI "planner" drives scanners and exploits were used by China-linked espionage groups (per Microsoft and CrowdStrike), North Korean clusters (per Microsoft) and criminal groups targeting banks and e-commerce.

Why it matters
The skill barrier and the time to compromise both drop. Exposed services, weak passwords and unpatched VPNs or APIs are found and exploited within hours.

AI-Targeted AI tools are now a door into the company

What happened
Infostealers now harvest Claude Code, Codex and Cursor credentials. A tampered LiteLLM release, a trojanized Claude Code extension and malicious agent "skills" reached developers directly.

Why it matters
AI keys and agent tokens open code repositories, cloud accounts and paid model quotas. They are privileged credentials and need the same controls.

AI-Enabled Cybercrime keeps getting cheaper

What happened
A ransomware group spent over $7,000 on AI to write its attack scripts (SOCRadar). Criminal mentions of guardrail-stripped open models rose 745% (Flare).

Why it matters
Mid-tier groups can now do work that used to need skilled operators. Expect more average-quality attacks, launched faster.

Three questions to ask your team this month

1. Has our red team tested us with the AI attack frameworks that attackers now use?
2. Do we have an inventory of AI agents, AI API keys and AI platforms, with an owner for each?
3. Could we detect and contain an intrusion that reaches admin rights in three hours?

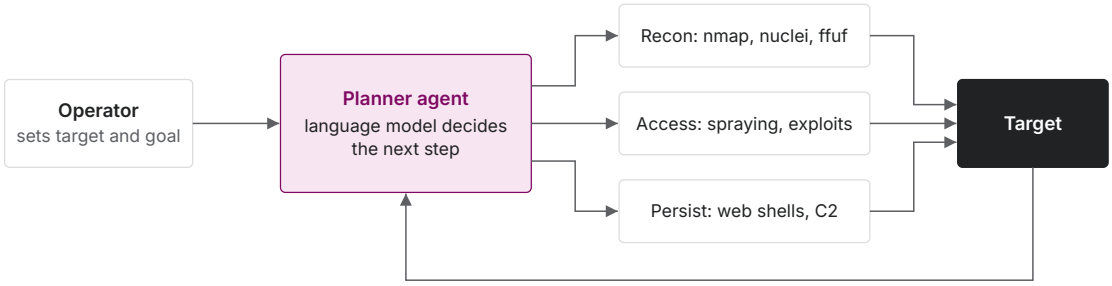
TDR Focus: autonomous attack frameworks

Each month, TDR picks one trend and gives its view.

An autonomous attack framework is a program in which an AI model plays the role of the attacker's team lead. The operator gives it a target and a goal. A "planner" agent, backed by a language model, breaks the goal into steps and hands them to "worker" agents that run ordinary hacking tools: port scanners, vulnerability scanners, password sprayers, exploit scripts. Results come back to the planner, which decides the next step.

> **The closest comparison is Cobalt Strike.** Most of these frameworks were released as open-source tools for penetration testers, then picked up by attackers, the way Cobalt Strike was a decade ago.

The difference is that they need neither a skilled operator nor a team. AhnLab found ARTEX forked and rebranded by several operators, and Trend Micro observed one platform coordinating 44 agents at once.



The frameworks seen most in October

October reports naming each framework

ARTEX	**15**	Storm-0940, VAPOR PANDA and others
Cairn	**6**	Criminal clusters against e-commerce and payment sites
PentAGI	**4**	400+ Latin American banks targeted via CARLIY0S
CyberStrikeAI	**2**	44 agents coordinated against Japan
Hermes Agent	**2**	Jailbroken DeepSeek for fraud and credential theft
4 others	**1 each**	Phantom, Ultra Suite, Hexstrike, Strix

A report can name several frameworks. Models most often named alongside them: DeepSeek (11 reports), GLM and Kimi (5 each), Grok (4).

Who used them in October

● APT ○ Cybercrime

Framework	APT	Cybercrime
ARTEX	Storm-0940 China · Microsoft, high VAPOR PANDA China · CrowdStrike, medium Eight North Korean clusters North Korea · Microsoft, high	Unnamed actor vs South Korean finance CrowdStrike, not attributed
Cairn	None reported	Three criminal clusters vs e-commerce NorForge, not attributed Unnamed actor, server takeover ReliaQuest, not attributed
PentAGI	None reported	CyberCartel (CARLIY0S) Mexico · Straiker, low CyberXero, access broker Russia · SOCRadar, medium PaperCut exploitation actor Russia · GreyNoise, low
CyberStrikeAI	None reported	No actor named (targets in Japan) Trend Micro
Hermes Agent	None reported	BUGSTER Unit 42, not attributed

TDR analysis

The limit moves from people to compute. Until now, an intrusion campaign was bounded by the skills of the attackers and the hours they could put in. Autonomous frameworks sharply lower the skill needed to run a full intrusion, and they do not sleep. We expect campaigns to run at a larger scale and almost around the clock, limited less by the number of human operators than by the cost of tokens and infrastructure.

This list is already incomplete. It only covers frameworks that made it into public reporting. New ones appear almost daily, because vibe coding (building software by prompting a model) makes them quick to build. We expect the most capable attackers to have their own model build a custom framework modeled on ARTEX, Cairn and similar tools. A one-off tool leaves fewer traces, so defenses that look for known tool names will miss it.

Sekoia on the front line

What Sekoia sees first-hand, in its own honeypots and investigations.

Attack on an AI agent platform, captured by a Sekoia honeypot

AI-Targeted

TDR exposed a decoy Flowise server, a tool companies use to build AI agents. An attacker found it, took the AI API keys and database credentials it held, then came back to plant a malicious workflow and to turn the AI agent into a reconnaissance assistant.

0 to 3 min — Discovery and theft
Fingerprints the server, lists workflows and credentials, takes the API keys and database access.

+2 h — Pause
No activity for just over two hours.

Return — Manual reconnaissance
Scans the platform's API, takes the credentials again, tests the AI with a simple prompt.

Malicious workflow
Creates a flow named "System Status" that calls the attacker's server.

Abuse of the AI agent
Tries to jailbreak it, extract its instructions and get it to analyze a network to attack.

Every 10 min — Polling
Keeps requesting stored variables and the list of workflows.

Who. TDR assesses with high confidence that the attacker is a fluent, likely native, Chinese speaker. The activity fits opportunistic cybercrime against unregulated online gambling sites in Hong Kong and Taiwan. No known group.

What to do

1. Treat AI agent platforms as credential stores: strong authentication, no secrets reachable from the internet.
2. Limit what each agent can reach, especially databases and document stores.
3. Alert on new workflows and on outbound calls from AI platforms to unknown hosts.

Observed late August 2026 Sent to Sekoia customers on 7 September Published 19 October Read the full Sekoia report

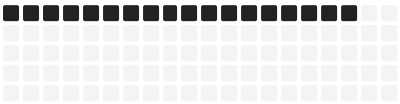
October in numbers

Each figure comes from the source named under it. Figures from AI Threat Watch describe the collection, not the threat.

18.5%

AI-Targeted

of 35,878 AI agent skills analyzed were malicious. 62% of those contained no code, only instructions.

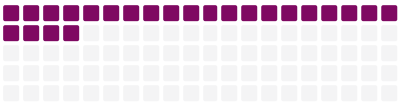


VirusTotal · each square is 1%

24 of 101

AI-Enabled

October reports, from 64 sources, name an autonomous attack framework.



AI Threat Watch · each square is 1%

745%

AI-Enabled

rise in criminal mentions of guardrail-stripped open-weight models.



after: 8.45 times as many

Flare

Speed and scale

~20 min

AI-Enabled

for an AI agent to move laterally and harvest domain controller credentials

CrowdStrike

400+

AI-Enabled

Latin American banks and fintechs targeted by the CARLIY0S multi-agent system

Straiker

Also this month

3,400+

AI-Targeted

servers infected by PoeLLM, which targets self-hosted LLM services such as Ollama and LiteLLM

Black Lotus Labs (Lumen)

$7,000+

AI-Enabled

spent on AI usage by the n0n ransomware group to write and debug its scripts

SOCRadar

140

CVEs and disclosures found by Microsoft's own AI vulnerability-hunting system

Microsoft

Method

This brief draws on the 101 entries AI Threat Watch collected with an October 2026 date. Each entry is a public report summarized and classified automatically, then checked against the site's data rules. Figures attributed to a source are that source's; counts attributed to AI Threat Watch describe the collection. Attribution is reported as stated by each source, with its confidence level. AI Threat Watch does not attribute on its own.