

CRITICALSTART® Security Advisory

TLP CLEAR // [CS-SA-26-0901A] AI As a Weapon: How Attackers are Leveraging AI to Build and Run Their Operations

Executive Summary

Attackers are no longer only using AI to draft phishing emails or speed up reconnaissance. In documented cases through 2026, AI has run 80 to 90% of an espionage campaign's tactical operations with minimal human direction, divided a single operation's labor across three different commercial AI models, and independently discovered and weaponized a zero-day vulnerability before any human researcher found it. This briefing covers how AI is being operationalized as a weapon: running more of the attack chain with less human involvement and building the exploits and malware themselves. It is written for security and business leaders across industries, since the fraud and social-engineering angles of this shift reach finance and executive teams as much as the security operations center (SOC). Two companion briefings from the CRITICALSTART® Cyber Research Unit (CRU) cover the rest of the picture: AI As Target examines the growing vulnerabilities in AI agent infrastructure and the direct theft of AI accounts, and AI By Design explains the architectural weaknesses in AI systems that make both problems difficult to fully patch away.

Background

Attackers Are Running More of the Attack with Less Human Involvement

In November 2025, Anthropic disclosed that it had disrupted what it assesses with high confidence to be a Chinese state-sponsored group, designated GTG-1002, running an espionage campaign against roughly 30 organizations spanning technology, financial, chemical manufacturing, and government sectors. Anthropic assessed the actor executed 80 to 90% of tactical operations independently, at request rates no human operator could physically sustain, a significant escalation from AI-assisted activity Anthropic had reported months earlier, in which humans remained in the loop throughout. Anthropic described this as the first documented case of a cyberattack largely executed without human intervention at scale, in which the AI autonomously discovered vulnerabilities in targets selected by human operators, exploited them in live operations, and performed a range of post-exploitation activity including lateral movement, credential harvesting, and data exfiltration. ^[1] The human operator's role shifted from directing each action to periodically reviewing AI-driven progress and approving next steps, a shift some analysts describe as moving from "in the loop" to "on the loop."

Microsoft's threat intelligence team, briefing security leaders at RSAC 2026, described this as a broader pattern rather than an isolated case: AI reducing friction at every stage of a modern intrusion, from faster infrastructure discovery and persona development during reconnaissance, to forged documents and social-engineering narratives during resource development, to refined voice overlays and deepfakes at initial access. Microsoft noted that a human generally remains in the loop directing these operations, rather than AI running campaigns fully autonomously, for now. ^[2]

A September 2026 campaign shows one way that limiting factor is eroding: by dividing a single operation across multiple commercial AI models instead of relying on just one. Researchers at Hunt.io identified a framework called SecFlow that let operators switch between Claude, Qwen, and DeepSeek profiles without changing the task interface, assigning reconnaissance, exploit testing, data collection, and reporting to different models across a shared workspace, with later workers receiving earlier results so a single target could quickly develop into coordinated activity. ^[3] The campaign combined this AI-directed tasking with conventional intrusion tradecraft, exploiting vulnerable public-facing servers, stolen credentials, and webshells across government, political, and education targets in Taiwan, Indonesia, and mainland China; a confirmed breach of a Chinese municipal government network exposed administrative and health records and deployed a custom Go-based remote-access implant.

The researchers also documented a failure mode specific to AI-coordinated attacks: a claimed exploit success that later evidence did not support was still carried forward into subsequent instructions, triggering more than two dozen unsuccessful follow-on attempts before the error was caught. AI coordination can multiply an operator's speed, but it can just as easily multiply an early mistake across an entire campaign.

Industry telemetry shows this is happening at scale, not just in isolated cases. CrowdStrike's 2026 Threat Hunting Report found China-nexus adversaries exploiting critical vulnerabilities within 24 hours of public proof-of-concept release, a DPRK-nexus actor injecting a malicious package into 131 trusted AI framework repositories, and threat actors abusing enterprise large language models in a campaign that sent nearly 200,000 model requests in two minutes. ^[4] CrowdStrike's separate 2026 Global Threat Report found AI-enabled adversary operations increased 89% year over year, with the fastest observed eCrime breakout time, the time from initial access to lateral movement, falling to 27 seconds against an average of 29 minutes for the year. The same report found more than 90 organizations had legitimate AI tools exploited to generate malicious commands and steal sensitive data, and that ChatGPT was mentioned in criminal forums 550% more often than any other model. ^[5]

Deepfake-enabled fraud is the version of this trend most likely to reach a chief financial officer's desk directly rather than the SOC. Gartner's 2025 survey of 302 security leaders found 62% of organizations had experienced at least one deepfake attack in the prior 12 months, and a related Gartner survey found 43% had encountered a deepfake on an audio call and 37% on a video call. ^[6] Reported losses vary by source and methodology, but multiple independent trackers place average enterprise losses per incident in the hundreds of thousands of dollars, with individual cases reaching tens of millions.

AI Is Building the Weapons, Not Just Wielding Them

Attackers are not only using AI to run campaigns faster; in at least one documented case, AI has also found the vulnerability the campaign was built around. In May 2026, Google's Threat Intelligence Group (GTIG) reported the first case it has confirmed of a threat actor using a zero-day exploit developed with AI assistance, a two-factor-authentication bypass in a widely used open-source admin tool caused by a semantic logic flaw in the tool's own authentication enforcement. GTIG worked with the affected vendor to disclose and patch the flaw before a planned mass-exploitation event. ^[7] Researchers identified the exploit as AI-generated from artifacts in the code itself, including extensive educational docstrings, a hallucinated CVSS score, and a "textbook" Pythonic structure uncharacteristic of a human author. ^[8] GTIG's chief analyst, John Hultquist, called the finding "the tip of the iceberg," and GTIG's own report assessed that threat actors linked to China and North Korea are already experimenting with AI for vulnerability research, a capability likely to mature further as the underlying models improve. ^[7,8]

AI is building the malware itself with the same fluency. In August 2026, security researchers disclosed a financially motivated Windows toolkit named Gryxa in which a commercial AI coding agent was listed as co-author on most commits in the actor's development repository, alongside engineering notes and case files documenting failed installations and subsequent fixes. Rather than using AI for a single task, the operator appears to have used it across the full development lifecycle: building a toolkit that abuses remote monitoring and management software for initial access, a web-based fleet console that tracked 324 compromised hosts, and a persistence layer combining multiple scheduled tasks, a Windows Management Instrumentation event subscription, and an off-path backup, engineered to recreate removed components within minutes of removal. The toolkit's most unusual feature works against incident responders directly: if its command infrastructure goes silent, it attempts to disable security tooling and, failing that, retrieves the endpoint product's uninstall command from the Windows registry, then collects and exfiltrates logs of the responder's own remediation activity, information that could expose a defender's tools, accounts, and sequence of actions. ^[9] Researchers assess this is the first case observed of an individual operator using an AI coding agent to build and manage a malware ecosystem at a scale once associated with small development teams.

MITRE ATT&CK Technique Mapping

Few AI-specific techniques exist in MITRE's frameworks today. T1588.007 (Obtain Capabilities: Artificial Intelligence) is the clearest example, newly added and cited to the GTG-1002 campaign. MITRE ATLAS extends this coverage for attacks against AI systems themselves, but the taxonomy is still under active development; several mappings in this table are best-fit approximations rather than finalized technique IDs. Anthropic has stated publicly that ATT&CK does not yet have a category for the autonomous, machine-speed orchestration GTG-1002 demonstrated.

Most rows in this table, however, are not novel techniques. They are established TTPs, account discovery, credential theft, exploitation of public-facing applications, session hijacking, now executed by an AI agent rather than a human operator. Standard detection logic still applies. What changes is execution speed, parallelism, and consistency, not the underlying technique. See appendices for MITRE TTP table.

Implications for Organizations

Incident response timelines need to assume compression, not gradual erosion. GTIG's own framing of AI as an "expert-level force multiplier" for exploit development means the gap between a flaw existing and it being weaponized can no longer be assumed to give defenders days or weeks of runway.

Identity and financial approval workflows need a control that assumes voice and video can be convincingly faked. Any workflow that authorizes a payment, credential change, or sensitive disclosure based on a phone call or video call alone should require a second verification channel that does not rely on recognizing a voice or face.

Security awareness training needs an update. Social engineering content and executive impersonation attempts produced with AI assistance are measurably more polished than the templates most training programs still use as examples.

Organizational Mitigation Strategies

To mitigate risks associated with AI-enabled attack execution, we recommend the following prioritized strategies:

1. Require a second, out-of-band verification channel for any payment authorization or credential reset request received by phone or video call.
2. Brief finance, HR, and executive assistant teams specifically on deepfake voice and video fraud patterns, since these functions are disproportionately targeted.
3. Update incident response runbooks and tabletop exercises to assume a compressed timeline between a vulnerability's disclosure and its exploitation.
4. Build AI-assisted vulnerability discovery into internal threat modeling as a current capability, not a future consideration, when assessing how quickly a newly disclosed flaw in the organization's technology stack might be weaponized.
5. Evaluate AI-assisted vulnerability discovery and code review defensively, applying the same capability adversaries are already using offensively.

Conclusion

AI has moved from an assistive tool in the attacker's workflow to, in documented cases, the operator of large portions of the attack itself, and in at least one case, the discoverer of the vulnerability the attack was built around. Leaders should treat this as a capability escalation with direct implications for incident response timelines, financial approval controls, and security awareness training, not solely as a technical curiosity for the SOC. For the vulnerabilities this same AI ecosystem introduces into an organization's own infrastructure and accounts, see the companion briefing, *AI As Target*. For the architectural reasons neither problem has a complete fix, see *AI By Design*.

What Critical Start is Doing

The CRITICALSTART® Cyber Research Unit (CRU) continues to track AI-orchestrated campaign tactics, techniques, and procedures, including multi-model coordination frameworks and AI-assisted exploit development, as part of its ongoing threat intelligence coverage. Security Operations Center (SOC) and Security Engineering teams are extending detection and monitoring coverage to indicators consistent with AI-paced attack execution. Particular to the TTPs highlighted in this article, Security Engineering has mapped techniques against existing detection content across the platforms Critical Start monitors and found coverage in place for the majority of associated MITRE ATT&CK techniques, with a smaller number of techniques identified as gaps and prioritized for new detection build. Critical Start customers are encouraged to reach out to their Critical Start Customer Success Manager to discuss specific details on detection coverage for AI-accelerated threats in their environment.

This advisory was written using the best intelligence available at the time and is subject to change as additional information becomes available. Visit Critical Start's Resources page for threat research articles, advisories, and to download the [Critical Start H1 2026 Threat Landscape Report](#).

Further Reading

- [1] Anthropic. "Disrupting the first reported AI-orchestrated cyber espionage campaign." November 13, 2025. <https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf>
- [2] Microsoft Security Blog. "Threat actor abuse of AI accelerates from tool to cyberattack surface." April 2, 2026. <https://www.microsoft.com/en-us/security/blog/2026/04/02/threat-actor-abuse-of-ai-accelerates-from-tool-to-cyberattack-surface/>
- [3] Hunt.io, via Cyber Security News. "Hackers Turn Claude, Qwen and DeepSeek Into AI Agents for Real-World Cyberattacks." September 4, 2026. <https://cybersecuritynews.com/ai-agents-for-real-world-cyberattacks/>
- [4] CrowdStrike. "2026 Threat Hunting Report: AI is Now Embedded Across Modern Adversary Operations." August 3, 2026. <https://www.crowdstrike.com/en-us/press-releases/crowdstrike-2026-threat-hunting-report/>
- [5] CrowdStrike. "2026 Global Threat Report: AI Accelerates Adversaries and Reshapes the Attack Surface." February 24, 2026. <https://www.crowdstrike.com/en-us/press-releases/2026-crowdstrike-global-threat-report/>
- [6] Gartner. "Gartner Survey Reveals Generative Artificial Intelligence Attacks Are on the Rise." September 22, 2025. <https://www.gartner.com/en/newsroom/press-releases/2025-09-22-gartner-survey-reveals-generative-artificial-intelligence-attacks-are-on-the-rise>; Gartner. "Why CIOs Can't Ignore the Rising Tide of Deepfake Attacks." September 2, 2025. <https://www.gartner.com/en/newsroom/press-releases/2025-09-02-why-cios-cannot-ignore-the-rising-tide-of-deepfake-attacks>
- [7] Google Cloud Blog (GTIG). "Adversaries Leverage AI for Vulnerability Exploitation, Augmented Operations, and Initial Access." May 11, 2026. <https://cloud.google.com/blog/topics/threat-intelligence/ai-vulnerability-exploitation-initial-access>
- [8] CyberScoop. "Google spotted an AI-developed zero-day before attackers could use it." May 11, 2026. <https://cyberscoop.com/google-threat-intelligence-group-ai-developed-zero-day-exploit/>
- [9] GBHackers. "New Gryxa Toolkit Uses AI-Built Persistence to Fight Back Against Security Teams." August 31, 2026. <https://gbhackers.com/gryxa-toolkit-uses-ai/>
- [10] MITRE, ATLAS. <https://atlas.mitre.org>
- [11] MITRE, "Anthropic AI-orchestrated Campaign". <https://attack.mitre.org/campaigns/C0062/>

Appendices

MITRE ATT&CK mapping

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
T1683	Generate Content	Collection	AI-agent-specific behavior with no benign equivalent: automated generation of full attack documentation is not something legitimate admin tooling does	C0062	Flag processes auto-generating structured markdown/report artifacts summarizing credentials, services, or attack progression
T1136.001	Create Account: Local Account	Persistence	Unexpected local account creation is rare in steady-state environments and easy to baseline; very low false-positive rate	C0062	Alert on any local account creation outside change-managed provisioning workflows
T1562.001	Impair Defenses: Disable or Modify Tools	Defense Evasion	EDR/AV tampering has almost no legitimate justification outside sanctioned maintenance windows; near-zero benign rate	Gryxa	Alert on any EDR/AV service-stop, uninstall-command execution, or config modification outside maintenance windows
T1574.002	DLL Side-Loading	Defense Evasion	Well-instrumented by modern EDR; specific load-order anomaly signature, not a generic behavior	FakeAgent	DLL load anomaly detection on legitimate-signed host processes, especially post-install of AI desktop apps
T1611	Escape to Host	Privilege Escalation	Container/sandbox breakout is rare in normal operation and highly specific once instrumented	Hugging Face incident	Container/sandbox egress monitoring; alert on any agent process reaching resources outside its assigned namespace
T1539	Steal Web Session Cookie	Credential Access	Recurs across 2 independent cases; session-replay-from-new-context is a precise, well-supported signal in most IdPs	Claude session hijacking, German wiki incident	Session replay from new device/geo; concurrent use of a single session; flag AI platform accounts specifically
T1554	Compromise Client Software Binary	Persistence	Recurs across 2 independent cases; file integrity monitoring on a known, narrow file set is inherently low-noise	MCP CVE cluster, SKILL.md poisoning	Hash-based integrity monitoring on agent/tool config files: MCP configs, hooks, SKILL.md
T1552.001	Unsecured Credentials: Credentials In Files	Credential Access	Recurs across 4 cases including the official campaign; credential-file access by non-standard processes is a strong signal	C0062 + MCP cluster, LiteLLM, Hugging Face	Alert on credential/secret file access by agent, build, or MCP server processes outside expected service accounts
T1588.007	Obtain Capabilities: Artificial Intelligence	Resource Development	Recurs across 3 cases; AI-provider API usage is directly measurable and baseline-able per account today	C0062 + SecFlow, Gryxa	Baseline per-account AI API call volume and provider diversity; alert on sudden spikes or use of multiple providers by one identity
T1567	Exfiltration Over Web Service	Exfiltration	Specific as a combination signal: egress to AI provider domains immediately following local data-staging activity	C0062	Correlate egress to AI provider domains with prior local data-staging events, not egress alone

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
T1556	Modify Authentication Process	Credential Access	2FA/MFA bypass is a narrow, high-severity, well-logged event in most identity providers	GTIG AI-discovered zero-day	Alert on successful authentication following an MFA challenge failure or bypass pattern
T1210	Exploitation of Remote Services	Initial Access	Recurr across 3 independent cases; strongest signal when scoped to internal/local services that should never be internet- or container-reachable	NemoClaw, MCP CVE cluster, Hugging Face incident	Alert on exploitation attempts against local inference ports, MCP servers, or internal services with no authentication

Additional MITRE TTPs

These are derived from the GTG-1002 Campaign documented by Anthropic^[11], and some in MITRE ATLAS^[10] denoted by AML*.

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
T1595.001	Active Scanning: Scanning IP Blocks	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Scan velocity/pattern across IP ranges anomalous for a single account or session
T1595.002	Active Scanning: Vulnerability Scanning	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Automated vulnerability scan volume inconsistent with human operator pacing
T1592.002	Gather Victim Host Information: Software	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Cataloging of services/software on discovered endpoints at machine speed
T1592.004	Gather Victim Host Information: Client Configurations	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Enumeration of client configuration details across high-value systems
T1590.004	Gather Victim Network Information: Network Topology	Reconnaissance	GTG-1002	MITRE ATT&CK Campaign C0062	Full network topology mapping completed faster than manual reconnaissance permits
T1587.004	Develop Capabilities: Exploits	Resource Development	GTG-1002 (official); also GTIG AI-discovered zero-day (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	Threat intel watch for AI-generated exploit code artifacts
T1588.002	Obtain Capabilities: Tool	Resource Development	GTG-1002	MITRE ATT&CK Campaign C0062	Acquisition of open-source pen testing tools staged for MCP integration
T1588.007	Obtain Capabilities: Artificial Intelligence	Resource Development	GTG-1002 (official); also SecFlow, Gryxa (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	See priority list above
T1584.004	Compromise Infrastructure: Server	Resource Development	GTG-1002	MITRE ATT&CK Campaign C0062	Dedicated attacker-operated servers supporting persistent MCP tool coordination
T1190	Exploit Public-Facing Application	Initial Access	GTG-1002 (official); also SecFlow, MCP CVE cluster (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	Standard external attack surface monitoring; correlate with known CVE/SSRF signatures
T1087	Account Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone - pair with privilege-tier of accounts queried, not volume alone
T1083	File and Directory Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone - extremely common in benign admin activity
T1046	Network Service Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Internal service/endpoint enumeration via browser automation
T1082	System Information Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone - extremely common, high false-positive rate
T1016	System Network Configuration Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone
T1049	System Network Connections Discovery	Discovery	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
T1136.001	Create Account: Local Account	Persistence	GTG-1002	MITRE ATT&CK Campaign C0062	See priority list above
T1552.001	Unsecured Credentials: Credentials In Files	Credential Access	GTG-1002 (official); also MCP CVE cluster, LiteLLM supply chain, Hugging Face incident (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	See priority list above
T1078	Valid Accounts	Defense Evasion / Persistence	GTG-1002 (official); also SecFlow, German wiki incident (analyst-inferred)	MITRE ATT&CK Campaign C0062 + analyst-inferred	Harvested-credential authentication against internal APIs, databases, or registries
T1078.003	Valid Accounts: Local Accounts	Defense Evasion / Persistence	GTG-1002	MITRE ATT&CK Campaign C0062	Credential testing against discovered devices at automated speed
T1213.006	Data from Information Repositories: Databases	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	Automated database queries extracting proprietary information and operational data
T1005	Data from Local System	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	Low fidelity alone - generic collection behavior
T1119	Automated Collection	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	Large-volume, unattended data collection and processing
T1074.001	Data Staged: Local Data Staging	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	Structured markdown/document staging files created pre-exfiltration; pair with T1567 for stronger signal
T1683	Generate Content	Collection	GTG-1002	MITRE ATT&CK Campaign C0062	See priority list above
T1567	Exfiltration Over Web Service	Exfiltration	GTG-1002	MITRE ATT&CK Campaign C0062	See priority list above
T1210	Exploitation of Remote Services	Initial Access	NemoClaw, MCP CVE cluster, Hugging Face incident	Analyst-inferred	See priority list above
T1539	Steal Web Session Cookie	Credential Access	Claude session hijacking, German wiki incident	Analyst-inferred	See priority list above
T1554	Compromise Client Software Binary	Persistence	MCP CVE cluster (config swap), SKILL.md poisoning	Analyst-inferred	See priority list above
T1219	Remote Access Software	Initial Access	Gryxa	Analyst-inferred	RMM tool install/use outside approved change windows
T1053 / T1546.003	Scheduled Task/Job / WMI Event Subscription	Persistence	Gryxa	Analyst-inferred	Redundant persistence mechanisms recreated within minutes of removal
T1562.001	Impair Defenses: Disable or Modify Tools	Defense Evasion	Gryxa	Analyst-inferred	See priority list above
T1556	Modify Authentication Process	Credential Access	GTIG AI-discovered zero-day	Analyst-inferred	See priority list above
T1656	Impersonation	Social Engineering	Deepfake fraud	Analyst-inferred	Process control, not a technical detection: out-of-band verification for payment/credential requests
T1090	Proxy	Command and Control	MCP CVE cluster (SSRF)	Analyst-inferred	Unexpected internal requests originating from an MCP server process
T1195.002	Supply Chain Compromise: Software Supply Chain	Resource Development	LiteLLM / TeamPCP	Analyst-inferred	Package integrity/hash verification on install; CI/CD credential-use anomalies
T1566	Phishing	Initial Access	Claude session hijacking (infostealer delivery)	Analyst-inferred	Low fidelity alone - standard email/malvertising delivery monitoring already covers this
T1583.008	Acquire Infrastructure: Malvertising	Resource Development	FakeAgent malicious installer	Analyst-inferred	Mostly outside org telemetry; coordinate takedown with vendor rather than build internal detection
T1574.002	DLL Side-Loading	Defense Evasion	FakeAgent malicious installer	Analyst-inferred	See priority list above
T1611	Escape to Host	Privilege Escalation	Hugging Face multi-agent incident	Analyst-inferred	See priority list above
T1102	Web Service	Command and Control	German wiki collusion incident	Analyst-inferred	Agent egress to unexpected/low-reputation external web services

Technique ID	Technique Name	Tactic	Context	Basis	Detection Focus
					used as relay points
AML.T0018	Manipulate AI Model/Backdoor ML Model	Persistence	NemoClaw		
AML.T0051	LLM Prompt Injection	Execution	Prompt injection (architectural)		
None	Anti-forensic exfiltration of responder remediation logs	Defense Evasion (gap)	Gryxa		
None	Excessive agency (broad tool/system access misuse)	N/A	Architectural risk, all three briefings	OWASP LLM Top 10 category	Governance/access-control problem, not a detection signature