

OCTOBER

pellera

THREAT INTEL REPORT 2026

Prepared by: Pellera Threat Intel Team
pellera.com | 800.747.8585

Driving Momentum. Accelerating Change. Empowering IT Transformation.

Pellera Technologies was born out of the combined expertise of Converge Technology Solutions and Mainline Information Systems, two industry leaders with over 35+ years of experience and a shared vision for innovation. Together, we empower businesses to achieve greater efficiency, adaptability, and growth for today and tomorrow.

Our commitment is to reshape what's possible with IT, offering advanced solutions in digital infrastructure, cloud, cybersecurity, and AI. We don't just deliver technology—we partner with you to build tailored strategies designed to simplify complexities, unlock opportunities, and drive transformational outcomes.

At Pellera, momentum builds here through collaborative, people-first technology designed to fuel progress and deliver measurable impact.

pellera

Observations for October 2026

The attacks that defined 2026 doesn't need a software flaw. It needs a phone call. A caller rings an employee, says IT is updating the company's MFA settings, and sends them to a sign-in page dressed up as the company's own. The employee types in their password and code, and the caller registers a device of their own on the account. From there the crew searches the company's cloud apps for documents containing words like "confidential" and "vpn" and pulls customer records out of Salesforce. Then a **ShinyHunters-branded extortion email arrives**: pay in bitcoin within 72 hours. Google's investigators found none of it relied on a flaw in anyone's software. **A borrowed login, stolen data and a deadline: that is 2026 in one case.**

Every step of that chain got easier this year. Getting in is the part attackers have turned into a product. **Tools that beat MFA are rented for days at a time** rather than built, and when a tool fails, a phone call to the help desk works instead. Once inside, attackers move fast. **The average intruder reaches the rest of the network in under half an hour, and one crew went from taking over an account to stealing data in under five minutes**. AI is part of the reason, but it is speeding up familiar attacks, not inventing new ones.

The payoff has moved too. **Encryption is now optional; the threat is publication.** August was the busiest month of the year on criminal leak sites, and no single gang runs the market, so taking one down changes little. Refusing to pay doesn't end it. Berlin refused, and its whole archive went public. Paying doesn't reliably end it either, since more than a third of payers get a second demand. Backups fix encrypted servers. They do nothing about data that is already gone.

That puts the year-end priorities on the login and the data. Move the accounts that reach the most systems onto a sign-in a fake page can't capture, and treat a password reset as unfinished until every session is revoked. Give staff a way to check the help desk is really the help desk. **Patch what faces the internet, keep logs where an attacker with admin rights can't wipe them, and find out what your own AI tools can read and do.** None of this is new, and that is the point. The attacks that worked this year mostly didn't need anything new either.

This October, Pellera is equipping organizations with expert insights, threat intelligence, and security best practices to build stronger cyber resilience.

Visit Pellera's Cybersecurity Awareness Month hub for resources, educational content, and actionable guidance to help defend against today's evolving threats.

Learn More



Executive Overview

HIJACKED SIGN-INS

Audience

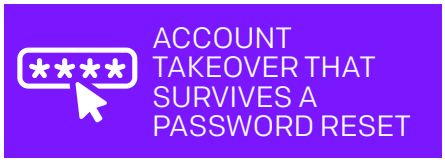
- CISOs and security leaders
- Identity and access management teams
- SOC analysts and incident responders
- IT help desk and end-user support managers
- Microsoft 365 and cloud platform administrators



RISK



GEOGRAPHIC SCOPE



BUSINESS IMPACT

Attackers don't break in anymore. They log in, usually as someone else, with a stolen or guessed password or a session they phished. Often there is no second check at all. Where MFA is on, it no longer settles the matter, because the tools that beat it are now sold by subscription to anyone who wants one.

Those tools sit between the employee and the real Microsoft sign-in page. They pass the password and the MFA code through and keep the signed-in session that comes back, which keeps working after a password reset. A newer method shows no fake page at all: the employee is sent to Microsoft's genuine site to type in a code and unknowingly hands the attacker a working token. When the tools fail, crews call or message staff as the help desk and talk them into giving up a code, enrolling the attacker's device, or handing over remote control. The prize is no longer the password. It is the live session. If your plan for a phished account is a password reset, the attacker is still inside.

READ MORE

RANSOMWARE WITHOUT A KINGPIN

Audience

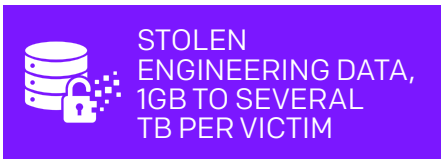
- CISOs and security leaders
- SOC analysts and incident responders
- Vulnerability management and patching teams
- Network, identity, and infrastructure administrators
- Backup and disaster recovery owners



RISK



GEOGRAPHIC SCOPE



BUSINESS IMPACT

Ransomware used to mean locked files. More and more, it means leaked ones. Crews break in, take what they can, and name the victim on a public leak site if it won't pay. August was the busiest month of the year for those postings, and the market no longer has a leader. Dozens of crews are active, many of them less than a year old, and half of August's most active groups weren't on the list in July. There is no single gang whose tools a defender can study and block, so the defense has to be general.

Because the threat is the stolen data, backups don't end the incident. Berlin's state government refused to pay and watched its whole archive go public, and a large share of the companies that do pay get a second demand. Ransom demands are falling while the cost of recovery keeps rising. The way in is ordinary: a phishing email, a stolen login, or

an internet-facing system nobody patched, sometimes attacked the day after the flaw becomes public. One unpatched product can produce dozens of victims at once, as Clop showed again with PTC's product-design software, naming Shell, GE and Philips among them. If your recovery plan ends at restoring from backup, it ends too early.

READ MORE

Audience

- **CISOs and security leaders**
- **SOC analysts and threat hunters**
- **Email security and messaging teams**
- **DevSecOps and platform engineering leads**
- **Cloud and identity administrators**
- **AI platform owners and agent developers**

ATTACKS THAT MOVE IN MINUTES



AI hasn't handed attackers anything new. It has handed them speed, and made the old tricks cheaper and more convincing. The average intruder now reaches the rest of a network in under half an hour, and one attacker used an AI agent to plan, build and run a credential-theft campaign in under six hours. Less skilled operators can now run attacks that once took a capable team. Phishing written with AI is harder for staff to spot, and most ransomware victims Proofpoint surveyed said AI made the attack more effective. What nobody has seen yet is a machine running a whole attack alone; AI is helping at every stage, with a person still directing it.

The other change is that the AI tools companies have adopted are now targets themselves. Coding agents went from almost no companies to most of them in a year, and they read code, secrets and internal data on a developer's behalf. Attackers are poisoning the add-ons those agents trust, slipping instructions into what they read, and stealing the AI accounts and cloud capacity they run on. A company that adopted AI assistants quickly may not know what they can reach, or who else is giving them instructions. Treat them like any other privileged account. Know what they can touch and who can talk to them.

READ MORE

Tactical Guidance

HIJACKED SIGN-INS

Overview & Impact

Most of this runs on an adversary-in-the-middle proxy. Tycoon2FA sat between the user and the real Microsoft sign-in page, relayed the password and MFA code in real time, and kept the session cookie that came back. That cookie kept working after a password reset unless someone revoked the session. It beat SMS codes, one-time passcodes, and push prompts alike, behind rotating CAPTCHAs and Cloudflare-hosted domains that lived 24 to 72 hours. BigBear 2.0 went further: CloudSEK found custom JavaScript that breaks the browser's FIDO2 prompt so the victim falls back to a weaker method.

Device code phishing skips the fake page entirely. The attacker's backend asks Microsoft for a real device code, shows it to the victim on a DocuSign- or voicemail-themed page, and sends them to the genuine microsoft.com/devicelogin to type it in. There is no bad domain to spot, and the attacker walks away with tokens that open Outlook, Teams, and OneDrive with no password and no second prompt. Proofpoint saw new kits appear weekly, and the FBI warned publicly about one, Kali365. The human-run version is vishing. Mandiant's UNC6661 called employees claiming IT was updating MFA settings and harvested SSO credentials on companynamesso.com lookalikes; Okta's O-UNC-066 appears to do the same with passkeys. Microsoft's Teams campaign skips credentials altogether: an external tenant poses as the help desk, talks the user into Quick Assist, and installs a Node.js implant.

- The pressure doesn't stay in the inbox. The extortion crew that follows these vishing intrusions has also sent text messages to employees, and Google has reports of victims' websites being hit with denial-of-service attacks.
- A single fake help-desk chat on Teams ended with the attacker taking screenshots of the user's desktop and moving over WinRM to domain controllers and certificate authorities.
- Attackers compromised around 5,000 Dropbox accounts in the first three weeks of August through a Lenovo ID sign-on path with no enforced two-factor authentication, and viewed or downloaded files from about 1,500 of them.

Observations

- In Sophos's three-wave phishing case, a 90-day default session token lifetime let a compromised account skip further MFA prompts, and the campaign touched over one-fifth of employees.

Guidance

Strategic Intelligence

Trend

- **MFA bypass by subscription:** Tycoon2FA panel access started at $120 for 10 days on Telegram and Signal. EvilTokens and Kali365 sell on Telegram too, and BigBear leases its panel to at least five affiliates. The operator needs a wallet, not skill.
- **Enrollment is the target:** Mandiant's UNC6661 talked victims through a fake MFA setup and registered its own device on the real account, and Okta assesses O-UNC-066's kit is built to do the same with passkeys. Once that factor exists, a password reset changes nothing.

Operational Intelligence

Threat Vectors

- **Proxy sign-in pages:** Tycoon2FA and the Evilginx2-based BigBear relay credentials and MFA codes to Microsoft in real time and keep the session cookie. Lures arrive as .svg, .pdf, .html, and .docx attachments with QR codes and bounce through Azure Blob Storage, Firebase, or Wix redirects first.
- **Device codes:** The victim types an attacker-obtained code into the real microsoft.com/devicelogin. Lures impersonate DocuSign reviews, Teams voicemail alerts, and Mimecast device enrollment, hosted on AWS S3 buckets and Cloudflare Workers domains.
- **Consent prompts:** The FBI reports OAuth consent phishing over commercial messaging apps since late 2025, aimed at prominent people and their families. The granted token only dies when revoked in app security settings, not on a password change.
- **Phone and Teams:** Crews call as IT staff about updating MFA settings or registering a passkey, or open an external Teams chat as the help desk and walk the user into Quick Assist. PowerShell then installs an MSI with msiexec /qn so the user sees nothing.

Monitoring & Detection Gaps

- **Real domain, real token:** Device code victims authenticate on Microsoft's own infrastructure and the tokens are genuine, so domain reputation checks pass. Proofpoint's warning: training users to check the URL doesn't help here.
- **Deleted evidence:** in at least one case UNC6661 deleted Okta's "Security method enrolled" email and its own outbound phishing; Tycoon2FA operators add inbox rules. Vishers keep instructions on the phone so nothing lands in Teams chat logs.

Response Actions

- **Revoke, don't just reset:** Reset the password and revoke every active session and refresh token, then force re-authentication for high-privilege accounts. Dropbox's fix was to expire every session authenticated through Lenovo IDs.
- **Clear the hiding places:** Review and delete unauthorized inbox rules, which attackers add to hide their activity, and revert any altered direct deposit details. A missing enrollment alert proves nothing: one crew deleted Okta's 'Security method enrolled' email outright.
- **Revoke app consents:** Audit third-party OAuth consents across the Microsoft 365 tenant for suspicious names, recent authorization dates, or unknown publishers and revoke them; consent-phished tokens survive password changes.
- **Assume host compromise:** For a Teams or Quick Assist intrusion, treat the endpoint as a network foothold and rotate every credential reachable from it, including domain admin accounts if the host was domain-joined.

Tactical Intelligence

Mitigation Strategies

- **Phishing-resistant MFA first:** Move to FIDO2 security keys, Windows Hello for Business, or Microsoft Authenticator passkeys, and use Entra Conditional Access authentication strength to require them for privileged roles and critical apps.
- **Bind the session:** Turn on Token Protection to bind session tokens to the device and enforce frequent MFA re-authentication; a long default token lifetime is what let a compromised account skip further prompts in the Sophos case above.
- **Lock down enrollment:** Constrain who can add or modify authenticators by network context, device management status, and existing factors, and notify users every time a sign-in method is added or changed, so an attacker's passkey gets seen.

Tactical Guidance

Preventive Measures

- **Block device code flow:** Create a Conditional Access policy using the Authentication Flows condition to block device code flow for all users. Run it in report-only mode first, audit existing usage, exclude emergency access accounts, and block authentication transfer too.
- **Help-desk pass phrases:** Establish internal help-desk authentication phrases, publish how staff verify a caller claiming to be IT, restrict Teams external access to trusted domains, and make the rule absolute: no remote access for unsolicited contacts.

Threat Hunting Hypotheses

Replayed session cookies after proxy phishing

Hypothesis: A phished session cookie is being replayed from an unmanaged browser at a risky sign-in, and the attacker follows up with an inbox rule or a new authenticator.

Investigation Steps

- **Risky unmanaged sign-ins:** In AADSignInEventsBeta over 7 days, filter IsManaged != 1, IsCompliant != 1, RiskLevelDuringSignIn in (50, 100), ClientAppUsed == "Browser", and empty DeviceTrustType, DeviceName, and AadDeviceId.
- **Click before sign-in:** Pivot to UrlClickEvents for ActionType ClickAllowed by the same AccountUpn just before the risky sign-in, and check Entra ID Protection for Anomalous Token and Unfamiliar sign-in properties for session cookies.
- **Stolen-cookie alerts:** Look for Defender XDR Stolen session cookie was used, Defender for Cloud Apps Suspicious inbox manipulation rule, and authenticator apps registered after the sign-in.

Sources

- Sophos: Sophos Active Adversary Report 2026: Identity Attacks Dominate as Threat Groups Proliferate
- FBI IC3: Kali365 Phishing-as-a-Service Kit Hijacks Microsoft 365 Access Tokens
- Microsoft: Inside Tycoon2FA: How a leading AiTM phishing kit operated at scale
- Sophos: Nowhere, man: The 2026 Sophos Active Adversary Report
- Proofpoint: Device Code Phishing is an Evolution in Identity Takeover
- FBI IC3: Malicious Cyber Actors Gain Access to Victim Accounts Through Consent Phishing
- Microsoft: Breaking the code: Multi-stage 'code of conduct' phishing campaign leads to AiTM token compromise
- Microsoft: Impersonating IT support: how threat actors turn a remote session into enterprise-wide access
- Okta: Vishing actors target Entra passkey enrollment
- Google Threat Intelligence: Tracking the Expansion of ShinyHunters-Branded SaaS Data Theft
- Mimecast: OAuth Device Code Phishing Campaigns Surge with EvilTokens Toolkit
- BleepingComputer: BigBear Microsoft 365 phishing service bypassed MFA at 258 organizations
- ITPro: Everything we know about the Dropbox breach so far

RANSOMWARE WITHOUT A KINGPIN

Overview & Impact

The 2026 ransomware market grew at the bottom while concentrating at the top. Black Kite counted 146 active groups by June and put the five largest at 43.6% of all victims. In August, BreachSense had Qilin alone at 157 victims, its best month of the year; ThreatVectr had Qilin at 165, 15% of every listing it saw. Both trackers count leak-site claims rather than confirmed breaches. Manufacturing led ThreatVectr's August count at 169 victims; healthcare topped BreachSense's at 91.

How the crews get in hasn't changed much, but the pace has. Microsoft's Storm-1175, which deploys Medusa, has exploited more than 16 vulnerabilities in web-facing products since 2023, weaponized a SAP NetWeaver flaw (CVE-2025-31324) the day after disclosure, and hit SmarterMail and GoAnywhere MFT a week before either flaw was public. The gap between first access and ransomware can be a single day. Clop is reported to have exploited CVE-2026-12569 in PTC Windchill and FlexPLM to plant JSP web shells and steal product data, then emailed extortion notes from compromised accounts to hundreds of employees per victim. A cPanel authentication bypass (CVE-2026-41940) that Shadowserver says compromised at least 44,000 IP addresses is pushing a Go-based Linux encryptor called Sorry. DeadLock, a Rust encryptor deployed by an affiliate of the Lynx and INC ecosystems, keeps its victim chat and leak blog on the Polygon blockchain, so that store needs no domain to seize.

- **Refusing to pay doesn't end it.** Berlin's whole 5.8-terabyte archive went public, and the Chaos Computer Club says it holds water-supply data, personal data on administrative staff, and emergency plans.
- **The Manchester Airports Group records are now free to any criminal.** Huntress expects the vehicle registrations, postcodes and contact details to feed convincing phishing, and Comparitech warns they show when people are away from home.
- **What Clop says it took from Shell, GE, and Philips is intellectual property, not customer records:** backups, project plans, facility photos, drawings, and blueprints, 89GB from Shell alone.
- **Paying rarely closes the case.** Proofpoint found 54% of affected organizations paid and 37% of those got a second demand; 65% had data stolen.
- **Small companies fare worse.** Only 34% of organizations with 100 to 250 employees stopped an attack before encryption, against 46% of larger ones, in Sophos's survey.

Guidance

Strategic Intelligence

Trend

- **Market, not kingpin:** Every prior year had a main character, LockBit one year and RansomHub the next. 2026 has a crowded middle instead, with 61 new groups in a single year. A new name can reach 40 victims in its first month.
- **One flaw, many victims: Clop's playbook turns a single enterprise bug into a batch of victims:** MOVEit hit more than 2,770 organizations, Windchill and FlexPLM produced 43 in one listing. Storm-1175 works the same clock against flaws that already have patches, and Black Kite says Everest built its pattern on accumulated patch debt.

Operational Intelligence

Threat Vectors

- **Email and phishing:** Malicious email (26%) and phishing (24%) were half the incidents in Sophos's State of Ransomware survey of 2,158 victims, the first time in four years that exploited vulnerabilities lost the top spot. Proofpoint had 47% of incidents starting with a malicious link, and 40% looked legitimate enough that staff didn't suspect them.
- **Exposed edge systems:** Where the way in was an exploit or a credential, Sophos traced 38% to exposed applications, 30% to user devices, 21% to firewalls and 8% to VPNs. Storm-1175's list reads Exchange, PaperCut, Ivanti Connect Secure, ScreenConnect, TeamCity, SimpleHelp, CrushFTP, GoAnywhere MFT, SmarterMail, BeyondTrust; Clop added PTC Windchill and FlexPLM, and cPanel and WHM joined in May.

Monitoring & Detection Gaps

- **Logs that disappear:** DeadLock clears Application, Security, System, and Windows PowerShell logs, sets Enabled to 0 on every channel under WINEVT\Channels, and rewrites ChannelAccess so nothing new is recorded. If your only copy lives on the host, it's gone.

Response Actions

- **Isolate, image, rotate: ReliaQuest's Windchill and FlexPLM guidance applies broadly:** take the affected servers off the network, preserve their disks and logs before you rebuild, and rotate every credential they held before you restore service.
- **Assume credentials are gone:** Storm-1175 dumps LSASS, pulls NTDS.dit and the SAM hive from a domain controller, and runs a script to recover Veeam passwords. Plan on rotating domain and backup credentials, and reset every account that installed an RMM agent you didn't approve.

Tactical Intelligence

Mitigation Strategies

- **Tamper protection on:** Turn on tenant-wide tamper protection and set DisableLocalAdminMerge so a local admin can't add antivirus exclusions. Storm-1175 edits Defender's registry settings and excludes the whole C:\ drive with encoded PowerShell.
- **Block lsass and webshells:** Block credential stealing from lsass.exe, block process creations from PsExec and WMI, block webshell creation for servers, and block executables that don't meet a prevalence, age, or trusted-list bar. Microsoft saw the PsExec/WMI rule and the prevalence rule stop DeadLock before hands-on-keyboard activity. Add a Controlled Folder Access policy, tuned in audit mode, and automatic attack disruption in Defender XDR.
- **Unreachable backups:** Storm-1175 recovers Veeam passwords once it has admin, and DeadLock disables vss, swprv, and wbengine before it encrypts, then empties every recycle bin. Shadow copies and a domain-joined backup server won't survive; keep a copy those credentials can't touch.

Tactical Guidance

Preventive Measures

- **Patch by exploitability:** Storm-1175 exploited SAP NetWeaver the day after disclosure, and CISA gave federal agencies three days on CVE-2026-12569. Black Kite's rule is exploitability, not convenience. Windchill, FlexPLM, cPanel, and WHM should already be done.
- **Ship logs off-host:** DeadLock wipes and disables every local event log channel, so on-host logs are gone by the time you look. Forward Security, System, and PowerShell logs to a SIEM or log collector the ransomware account can't reach.

Threat Hunting Hypotheses

New admin and open RDP on a web-facing server

Hypothesis: An internet-facing application server (SmarterMail, Exchange, GoAnywhere MFT, ScreenConnect, TeamCity, SimpleHelp, or PTC Windchill) was exploited, and within days the attacker created a local account, added it to Administrators, and changed Windows Firewall policy to allow Remote Desktop.

Investigation Steps

- **Pull account creations:** Query Defender for Endpoint DeviceEvents and the Security log on those servers for net user /add and net localgroup administrators, and review alerts for User account created under suspicious circumstances and New local admin added using Net commands.
- **Check the firewall change:** Look for firewall policy changes that enable Remote Desktop on servers where RDP is normally blocked; Storm-1175 wrote the command output to a TXT file, so look for a new .txt written on the server around the same time. Diff the web root for new scripts at the same time, JSP files on Windchill and FlexPLM in Clop's case.
- **Trace the tunnel:** Flag conhost.exe running from any path other than System32 or making outbound connections, lsp.exe matching SHA-256 2f7564526016b78972b2cd49a610354d672523c (Rclone), and any traffic to 185.135.86[.]149, 134.195.91[.]224, or 85.155.186[.]121 (SimpleHelp C2).

Sources

- Sophos: The State of Ransomware 2026: Payments Drop as Encryption Climbs
- Black Kite: 2026 Ransomware Report: 7,551 Victims, Up 24.9%
- Verizon: 2026 Data Breach Investigations Report (DBIR)
- Proofpoint: 2026 AI-Era Ransomware Report
- Microsoft: Storm-1175 focuses gaze on vulnerable web-facing assets in high-tempo Medusa ransomware operations
- TechRadar: Ransomware hackers dump 1.4 million stolen records from German government
- Microsoft: DeadLock ransomware: Breaking down a Rust-based encryptor with decentralized recovery infrastructure
- BleepingComputer: Clop ransomware targets Windchill, FlexPLM in data theft attacks
- BleepingComputer: Critical cPanel flaw mass-exploited in "Sorry" ransomware attacks
- TechRadar: Manchester Airports hack: Experts weigh in on 8.7M data leak
- BleepingComputer: Philips and GE investigating Clop ransomware data theft claims
- Help Net Security: Ransomware in 2026: More groups, more victims, no slowdown
- BreachSense: August 2026 Ransomware Report: 964 Victims, 83 Groups
- ThreatVectr: State of Ransomware: August 2026

ATTACKS THAT MOVE IN MINUTES

Overview & Impact

Attackers have moved from asking a chatbot for help to handing it the work. CrowdStrike's 2025 data has FANCY BEAR running LLM-enabled malware called LAMEHUG and FAMOUS CHOLLIMA scaling insider operations with AI-generated personas. Google's Q2 2026 tracker goes further. The financially motivated group UNC6780 published trojanized MCP servers to PyPI and shipped a stealer called DUSTMAKER, which steals GitHub Actions tokens to publish packages carrying valid signed attestations that pass AI coding agents' trust checks. Google also found an exposed server running an agent framework called Recon, its dashboard managing over 23,800 harvested secrets. Anthropic's GTG-1002, the espionage actor it disrupted in November 2025, ran Claude Code on a Kali machine with penetration testing tools wired in as MCP servers. The AI exploited an SSRF flaw, harvested SSH keys and AWS Secrets Manager tokens, and moved laterally. The human set direction and kept the final download decision.

Attackers also went after the AI itself. Google saw a PRC-nexus group, UNC6508, target proprietary AI research and, in activity it suspects is the same group, run a local open-weight model inside compromised cloud to avoid API monitoring. Underground prices for stolen AI accounts, mostly Claude and Gemini, more than doubled in 2026. Microsoft watched a technique born in AI red-teaming, ASCII smuggling with invisible Unicode tag characters, cross over into a phishing campaign that fired more than 1.3 million messages in a single day.

- The response window is now minutes. SNARKY SPIDER went from account takeover to data theft in under five minutes, and in another intrusion CrowdStrike saw data leaving four minutes after initial access.
- A breached cloud account is now also a resource to spend. CrowdStrike saw cloud-focused eCrime rise 171%, with cryptomining and LLM abuse running alongside credential theft.
- Companies building AI can lose the AI itself. In Q2 Mandiant worked several extortion cases in which attackers stole models, prompts, skills, source code, and related research.

Guidance

Strategic Intelligence

Trend

- **Minutes, not days:** 88% of proof-of-concept exploitation lands inside 48 hours, and CrowdStrike's Adam Meyers says AI is compressing the time between deciding to attack and doing it. Patch and response plans measured in days are now too slow.
- **Orchestration beats skill:** Anthropic's median actor used 16 techniques, a breadth that five years ago might have signaled a well-resourced team. What separates high-risk actors now is the scaffolding around the model; technical sophistication correlated with risk at only r = 0.28.

Operational Intelligence

Threat Vectors

- **Poisoned coding tools:** Attackers publish trojanized MCP servers such as tiktoken_mcp on PyPI and inject code into projects like azure-functions-mcp-extension on GitHub. DUSTMAKER drops files into .claude/, .vscode/ and .cursor/ that run when the IDE opens the workspace, and MIDNIGHT NEPTUNE alters Claude CLI hooks to drop its SOMBERMEME backdoor.
- **Agents on your cloud:** Attackers run markdown instruction sets as playbooks, scan autonomously, rotate IPs and troubleshoot in real time, all from compromised cloud infrastructure, so the attack traffic comes from legitimate addresses.

Monitoring & Detection Gaps

- **Developer noise:** DUSTMAKER lives in IDE workspace folders instead of registry keys or /etc/cron.*, where EDR watches harder, and it deletes its GitHub Actions workflow logs after running.
- **What AI returns:** Netskope found AI assistants returning data the user wasn't cleared for. The rarest of those returns is malicious code handed back to a coding agent, and it's the most severe, because the agent may run it.

Response Actions

- **Rotate developer secrets:** Assume any GitHub PAT, GitHub Actions OIDC token, or key in a Cline secrets.json or Continue config.yaml is exposed. Rotate them, then check Artifact Registry, Cloud Run IAM bindings to allUsers, rogue Editor service accounts and pending quota increases.
- **Sweep the workspace folders:** Diff .claude/, .vscode/ and .cursor/ across repos for hooks, startup commands or a setup.mjs nobody added, and review Claude CLI hooks for changes.
- **Check what Defender caught:** Confirm Prompt Injection Protection is on — it needs Defender for Office 365 Plan 2 or E5 — then filter EmailEvents on DetectionMethods to see what it has already caught on inbound mail.

Tactical Intelligence

Mitigation Strategies

- **Normalize before you match:** Strip or fold Unicode tag characters and other zero-width code points from subject and body before content signatures run, and apply the same normalization upstream of any AI assistant that reads mail.
- **Gateway every AI call:** Route all AI traffic, including shadow AI, local Ollama instances and remote MCP, through an AI gateway or proxy that logs every prompt and scans every response for malicious code, with threat scanning in CI/CD behind it.
- **Layer the signals:** Keep sender, IP, URL and domain reputation, machine-learning spam classification, brand-impersonation detection and authentication checks running ahead of content signatures. Over 99% of the smuggling campaign was caught by layers that never saw the tag characters.

Preventive Measures

- **Verify beyond the signature:** DUSTMAKER's packages carried valid SLSA Build 3 attestations and still passed agent trust checks, so a signature is not a decision. Require a named release process behind every OIDC-authenticated publish to PyPI or npm, and check an agent's workspace is isolated from production before it ships.
- **Fund the front door:** Proofpoint's top two bypass reasons were employees trusting authentic-looking messages and users interacting with malicious content, not technical failure. Fund email filtering, phishing-resistant sign-in and a verified way for staff to check an internal message before you fund more endpoint recovery.

Threat Hunting Hypotheses

Backdoored AI coding workspace

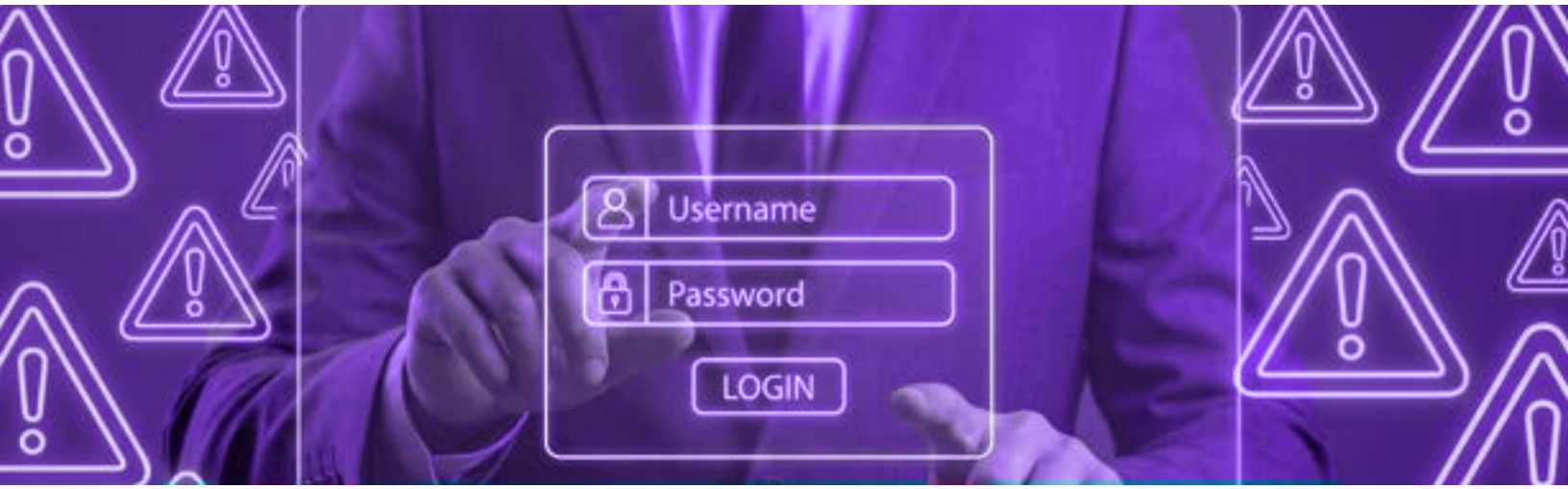
Hypothesis: A malicious package or cloned repo has dropped files into AI assistant and IDE workspace directories that execute commands through the coding agent or on workspace open, and is using CI/CD tokens to republish poisoned packages.

Investigation Steps

- **Inventory hidden dirs:** Across developer endpoints and repos, list the contents of .claude/, .vscode/ and .cursor/ and flag config files that instruct the assistant to run scripts such as setup.mjs, or hooks no known developer committed.
- **Check Actions history:** In GitHub, look for workflow tasks named like Copilot Setup, runs whose logs were deleted through the API, and OIDC-authenticated publishes to PyPI or npm that no release process triggered.
- **Read the loader comments:** For suspect JavaScript loaders like _index.js, inspect the top-of-file comments; extreme weapons-themed text is there to make an LLM scanner refuse, and is itself the indicator.

Sources

- CrowdStrike: 2026 CrowdStrike Global Threat Report: AI Accelerated Adversaries
- Google Threat Intelligence: GTIG AI Threat Tracker: From Prompting to Autonomy – The Evolution of Adversarial AI
- CrowdStrike: CrowdStrike 2026 Threat Hunting Report: AI is Now Embedded Across Modern Adversary Operations
- Anthropic: Mapping AI-enabled cyber threats
- Netskope: Netskope AI Report: 2026
- Proofpoint: Proofpoint Research Finds 65% of Organizations Affected by Ransomware Say AI Made Attacks More Effective
- Microsoft: ASCII smuggling crosses over from AI prompt injection to phishing evasion



pellera

Contact the Pellera Threat Intel Group at getsecure@pellera.com

pellera.com

