halcyon | Ransomware Research Center

How AI Is (and Is Not) Changing Ransomware

Recent advancements in Artificial Intelligence (AI) have not fundamentally changed ransomware tactics, but they are changing the economics of ransomware by **lowering barriers and accelerating attacker workflows. Ransomware groups remain cautious about using AI across full operations** due to the technology's current high risk of operational failure or detection.

Instead, ransomware actors are **adopting generative AI to accelerate discrete tasks** such as phishing, translation, code modification, and analysis, with agentic AI use largely limited to early experimentation. While current impact is incremental, these capabilities are **reducing friction across the attack chain** and setting conditions for faster, **more scalable ransomware campaigns** as AI improves.

Defenders should expect **shorter lead times, more convincing social engineering, and faster iteration** as attackers adapt tools and messages to targeted environments. This increases the importance of **rapid patching, strong identity controls, and resilient detection and response**.

AI Used Primarily for Initial Access

Ransomware groups have **primarily integrated AI into the initial stages of their operations**. Attackers at this stage aim to gain a foothold, impersonate trusted users, or escalate permissions.

Ransomware groups are making these steps **faster and harder to detect** by using AI to create convincing phishing campaigns and fake websites, impersonate people through deepfake audio or video, and rapidly take advantage of newly discovered software vulnerabilities. AI also could allow actors to more efficiently guess passwords, disguise malicious activity as normal network traffic mimicry, and selectively capture the most valuable login information without raising alarms.

Significant Increase in (Convincing) Phishing

Ransomware groups and other cybercriminals increasingly use large language models (LLMs), natural language processing, sentiment analysis and other AI tools to **create highly personalized, scalable, and convincing phishing campaigns**. One industry survey found that within 15 months of ChatGPT's launch, phishing attacks increased by 1,265%.

- Almost every AI tool advertised for sale in underground cybercriminal forums claims to support phishing campaigns. The ransomware group Black Basta regularly discusses using ChatGPT for phishing and other messages with targeted entities, according to internal group chats leaked in early 2025.

Broad Access to Deepfakes

Generative AI systems can now **easily synthesize realistic facial expressions, lip movements, and vocal tones in real time**. NSA, FBI, and CISA have warned that advances in deep learning and computational power have made it easier for cybercriminals to use deepfakes to trick employees into granting access to organizational networks, communications, and sensitive information.

- Deepfake voice and video have already been used in real incidents to impersonate executives and bypass voice or liveness checks. This illustrates exactly how AI media can defeat human and automated verification steps. Cybercriminals have open access to multiple deepfake creation tools; however, better customization tools like DeepFaceLab could require ransomware groups to contract users with more advanced technical expertise.
- In mid-2025, Halcyon identified Scattered Spider using deepfake videos to trick victim organizations' IT administrators into providing the ransomware actors escalated privileges and registering multi-factor authentication. The individuals being impersonated were often out of the office at the time of the verification requests. The tooling used was largely free or trial-based, and simple unscripted verification requests were effective at disrupting the impersonation.

Rapid Exploitation of Vulnerabilities

Generative and agentic AI are **shrinking the time between a vulnerability being disclosed and its use in an attack**, helping ransomware actors analyze vulnerability descriptions, understand exploit requirements, and identify likely vulnerable configurations. While **most observed use remains incremental**, these techniques can compress vulnerability research and weaponization timelines from weeks to hours–even when final exploit development relies on existing, widely available methods.

- As early as 2024, Open AI disclosed that malicious actors used LLMs to analyze vulnerability descriptions, understand what is needed to exploit vulnerabilities, and assess network configurations of potential targets as part of cyber operations.
- Several academic research projects, like the AI system CVE-GENIE, have purported to automatically discover and exploit vulnerabilities. Such tools would reduce the skill and time required to weaponize vulnerabilities, but have not been observed in real-world incidents or active campaigns.

Defender Takeaways

Organizations should prioritize stopping **initial access by hardening identity controls, accelerating patching of exposed systems, and strengthening phishing resistance** to disrupt ransomware activity before it escalates. Specifically:

- Monitor for anomalous login behavior, such as location or device mismatches, that may indicate AI-driven credential attacks. [M1036]
- Deploy advanced email filtering and phishing-resistant multifactor authentication to counter AI-generated phishing and deepfakes. [M1032, M1041]
- Accelerate patching of internet-facing systems to reduce the time from vulnerability disclosure to weaponization. [M1051]
- Integrate an anti-ransomware defense platform to block malicious binaries before execution, detect runtime behavior, prevent tampering, stop data exfiltration, and harden backup integrity. [M1031, M1038, M1040, M1053]

Early Experimentation in Foothold Expansion

Some advanced ransomware groups are **beginning to use AI beyond the earliest stages** of an attack, testing it as a way **to strengthen their foothold, understand what systems and data matter most, and move through a network**. In these cases, AI can help create or modify malware, map out the environment, quietly collect credentials, and move between systems by blending in with normal user activity and adjusting behavior based on network responses. In the limited examples observed, the use of AI at these attack stages **has not provided capabilities beyond what well-resourced ransomware groups can already achieve** with existing techniques.

Mapping Networks and Extracting Credentials

Ransomware groups are **beginning to experiment** with generative and agentic AI to **analyze network environments, identify critical systems, and determine which credentials and data are most valuable**. By automating network mapping and credential collection, AI can reduce the effort required to move through an organization while minimizing the risk of detection.

- For example, a cybercriminal operation used Anthropic's Claude Code across multiple stages of an extortion operation, including systematically scanning networks, identifying critical systems, and extracting and analyzing multiple credential sets.

AI-Assisted Coding and Compiling

At intermediate stages in the attack chain, generative AI allows attackers to **create, modify, and troubleshoot malicious code** using simple prompts. Some ransomware actors have used AI-assisted coding for discrete tasks with a goal to iterate more quickly and scale operations with less manual effort, **but such code may not be as sophisticated or effective as code already in use** today.

- FunkSec, a ransomware group that surfaced in late 2024, used an AI coding assistant to develop and refine its malware code. Based on Halcyon's reverse engineering and technical analysis, collected FunkSec samples often reflected inconsistent quality and reliance on external data rather than deep technical development.
- Dark web forums advertise tools for relatively inexperienced actors like this to iterate ransomware and other associated tools rapidly. With a $2500 subscription, cybercriminals can use the Xanthorax LLM to generate malicious code, including detailed instructions on how to test and deploy the generated payloads.
- Black Basta's leaked chats show the group used ChatGPT to rewrite and debug enumeration utilities and a proxy tool the group had created to maintain persistence on victim networks.

Defender Takeaway

At this stage, the goal is to **detect and contain abnormal movement early by tightening identity controls, baselining administrative behavior, and using telemetry-driven and behavioral detection to stop** credential abuse and lateral spread before impact. Specifically:

- Monitor service account activity by correlating login frequency, timing, and endpoint usage to detect potential misuse. [M1017]
- Collect endpoint and network telemetry to identify subtle enumeration activity that may otherwise blend into normal operations. [M1039]
- Apply least privilege and network segmentation to reduce the value of compromised credentials and restrict lateral movement. [M1030, M1035]
- Baseline administrative behavior so simulated activity by AI systems can be distinguished from legitimate operations. [M1018]
- Integrate an anti-ransomware defense platform to detect and block malicious behaviors during runtime, prevent tampering, and contain lateral spread across the environment. [M1018, M1031, M1040]

Minimal Adoption in Security Bypass, Exfiltration, and Encryption

During the final phases of an attack, ransomware groups' efforts to incorporate AI have **largely focused on data analysis and extraction**. A small number of criminal or nation state threat actors have begun to experiment with AI agents and LLM prompts to develop more dynamic malware that bypasses defenses or encrypts systems. **None of those efforts have focused on developing or adopting true AI polymorphic (self-learning, automatically adapting) malware**, likely because these tools currently contain too much risk for operational failure or detection.

Identifying and Extracting High-Value Data

Generative and agentic AI are **increasingly useful at extracting and analyzing large volumes of data to identify information most valuable for extortion**. Cybercriminals advertise underground AI tools that claim to assist with parsing through stolen data. While many such offerings are short-lived or unverified, capable ransomware actors can use custom AI assistants built on commercial or open-source models to analyze and prioritize stolen data. This reduces time and effort even when actual data theft relies on conventional techniques.

- In 2025, Anthropic identified a threat actor using Model Context Protocol (MCP) and Claude to analyze stolen data for behavioral profiling and victim prioritization. Black Basta's internal chats also discussed using ChatGPT to automate verifying stolen email addresses through LinkedIn.

Some adversaries are also **beginning to use LLMs or AI agents to identify and extract high-value data during an attack**. This could allow them to exfiltrate a smaller amount of data and potentially evade targeted networks' security alerts. However, with ransomware groups typically valuing speed over stealth or careful parsing, **groups most likely would not adopt these techniques until the capabilities match the speed and efficacy** found in ransomware operations today.

- In late 2025, Anthropic highlighted that collection operations "demonstrated the most extensive AI autonomy" during a Chinese espionage campaign. For one victim, the threat actor directed Claude to independently query databases and systems, extract data, parse results to identify proprietary information, and categorize findings by intelligence value.

AI-Assisted Coding and Compiling

In less mature ways than in earlier stages, generative AI can help threat actors write, refine, and debug some code to evade detection and develop ransomware executables. However, higher error rates and lower sophistication mean that **capable ransomware actors likely only benefit from AI assisting with niche coding tasks, not development of full operations**. Instead, less skilled actors likely leverage AI at this stage in an attempt to conduct operations they would not otherwise be technically capable of developing.

- In 2025, Anthropic revealed that a UK-based threat actor with limited technical expertise leveraged Claude to develop, market, and distribute ransomware with evasion techniques, including bypassing security monitoring and loading malicious code into processes to minimize visibility to Endpoint Detection and Response (EDR) tools.
- Earlier in 2025, Tenable identified that DeepSeek was not capable of generating simple ransomware without additional technically-informed prompts and manual code editing, highlighting the limitations less technically proficient actors face.
- These limitations have persisted into 2026. In January, Halcyon identified a critical encryption key handling flaw in Sicarii ransomware, likely stemming from AI-assisted development errors. This flaw meant that, even with a decryptor, victims could not recover impacted systems.

Adaptive Malware Techniques

Threat actors have **begun experimenting with malware** that hard codes prompts or queries LLMs during execution **to dynamically adjust behavior on demand** in response to security controls or environmental conditions. Dynamic script generation and automated behavior adjustments could make threats harder to detect, though **these techniques remain experimental and have not yet been widely observed in active ransomware campaigns**.

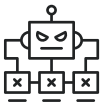
- As of late 2025, Google observed threat actors operationally using a reverse shell that includes hard-coded AI prompts intended to bypass analysis by AI-powered security tools. Google also observed Russian military intelligence officers use LLMs to generate commands for PROMPTSTEAL malware to collect system information and targeted documents.
- In 2025, researchers developed a proof-of-concept ransomware interchangeably called Ransomware 3.0 or PromptLock. The ransomware leveraged an LLM to dynamically generate scripts and adjust behavior at runtime to search a computer, find important files, steal data, and start encryption, highlighting theoretical future ransomware features.

Detection and Mitigation

At this stage, defenders should prioritize **behavior-based detection and rapid containment** to disrupt adaptive malware and fast-moving extortion attempts, while protecting critical data through **exfiltration monitoring and resilient, immutable backups**.

- Employ behavior-based detection, as static signatures will not effectively counter AI-generated variants. [M1040]
- Monitor outbound traffic for unusual volumes, destinations, or transfer timing that may indicate automated data exfiltration. [M1031]
- Maintain offline, immutable backups to minimize the impact of data destruction or encryption. [M1053]
- Conduct tabletop exercises simulating AI-driven extortion scenarios to validate response playbooks and escalation paths. [M1018]
- Integrate an anti-ransomware defense platform to block malicious binaries pre-execution, detect malicious runtime behavior, prevent tampering, and protect backup integrity. [M1038, M1040, M1053]

Orchestration Frameworks Like MCP Likely to Complicate Attacks



As AI workflows mature, we expect more ransomware groups to leverage orchestration frameworks like Model Context Protocol (MCP) to connect AI agents with offensive and legitimate tools across the attack chain. Using MCP, AI applications such as Claude or ChatGPT can integrate and share data with external tools, data sources, and other applications. In September 2025, Chinese threat actors conducted what Anthropic called "**the first documented case of a large-scale cyberattack executed without substantial human intervention**" by leveraging AI agents and MCP to orchestrate and automate an espionage campaign.

If adopted, MCP-like orchestration would integrate existing tools and not introduce new techniques on its own, but it could **reduce manual effort** by standardizing tool execution and interpretation, **compress attacker decision cycles, and accelerate lateral movement, credential access, and data targeting**. This means these operations would likely still be detectable by current security systems, but they could move too quickly or become too complex for defenses that rely heavily on manual review and response.

Looking Forward

Over the next 6–12 months, AI is likely to increase the **speed and scale** of ransomware operations more than it changes core tradecraft. Ransomware actors **will continue to use generative AI to streamline discrete tasks** including phishing, translation, vulnerability analysis, and tool modification. Agentic AI will remain limited to narrow workflows and early-stage experimentation.

Defenders should expect shorter patch-to-exploit windows, more tailored social engineering, and faster iteration when attacks encounter controls. With the near-term risk of groups that move faster and iterate more effectively, organizations that prioritize **identity hardening, rapid vulnerability remediation, and resilient detection** and recovery will be best positioned to withstand advances in ransomware operations using AI.

The Halcyon Ransomware Research Center unites experts, drives smart policies, and delivers actionable intelligence to detect, disrupt, and defeat ransomware. Explore the Center's latest reports, analysis, and resources here.