Public safety threat report:

Adversarial use of artificial intelligence in 2026

Disclosure protocol: GREEN: Restricted to the community
Date of publication: 17 November 2026

The Public Safety Threat Alliance (PSTA) threat intelligence team actively monitors and evaluates the threats to public safety. In this report, we examine the ways in which threat actors are leveraging artificial intelligence (AI) during cyberattacks. Our team used both open and closed sources as part of our investigation, including information from our ActiveEye Managed Detection and Response team, private and trusted vendors, and government reporting. We used numerous intelligence analytical techniques in the assessment of threat intelligence provided in this report.

Key points

- Sophisticated adversaries are beginning to use AI to conduct fully automated cyberattacks without a human in the loop.
- Top threat actors like *The Gentlemen* and *Qilin* use AI in their operations, and their public safety attacks likely involve malicious AI.
- Phishing remains the most common use case for malicious AI.
- AI isn't magic. Rigorous network security can still halt threat actors or slow them enough to get caught.

Executive summary

Threat actors in 2026 are using artificial intelligence to automate cyberattacks. Operations are shifting from human-guided assistance to fully autonomous agents that scan targets, steal data, and execute attacks with little human oversight. Key targets include government and critical infrastructure organizations. Attackers also relied on AI-driven phishing, voice cloning, and underground tools to bypass security controls. Standard network defenses such as hardware passkeys, strict access controls, and network isolation remain effective at stopping or slowing these threats.

Autonomous agents

Threat actors continue to employ artificial intelligence (AI) during cyberattacks, but their methods and focus changed over the course of 2026. At the beginning of the year, adversaries primarily used commercial large language models (LLMs) for tactical help. Threat actors were able to do this by breaking up tasks or jailbreaking flagship models via tricky prompting.[1] Humans were still present for most steps of the attack lifecycle, but AI began to do some of the heavy lifting (See Figure 1).[2]

[1] https://www.microsoft.com/en-us/security/blog/2026/03/06/ai-as-tradecraft-how-threat-actors-operationalize-ai/
[2] https://www.kaspersky.com/resource-center/threats/ai-phishing-scams

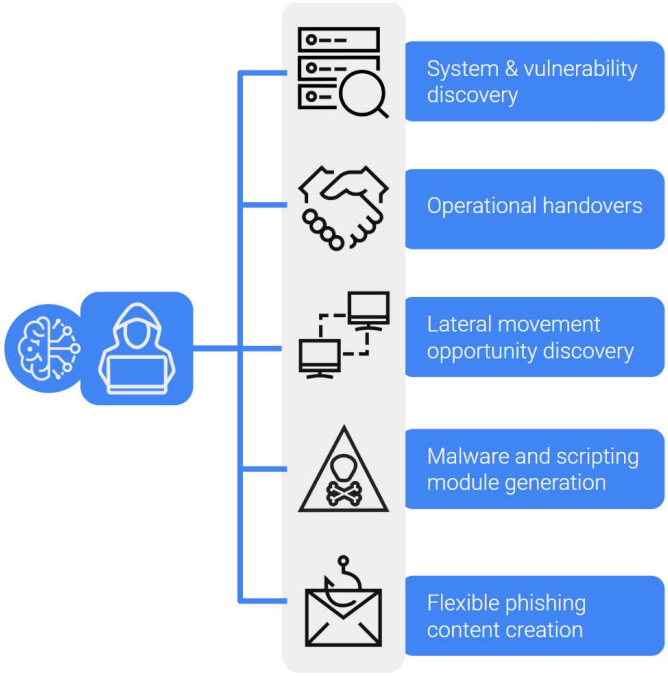


FIGURE 1: Example use cases for malicious AI with a 'human-in-the-loop'

Now, threat actors are beginning to leverage AI agents that can act alone. In these cases, humans give agents a playbook and the agent runs the attack entirely. As described by the Google Threat Intelligence Group (GTIG), "highly autonomous systems [are] capable of reasoning through complex tasks and making dynamic decisions without the need for human oversight."[3] They scan targets, fix their own errors, rotate IP addresses, and steal data without human intervention.[4]

How common is this? It's currently unclear. Many threat actors likely still operate with a person in the loop since AI agents continue to suffer from problems like hallucinations and can have immense computational costs. However, the most advanced operations are increasingly leveraging the more reliable frontier models autonomously.[5] Adversaries targeted the public, healthcare, and military sectors to hijack victim cloud environments for the purpose of powering these agents' workloads, solving one barrier to consistent reliance on autonomous systems.

Cybercriminals are also starting to experiment with unsupervised agents. GTIG saw a financially-motivated threat actor use an AI-coding chatbot to "plan, build, and execute a mass credential harvesting campaign in less than six hours," for example (See Figure 2). By deploying multiple agents, attackers were able to perform at the speed and scale of larger, more well-resourced groups. Similarly, the JADEPUFFER ransomware campaign was possibly the first in history to be fully automated by AI from start to finish; letting adversaries take advantage of machine speed and adaptive reasoning.[6]

[3] https://cloud.google.com/blog/topics/threat-intelligence/from-prompting-to-autonomy-the-evolution-of-adversarial-ai
[4] https://www.sysdig.com/blog/jadepuffer-agentic-ransomware-for-automated-database-extortion
[5] https://unit42.paloaltonetworks.com/ai-assisted-cyber-attack-inside-a-unit-42-investigation/
[6] https://www.forbes.com/sites/jonmarkman/2026/07/07/the-first-ransomware-attack-run-from-start-to-finish-by-an-ai-agent/

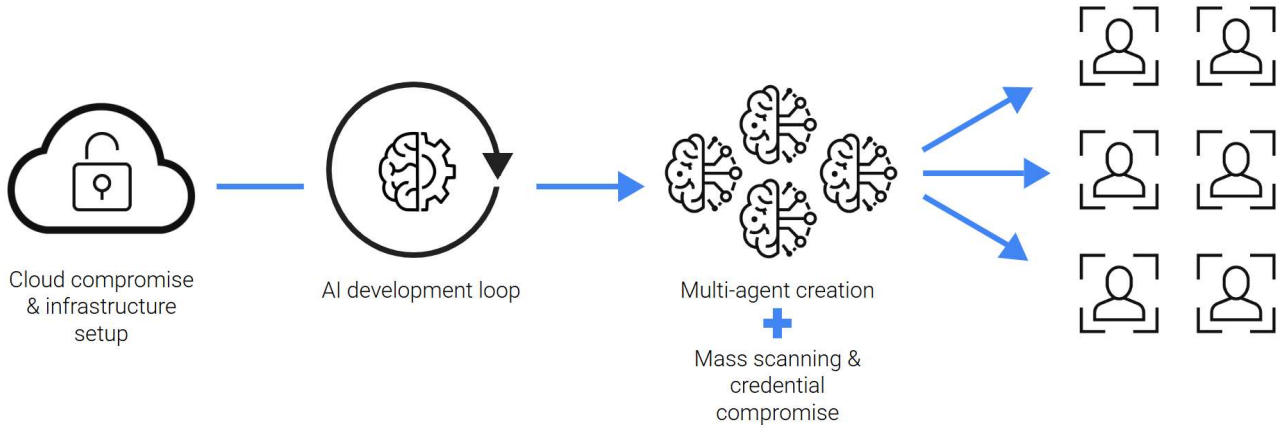


FIGURE 2: Overview of autonomous credential harvesting operation

Public safety attacks

This year, public safety threat actors like *The Gentlemen* leveraged a variety of LLMs for specific tasks, such as using China's DeepSeek for technical and coding objectives.[7] *The Gentlemen* are responsible for **12** attacks on municipalities, federal entities, and police this year. It cannot be confirmed whether these compromises involved malicious AI, but given *The Gentlemen's* observed experimentation with such tools, AI usage is highly likely.

Other top public safety adversaries like *Qilin* also used AI thanks to its effectiveness in lowering attack times, improving victim discovery and system enumeration, and scripting.[8] "At this point, we can assume that all threat actors are using AI in some capacity and their operations have benefited," John Hultquist, chief analyst at GTIG, told The Hacker News.[9] Similarly, INTERPOL found that over half of cybercrime in Africa involved AI, highlighting how popular AI has become for financially-motivated adversaries.[10]

Critical infrastructure

In August 2026, federal agencies reported AI-enabled threat actors targeting critical infrastructure organizations.[11] AI-generated scripts were disguised as legitimate monitoring software. Victim systems were identified via widespread internet scanning, and threat actors focused on devices with outdated software and those that were "otherwise poorly protected." The authoring agencies stated that the activity was "not a theoretical risk" but "an active threat" and was an "evolution" in threat actors' capabilities.

These types of attacks compromises can lead to safety incidents, equipment damage, theft of sensitive data, and the disruption of critical industrial processes. Additionally, critical infrastructure disruptions can directly impact public safety operations. When power grids or emergency communications go down, first responders lose vital tools and response times slow.[12]

[7] https://www.kelacyber.com/blog/the-gentlemen-ransomware-internal-chat-leak-analysis-2026/
[8] https://blog.talosintelligence.com/ransomware-incidents-in-japan-in-the-first-half-of-2026/
[9] https://thehackernews.com/2026/09/autonomous-ai-agents-compromise.html
[10] https://www.interpol.int/en/News-and-Events/News/2026/INTERPOL-report-finds-AI-linked-to-more-than-half-of-cybercrime-in-Africa
[11] https://www.cisa.gov/news-events/cybersecurity-advisories/aa26-231a?utm_source=SiemensS7SeriesPLC&utm_medium=GovDelivery
[12] https://carbyne.com/the-pulse-of-9-1-1/

AI-enabled phishing & vishing

Phishing remains one of the most reliable and common use cases for malicious AI, and a top access method for the public safety sector. The PSTA Threat Intelligence team regularly observes instances of phishing (AI-created or otherwise) used to obtain initial access to public safety—for example, the May 2026 phishing incident impacting the city of Roanoke, Virginia.[13] Or when adversaries targeted the Ukrainian government with AI-developed phishing pages.[14]

While AI involvement cannot always be confirmed, public safety phishing cases occur while AI-generated phishing content wildly expands globally.[15] Certain techniques like ClickFix have also become more common, and the consistency and frequency of these lures strongly suggests AI involvement. Phishing-as-a-service (Phaas) kits are likewise integrating AI, showing how the technology is now part of many cybercriminal offerings.[16]

Voice phishing (vishing) is a growing initial access tactic in 2026, and similarly directly benefits from advances in AI, in this case, voice cloning. The PSTA observed vishing campaigns targeting mission-critical systems where threat actors masqueraded as IT support in attempts to obtain initial access to 9-1-1 and emergency dispatch environments. While voice cloning was likely not used in these cases, vishing is becoming more popular globally and cybercriminal services are capitalizing on this demand with voice cloning advertisements and services.

For example, in March 2026, a threat actor advertised access to the DOOM call center platform, which allegedly enables users to use AI-driven voice agents that follow custom scripts and pose as human callers to trick victims into disclosing one-time password (OTP) codes and other sensitive information (See Figure 3).

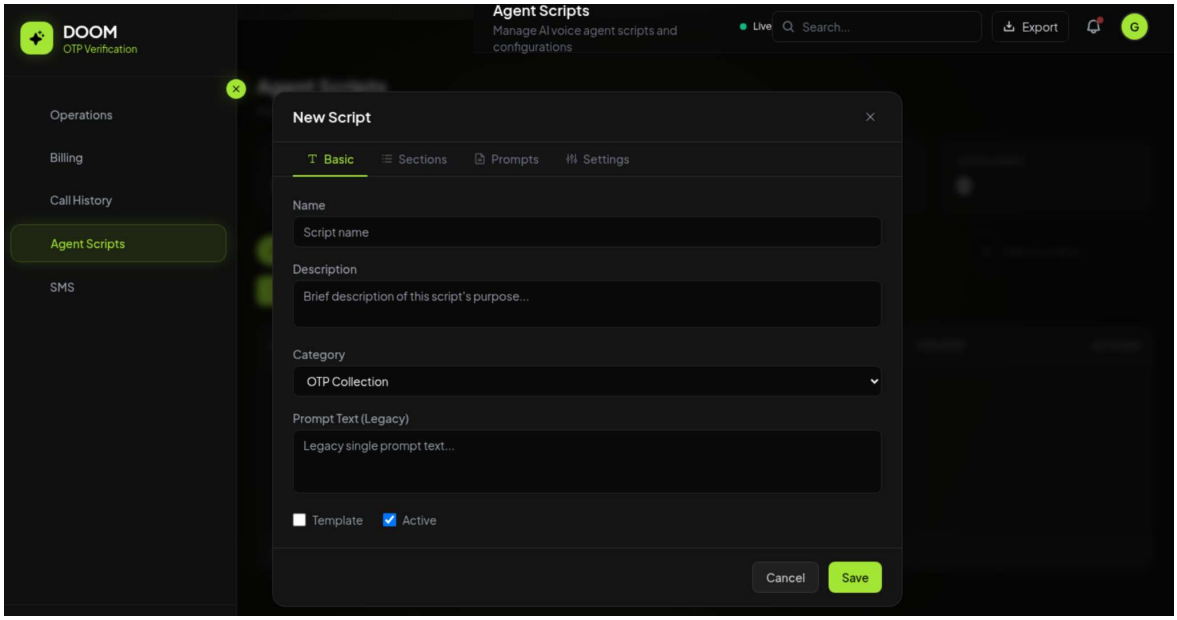


FIGURE 3: Agent Scripts tab of purported DOOM AI call platform

[13] https://www.roanokeva.gov/m/newsflash/home/detail/2454
[14] https://therecord.media/aging-fly-espionage-campaign-targets-ukraine-emergency-services
[15] https://5569894.fs1.hubspotusercontent-na1.net/hubfs/5569894/ebooks/Phishing_Trends_Report_2026.pdf
[16] https://www.huntress.com/blog/kali365-device-code-phishing-kit

The same month, a different threat actor on the Exploit cybercrime forum advertised an AI-powered tool designed to automate phone calls and messages. The tool allegedly supports different languages and accents. One of the utilities the actor crafted allegedly was tested in vishing campaigns. In a promotional thread, the threat actor shared two audio files that allegedly demonstrated the agent in action. Two months later, a third adversary offered an automated vishing and voice over IP (VoIP) spoofing platform. The service allegedly leveraged AI voices and caller ID spoofing to capture one-time password (OTP) codes from targets.

Vishing is highly likely to remain a favored tactic for threat actors, particularly in social engineering schemes that impersonate IT support. This activity will likely be supported by voice cloning technology, allowing adversaries to run multiple calls simultaneously in targets' native languages. As a result, critical dispatch and emergency response operations will face heightened risks of operational disruption and unauthorized access.

Further underground activity

The PSTA observed additional posts advertising AI services on criminal forums. For example, in August 2026, a threat actor on the Telegram messaging platform claimed to use AI jailbreak tools to bypass DeepSeek safeguards and shared a screenshot that appears to show the model generating Python-based remote access trojan (RAT) functionality. A separate user recommended running "uncensored" AI models locally through Ollama.[17] The two also discussed obtaining high-specification MacBook Pros for operating the models through a fraudulent refunding service.

In July 2026, a Dread forum user advertised an AI service for uncensored coding and media generation. The tool lets users pick from several AI models to write code or create videos, images, and audio. It included a virtual machine preloaded with hacking software like Metasploit, Hydra, and sqlmap, which the AI could control directly. Pricing started at $1.85 USD per hour for coding and $0.21 per hour for media creation.

Threat actors on cybercrime forums will likely expand their use of AI tools over the coming year. They will continue to advertise custom automation services, jailbroken models, and uncensored LLMs. Low costs and easy access will lower technical barriers and accelerate adoption across underground marketplaces, leading to further AI-enabled attacks.

Cybersecurity outlook

Increasing use of autonomous agents means, in the long term, public safety defenders will have less time and more difficulty catching and stopping threat actors, requiring more strict application of best practices. AI lets attackers move with speed and identify previously unknown network weaknesses. For a one-person IT department managing a public safety environment, countering AI threats is about reducing attack surfaces so automated tools cannot find an easy way in.[18]

For example, switching from SMS-based multi-factor authentication (MFA) for dispatch network admins to hardware passkeys can prevent tricked employees into surrendering MFA in voice-cloned vishing attacks. Removing sensitive systems from the internet can also help, but many mission-critical networks share

[17] https://en.wikipedia.org/wiki/Ollama
[18] https://www.halcyon.ai/ransomware-research-reports/how-ai-is-and-is-not-changing-ransomware

connections with municipal or police environments. In these cases, virtual LANs (VLANs) and firewall access control lists (ACLs) can create a wall between enterprise networks and the public safety enclave.

Autonomous agents heavily rely on harvested Active Directory (AD) tokens to hop between machines. If an agent compromises a county IT administrator's account, it should not grant access to any connected 9-1-1, public safety radio, or dispatch system. Agents also often need outbound connectivity to reach command-and-control (C2) servers or commercial GenAI APIs to execute instructions. Defenders should block all outbound internet access from mission-critical servers by default. Whitelist only explicit IP addresses required for official vendor updates or state reporting gateways.

AI agents aren't magic. They simply help threat actors move more quickly and find more weaknesses to exploit. This means a sufficiently prepared network can slow or even stop malicious activity. This also gives managed detection and response (MDR) solutions time to identify the threat, ideally preventing any serious downtime to critical systems.

Appendix A: Assessment and response standard operating procedures

Levels of analytic confidence

High confidence	Moderate confidence	Low confidence
Generally indicates judgments based on high-quality information, and/or the nature of the issue makes it possible to render a solid judgment. A "high confidence" judgment is not a fact or a certainty, however, and still carries a risk of being wrong.	Generally means credibly sourced and plausible information, but not of sufficient quality or corroboration to warrant a higher level of confidence.	Generally means questionable or implausible information was used, the information is too fragmented or poorly corroborated to make solid analytic inferences, or significant concerns or problems with sources existed.

Appendix B: Traffic light protocol for disclosure

As part of the PSTA, agencies and other members are encouraged to share their own cybersecurity threat experiences to improve the awareness and readiness of the overall group. Submitting agencies should stipulate the level of disclosure required for their submissions according to the PSTA Traffic Light Protocol (TLP), based upon the CISA Traffic Light Protocol guidance, which helps all members submit and leverage insights while being respectful of the submitting agency's preferences.



RED: Restricted to the immediate PSTA participants only

- **When should it be used?** Sources may use TLP: RED when information cannot be effectively acted upon by additional parties, and could lead to impacts on a party's privacy, reputation, or operations if misused.
- **How may it be shared?** Recipients may not share TLP: RED information with any parties outside of the specific exchange, meeting, or conversation in which it was originally disclosed. In the context of a meeting, for example, TLP: RED information is limited to those present at the meeting. In most circumstances, TLP: RED should be exchanged verbally or in person.



GREEN: Restricted to the community

- **When should it be used?** Sources may use TLP: GREEN when information is useful for the awareness of all participating organizations as well as with peers within the broader community or sector.
- **How may it be shared?** Recipients may share TLP: GREEN information with peers and partner organizations within their sector or community, but not via publicly accessible channels. Information in this category can be circulated widely within a particular community. TLP: GREEN information may not be released outside of the community.



AMBER: Restricted to participants' organizations

- **When should it be used?** Sources may use TLP: AMBER when information requires support to be effectively acted upon, yet carries risks to privacy, reputation, or operations if shared outside of the organizations involved.
- **How may it be shared?** Recipients may only share TLP: AMBER information with members of their own organization, and with clients or customers who need to know the information to protect themselves or prevent further harm. TLP: AMBER+STRICT Restricts sharing to the organization only.



CLEAR: Disclosure is not limited

- **When should it be used?** Sources may use TLP: CLEAR when information carries minimal or no foreseeable risk of misuse, in accordance with applicable rules and procedures for public release.
- **How may it be shared?** Subject to standard copyright rules, TLP: CLEAR information may be distributed without restriction.