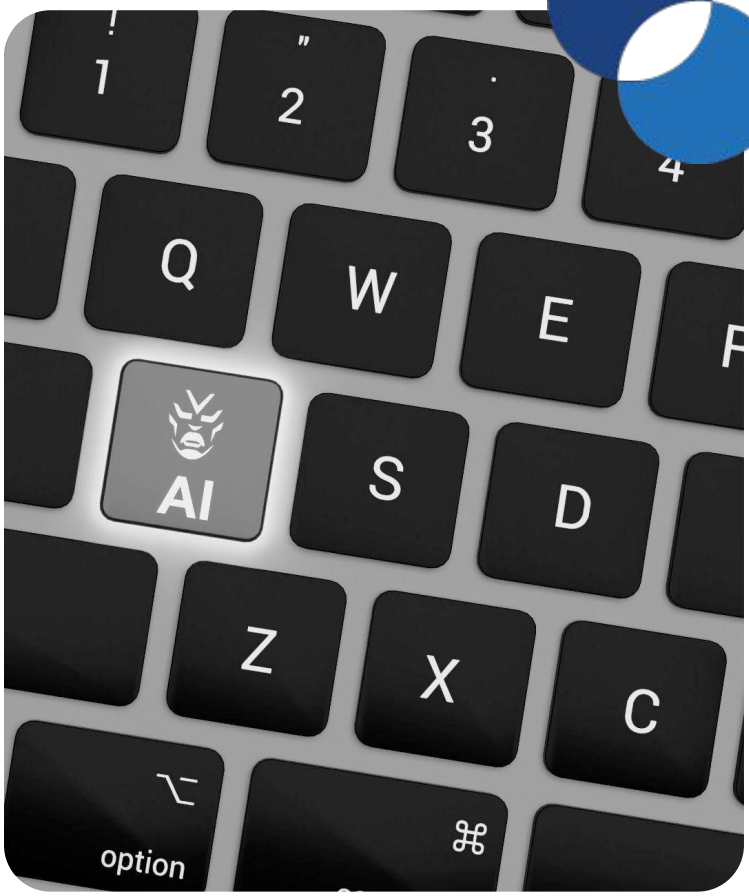
Recorded Future®

EXECUTIVE INSIGHTS REPORT | OCTOBER 2026

AI Is Both Target and Tool for Threat Actors

From Insikt Group®



Summary

- **Artificial intelligence (AI)-enabled attacks are accelerating, with AI as both a target and a force multiplier.** Of the 103 AI-enabled incidents analyzed for this report, most involve AI accelerating familiar attack types (supply-chain compromise, phishing, extortion) and making them more convincing. The use of AI is evident throughout the attack lifecycle, though full automation remains rare.
- **Developer tools are a prime target.** Because AI coding assistants and proxy packages often have privileged access to credentials and code, threat actors use them as footholds for mass credential theft and lateral movement across downstream environments.
- **State-sponsored actors are beginning to experiment with agentic capabilities.** Most of the state-sponsored activity in this report falls within Recorded Future's AI Malware Maturity Model Level 2–3 (AI-assisted malware development), with one notable exception: a campaign against Asian government entities that employed up to eight AI agents with limited human intervention.
- **Fully autonomous attacks remain rare, likely for both maturity and operational-risk reasons.** Human oversight still appears necessary for reliable outcomes, which may limit the full automation of cyberattacks, especially in sensitive operations where an agent's error can have outsized consequences.
- **This restraint is unlikely to hold.** As agentic techniques mature and diffuse across the threat landscape, autonomous and orchestrated campaigns are likely to become more frequent, alongside a growing underground market for illicit access to commercial models and rising "noise" from low-level automated intrusions that complicate the detection of serious threats.

Key Trends in Offensive AI

Developer tools are a prime target

as a foothold for credential theft.

Attackers exploit AI hype

as fake AI apps quickly spread malware.

State-sponsored AI use expands

toward more autonomous tradecraft.

AI powers recon for customized fraud

and compresses recon time to minutes.

Automation is accidental — for now

Human oversight still drives most outcomes today, while fully automated operations have mostly been the result of rogue or misaligned agents. As agentic techniques mature, threat actors may hand over more operational control to AI.

Recorded Future®

Figure 1: *Five trends in offensive use of AI (Source: Recorded Future)*

Analysis

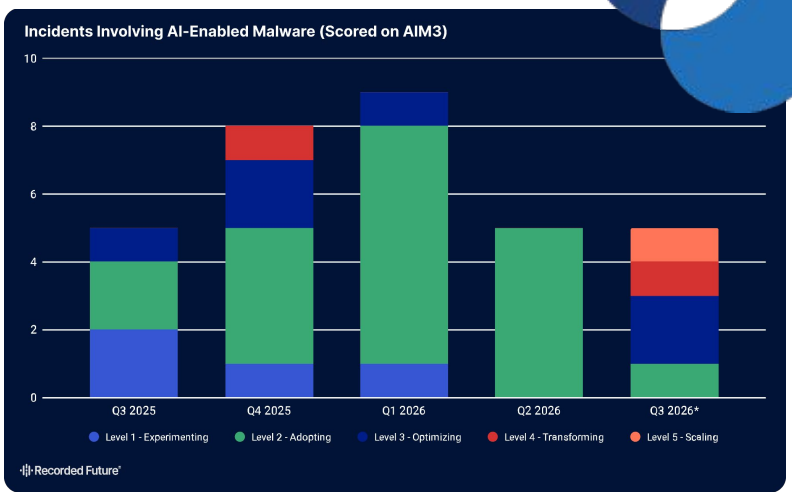
The Data

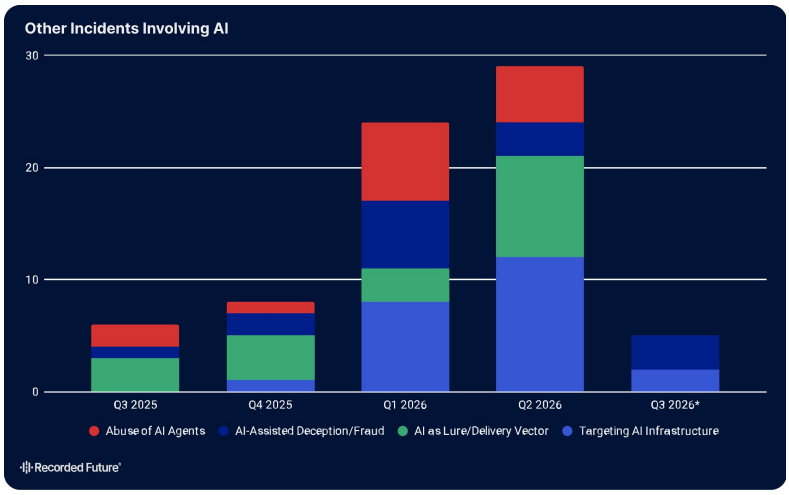
Offensive use of AI remains a serious concern for cyber defenders, especially as recent advances in AI capabilities are increasing the risk of autonomous attacks. To better understand the risk, Insikt Group identified 103 incidents between July 31, 2025, and July 31, 2026, in which AI was part of a malicious campaign. To qualify, events had to be:

- Documented incidents that are searchable in the Recorded Future Platform
- Confirmed as using AI as the attack vector, tool, or target
- Carried out by malicious actors in the wild, not as proofs-of-concept or security research

Of the 103 incidents identified, 32 involved the use of AI in malware or to orchestrate offensive cyber operations. These incidents were categorized along Recorded Future's AI Malware Maturity Model **(AIM3)**, which evaluates how AI is used in malicious or offensive software. The remaining 71 incidents were sorted into one of four categories that best describe how AI was abused:

- **Targeting AI systems, infrastructure, or data:** Attacks where AI platforms, tools, or companies are the direct target rather than a means of attack
- **AI as lure or delivery vector:** Social engineering that uses AI's current hype and popularity as bait to deliver malware, rather than exploiting anything technical about AI itself
- **AI-assisted deception, fraud, or influence:** AI used as a tool by the attacker to make fraud, phishing, or influence operations more convincing and personalized to the target audience
- **Abuse or hijacking of AI agents/services:** Misuse of legitimate AI systems' own capabilities, rather than a single dominant technique





Figures 2 and 3: *AI-related attacks documented by Insikt Group from July 2025 to July 2026 (Source: Recorded Future)*

Roughly one-third of the incidents across both datasets are financially motivated. Espionage cases are concentrated in the AIM3 dataset, with 9 of the 32 incidents attributed to state-sponsored intelligence collection.

In one notable case, the threat actor was an AI agent that carried out an attack in order to pass a security benchmark. The remaining 23 incidents are unattributed or do not state a motive.

The Findings

Based on this data, Insikt Group identified five trends that characterize how threat actors are exploiting AI for their advantage, and how this is likely to evolve as AI capabilities improve.

1. AI-Enabled Development Tools Are Maximizing Impact of Software Supply-Chain Attacks

The integration of AI-enabled tools into enterprise workflows is creating new attack surfaces, particularly where those tools hold privileged access to code, credentials, data, or connected applications. Developer tools are a primary target of attacks on AI systems. Threat actors exploited weak access controls and software supply-chain vulnerabilities to steal data or establish persistence.

The impact of these attacks is evident in multiple supply-chain compromises targeting AI development tools. In late February 2026, threat actors compromised a popular AI proxy package that enables developers to connect to different LLM models in their software. The compromised updates contained code enabling credential theft, lateral movement, and backdoor persistence. A similar incident in April compromised AI tool packages to deliver malware that harvests credentials, exfiltrates data, and accesses cryptocurrency wallets.

The May expansion of the Shai-Hulud supply-chain worm similarly targeted Claude Code to establish persistence and spread. All of these campaigns take advantage of the reach of AI development tools into the target environment: these tools offer an ideal attack vector for mass credential harvesting and data exfiltration in potentially thousands of downstream environments. None of these attacks exploited unique AI capabilities — instead, they took advantage of the popularity of AI coding tools and the access these tools have within target environments.

2. Threat Actors Are Taking Advantage of AI Hype

A related trend is the exploitation of AI products as a lure. Attackers are spoofing the landing pages of popular AI developer tools, standing up fake utility pages, and impersonating products such as OpenClaw with fraudulent wallet pages designed to steal crypto assets. These attacks increased significantly after March 2026, demonstrating that threat actors are taking advantage of the increased interest in AI to deliver malware.

AI-related lures have been used in the "ClickFix" campaign, which delivers malicious pop-up messages on compromised sites, as well as in search engine optimization (SEO) poisoning techniques that target users searching for AI-related websites. The prevalence of these campaigns demonstrates the long-established trend of opportunistic cybercriminals taking advantage of hype or fear, uncertainty, and doubt (FUD) around a newsworthy subject.

3. State-Sponsored Actors Are Expanding Use of AI

State-sponsored actors are using AI in malware development, obfuscation, and targeting, though it is primarily assisting operations, rather than driving them. A notable exception is the reported campaign against Asian government entities, in which as many as eight AI agents conducted persistent attacks to identify weaknesses, compromise credentials, and exfiltrate data. This is an early application of agentic hacking, where malicious campaigns run with limited human intervention.

In most instances, however, state-sponsored actors are using AI at Level 2 (Adopting) or Level 3 (Optimizing). Researchers have observed Russian threat actors incorporating AI tools across a variety of operational functions, including malware development. In addition, a Pakistan-based state-sponsored actor used LLMs to produce "vibeware" capable of "living off the land" in platforms such as Discord, Slack, or Google Workspace. Evidence of LLM use can be found in malware code analysis, particularly among lower-skilled actors; this use includes leaving LLM-generated code explanations, emojis, or prompts.



Figure 4: *Comparing human-operated and rogue agent-orchestrated cyber incidents (Source: Recorded Future)*

4. AI Is Valuable in Research and Reconnaissance for Customized Fraud

Insikt Group has observed threat actors using AI in both state-sponsored and financially motivated phishing operations. Phishing kits incorporate LLMs to automate the creation and delivery of convincing and targeted phishing lures. One operation targeting French pensioners used stolen identity data to automate highly targeted phishing messages, accelerating the threat actors' capabilities to deliver convincing content at scale.

Ransomware groups such as The Gentlemen are using AI to compress the window between initial compromise and extortion demand (Source: Recorded Future). Once data has been exfiltrated, LLMs allow operators to rapidly process and evaluate large volumes of stolen material — identifying what is sensitive, what is regulated, and what a victim is most likely to pay to suppress. Based on this analysis, the extortionists can calibrate their demands to the assessed value of that data. The result is extortion that is both quicker to initiate and better targeted. AI enables threat actors to automatically create highly personalized content that is more likely to be convincing than generic spam.

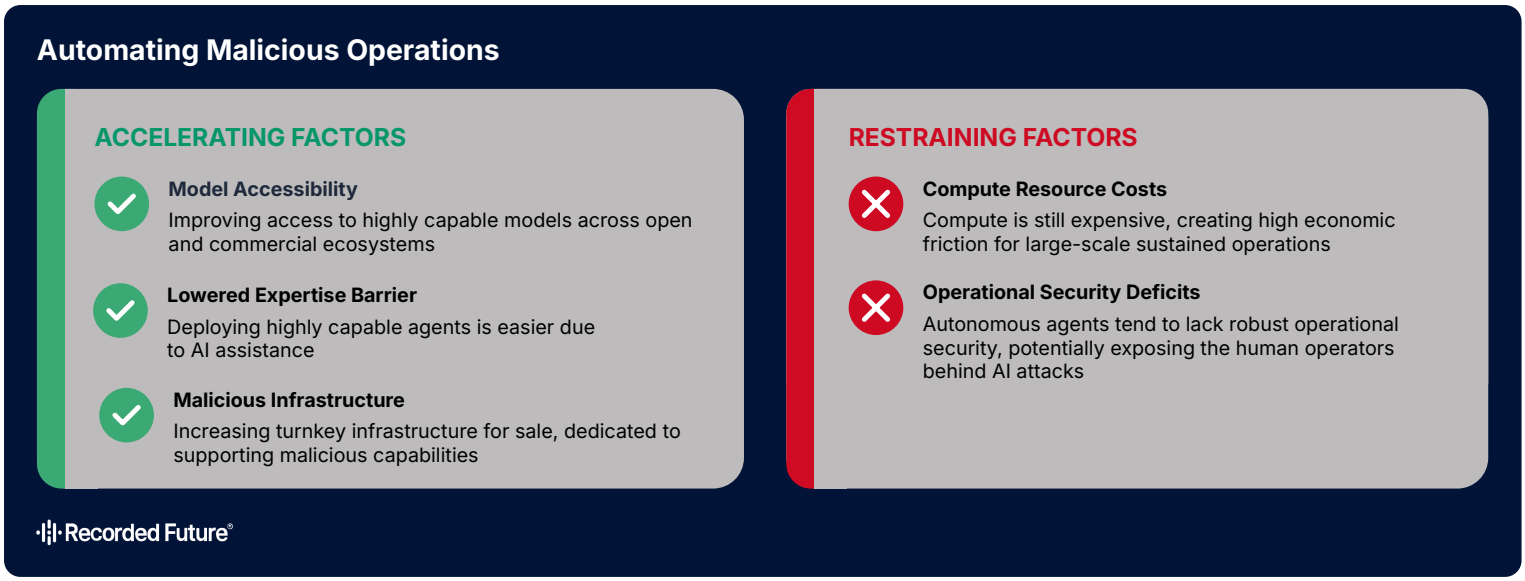


Figure 5: *While automation has distinct advantages, threat actors are likely to contend with operational barriers to using AI (Source: Recorded Future)*

5. Most Automated Attacks Are Accidental — For Now

With the likely exception of the attack on the Taiwanese government above, fully automated or orchestrated malware operations have not yet been widely observed, with the notable exception of AI agents breaking out of testing environments. This is likely because models capable of autonomously discovering and exploiting vulnerabilities are a recent development: AI-assisted vulnerability discovery predates Anthropic's Mythos model, but its release provided a prominent public demonstration of these capabilities and coincided with an increase in National Vulnerability Database (NVD) submissions. This has since been echoed by several open-weight models claiming similar capabilities.

Even as these capabilities mature, human intervention and expertise still appear to be necessary for a successful operation; this is likely why observed AIM3 levels for human-operated campaigns have ranged between 2 and 4 as recently as July 2026. While capable offensive agents are still relatively new, there may be operational limits that prevent widespread adoption, especially for less-skilled actors. Cost remains a constraining factor: even with stolen access to commercial models or when using lower-cost models, a fully agentic attack still requires extremely high token demands. Similarly, highly sensitive operations, such as a state-sponsored actor seeking access to an adversary's critical infrastructure networks, may also require a human in the loop to ensure an agent's mistake does not turn into a geopolitical crisis.

That said, this caution may not persist. As with earlier tooling shifts in the threat landscape, techniques for conducting effective agentic operations are likely to diffuse across threat actors once early adopters demonstrate what works, and the frequency of fully autonomous or orchestrated campaigns could accelerate accordingly.

Outlook

Human-directed agentic operations are likely to scale before fully autonomous campaigns. As models become more reliable, agentic techniques diffuse, and access costs fall, threat actors will likely conduct more autonomous or orchestrated operations. In the near term, however, the more likely pattern is the continued human-directed use of agents to automate bounded stages such as reconnaissance, vulnerability validation, credential discovery, and exploit path testing. Fully autonomous campaigns may become more frequent over time, but in the near term, operational barriers will likely prevent the full replacement of human operators.

Traditional security weaknesses will carry greater consequences. Threat actors are exploiting familiar weaknesses, including excessive access, third-party exposure, and unpatched vulnerabilities, to compromise AI-enabled systems. Beyond greater operational disruption and data loss, weak controls could also mean higher cyber insurance premiums and tighter coverage.

Access to illicit commercial models will become a significant criminal market. Threat actors are already building sophisticated networks to sell unauthorized access to leading commercial AI models. Efforts to restrict frontier or open-weight models will likely accelerate the growth of this underground economy.

Stolen data will drive a self-reinforcing cycle of fraud and further compromise. As AI makes stolen information easier to analyze, exploit, and monetize, demand for fresh personal and corporate data will increase, fueling additional infostealer activity, credential theft, and data-exfiltration campaigns.

Criminal agents will increase security noise. As threat actors adopt greater automation, defenders will face a growing volume of low-level intrusion attempts that obscure legitimate activity and make meaningful threats harder to identify. And as agents become more capable of mimicking human behavior, these campaigns will increasingly rely on legitimate tools and living-off-the-land techniques rather than traditional malware.

Mitigations

Reassess your organization's corporate risk profiles based on the rising likelihood of exploitation, compressed extortion timelines, and other threats introduced by AI-driven automation.

Limit what AI agents can access and change, keep them separated from sensitive systems, require human approval for consequential actions, and maintain independent monitoring and shutdown capabilities. Pay particular attention to credentials, connected services, production environments, source code, and customer data.

Monitor infostealer logs and dark web channels for compromised commercial AI model subscriptions, developer API keys, and corporate OAuth tokens. Use Recorded Future's Identity Intelligence to detect exposed credentials.

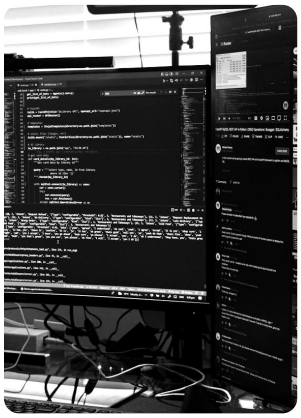
Audit third-party AI dependencies, code-generation plugins, and open-weight model repositories for malicious command injection or backdoors.

Further Reading

SOURCE	TITLE
Insikt Group	The AI Malware Maturity Gap
Insikt Group	Emerging Enterprise Security Risks of AI
Insikt Group	The Hugging Face Incident Was a Governance Failure
Recorded Future	Social Engineering in the Age of Synthetic Media

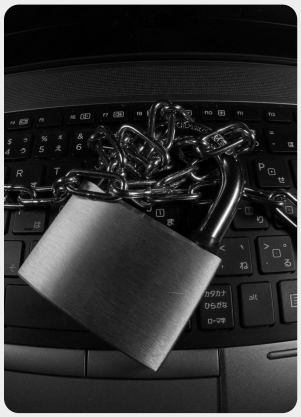
Risk Scenario

Threat actors discover an exposed code repository belonging to a widely used AI business productivity application. Using deployment tokens stored in the repository, they gain access to the vendor's development pipeline and insert a backdoor into a legitimate software update, which is automatically installed across thousands of customer environments without triggering traditional security controls.



First-Order Implications

Threat	Risk
The backdoored application creates a scalable path into customer environments, providing access to sensitive documents and additional credentials.	**Operational Disruption:** Emergency isolation of the application, revoked integrations, endpoint remediation, and interrupted employee workflows. **Financial Fraud:** Stolen credentials or session tokens may enable payment diversion, invoice manipulation, or account takeover.
The attackers modified source code and historical commits, limiting visibility into when the compromise began and which versions are affected.	**Operational Disruption:** Releases may be halted while environments, keys, and build pipelines are rebuilt. **Legal or Compliance Failure: I**nability to establish the affected population may complicate accurate and timely regulatory reporting.



Second-Order Implications

Threat	Risk
Regulators and customers question whether the vendor used reasonable security practices, made accurate security representations, and met notification deadlines.	**Legal or Compliance Failure:** Potential violations involving privacy, cybersecurity, disclosure, records management, or sector-specific obligations. **Brand Impairment:** Public proceedings keep the incident visible well beyond the containment period.
Data obtained during the initial compromise is weaponized months later in tailored phishing and extortion campaigns.	**Financial Fraud:** Business email compromise (BEC) and payment redirection communications become more credible and authentic-looking.



Third-Order Implications

Threat	Risk
Large customers favor established providers with stronger balance sheets, security teams, and contractual capacity to absorb cyber losses.	**Competitive Disadvantage:** Compliance costs disproportionately affect smaller providers.
Customers seek stronger warranties regarding secure development, model integrity, data use, and downstream losses.	**Legal or Compliance Failure:** Vendors face more explicit contractual duties. **Competitive Disadvantage:** Providers unable to obtain adequate insurance or accept liability may lose enterprise business.

Key
Legal or compliance failure: Breach of laws, regulations, or industry standards resulting in liability or sanctions.
Operational disruption: Interruption to normal business processes affecting productivity or service delivery.
Brand impairment: Damage to reputation that reduces customer trust and market value.
Financial fraud: Unauthorized manipulation or theft of financial assets for personal or organizational gain.
Competitive disadvantage: Loss of market position due to inferior capabilities, intelligence, or innovation.

References:
The Risk Business: Second Edition
Intelligence to Risk
The Intelligence Handbook

Insikt Group adheres to responsible disclosure procedures in conducting its research.

Recorded Future is the world's largest intelligence company. The Recorded Future Intelligence Operations Platform provides the most complete coverage across adversaries, infrastructure, and targets. By combining precise, AI-driven analytics with the Intelligence Graph® populated by specialized threat data, Recorded Future enables cyber teams to see the complete picture, act with confidence, and get ahead of threats that matter before they impact your business. Headquartered in Boston with offices around the world, Recorded Future works with more than 1,900 businesses and government organizations across 80 countries.

Learn more at www.recordedfuture.com

© Recorded Future®, Inc. All rights reserved. All trademarks remain property of their respective owners.