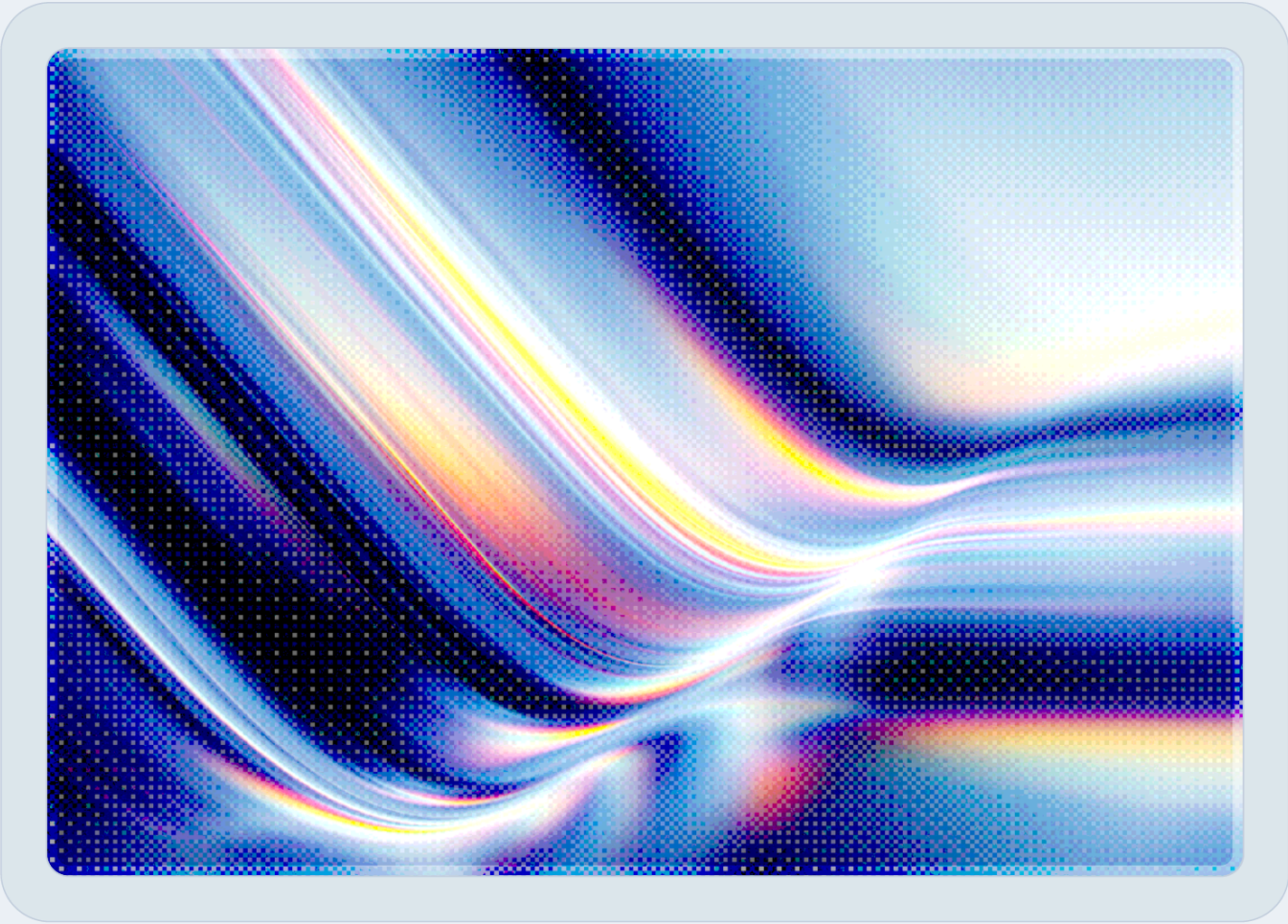
CSAI Foundation | Cloud Security Alliance

Jade Sleet Triggers macOS Backdoors Through Cursor AI

2026-09-21

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- The North Korean state-sponsored actor tracked as Jade Sleet, PUKCHONG, Slow Pisces, and TraderTraitor (also designated UNC4899) compromised a DevOps engineer at an India-based IT services provider using a fake job-interview lure containing a weaponized Terraform project; the lure's `.terraform.lock.hcl` file pointed to typosquatted HashiCorp provider registries, so that running the routine command `terraform init` silently downloaded and executed attacker-controlled code [1][2][3].
- Two Rust-based macOS backdoors, FLATROOF and ROOFDECK, remained dormant on the victim's disk for eleven days and activated only when the engineer opened a project workspace inside the Cursor AI coding assistant, which launched both implants as child processes. FLATROOF harvests browser credentials, keychain data, and terminal history and exfiltrates it over Telegram, while ROOFDECK is a more capable second-stage implant that resolves its command-and-control infrastructure through Nostr relay "dead drops" and authenticates operator commands with signed keys [1][3].
- The same backdoor family was previously linked to the April 2026 compromise of a LayerZero Labs developer, which cascaded into the theft of approximately $292 million in rsETH from the KelpDAO bridge [4][5].
- This campaign suggests that an AI coding assistant's implicit trust in a developer's workspace, and its willingness to execute build tooling on the user's behalf, can function as a largely unmonitored launch point for otherwise-conventional supply chain malware — a pattern this incident illustrates but does not by itself establish as industry-wide.

Background

Jade Sleet is one of several overlapping designations — alongside PUKCHONG, Slow Pisces, and TraderTraitor — used by different vendors to describe a financially motivated North Korean threat cluster operating under the broader Lazarus umbrella, with SentinelOne and Microsoft tracking the group's most recent activity as UNC4899 [1][3]. The group has spent several years targeting cryptocurrency exchanges, Web3 infrastructure providers, and the vendors that service them, and its signature technique is social engineering built around fabricated job interviews. Victims are approached through professional networking platforms and asked to complete a "coding assessment" or "candidate exercise"

hosted in a GitHub repository controlled by the attacker. Because the request arrives packaged as a routine technical interview task, it can bypass much of the skepticism a developer would typically apply to an unsolicited email attachment or link.

In the campaign detailed by SentinelOne and reported by The Hacker News and GBHackers, that lure took the form of Terraform infrastructure-as-code projects with names such as `gtn-candidate-repo`, `terraform-candidate-repo`, and `Northwind-IAC` [1][2]. Terraform's dependency resolution model made the technique effective: a `.terraform.lock.hcl` file inside the repository referenced provider modules from domains that closely mimicked HashiCorp's official registry, including `registry.hashicorp-aws[.]com` and `registry.hashicorp-terraform[.]io`. A developer who cloned the repository and ran the ordinary bootstrapping command `terraform init` — a step most infrastructure engineers perform without hesitation on any new project — caused Terraform to fetch and execute the attacker's provider code rather than a legitimate one. SentinelOne noted that at least one targeted candidate recognized the typosquatted domain and removed the suspicious provider before initialization, evidence that the technique is not foolproof, though the group's continued use of it across multiple campaigns suggests it succeeds often enough to be worth the operational cost [3].

This same tradecraft connects directly to a major 2026 DeFi loss: the approximately $292 million KelpDAO bridge hack [4]. LayerZero Labs' own incident report states that a developer was socially engineered on March 6, 2026 into cloning a malicious repository [5], giving the attackers a foothold that was later used to harvest session credentials for LayerZero's relayer infrastructure. On April 18, 2026, attackers exploited a single-point-of-failure verifier configuration on the KelpDAO rsETH bridge to fabricate proof of a token burn and drain roughly 116,500 rsETH from the bridge [4]. LayerZero's own postmortem and independent attribution from Mandiant and CrowdStrike tied the intrusion with high confidence to TraderTraitor/UNC4899, the same actor later observed using FLATROOF and ROOFDECK against the India-based IT provider [3][4][5]. The Indian provider had no direct cryptocurrency exposure, indicating the group may be widening its net beyond crypto-native targets, though this is so far based on a single observed non-crypto victim [1][3].

Security Analysis

The forensic timeline SentinelOne reconstructed on the India-based provider's compromised MacBook demonstrates how the attack chain interacts with modern developer tooling. FLATROOF and ROOFDECK were both present on disk as early as March 18, 2026, but neither backdoor executed or attempted to beacon out during the following eleven days [1][3]. That dormancy ended on March 29,

when the engineer opened a workspace named `cloudshield` inside Cursor, the AI-assisted code editor built on a fork of Visual Studio Code. Within seconds of the workspace opening, Cursor-spawned shell processes launched both implants via `nohup`, disguising them as renderer or system-update processes to blend in with the editor's own background activity [1][3]. Beaconing subsequently tracked closely with active Cursor sessions and went quiet whenever the editor was closed, strongly suggesting that a startup script, task configuration, or extension associated with that specific project workspace — rather than a persistent daemon — was the activation trigger.

This detail matters because it suggests AI coding assistants can function as a new and largely unmonitored launch surface for otherwise conventional malware staging techniques, a pattern illustrated by this incident though not yet established as industry-wide. The underlying persistence and evasion mechanisms in this campaign are not novel: FLATROOF strips the macOS `com.apple.quarantine` extended attribute from ROOFDECK to defeat Gatekeeper, and ROOFDECK installs itself as a Launch Agent via a plist in `~/Library/LaunchAgents`, using a `--type=renderer` command-line argument to resemble a legitimate Chromium-based helper process [1][3]. What is new is the reliance on an AI-assisted IDE's routine behavior — automatically running project tasks, executing shell commands the assistant suggests, or launching configured build steps when a workspace opens — as the dormant-to-active switch. A defender monitoring only for malicious downloads or unusual cron entries would miss this activation entirely, because the actual execution event looks, from the operating system's perspective, like ordinary developer activity inside a trusted application.

The two backdoors divide responsibilities along staging and second-stage access roles, consistent with tradecraft observed in other TraderTraitor/UNC4899 operations [3]. FLATROOF, also tracked as macOS.Gaslight, functions as the initial staging and reconnaissance tool: it can execute shell commands, upload and download files, and exfiltrate browser credential stores from Chrome, Brave, Firefox, and Safari, along with terminal history and system profiling data, using an embedded Telegram bot as its command channel [1][3]. ROOFDECK is the more capable second-stage implant. It resolves its command-and-control infrastructure through Nostr, a decentralized social protocol whose relay network was never designed for this purpose but offers attackers a resilient, difficult-to-block dead-drop mechanism, and it verifies that incoming operator commands are cryptographically signed using RSA-2048 before executing them, limiting the value of network interception to defenders [1][3].

Capability	FLATROOF (macOS.Gaslight)	ROOFDECK
Role	Initial staging, reconnaissance, credential theft	Second-stage interactive access

Capability	FLATROOF (macOS.Gaslight)	ROOFDECK
C2 channel	Telegram bot API	Nostr relay dead drops
Command integrity	None specified	RSA-2048-signed commands, verified before execution
Key functions	Shell execution, file transfer, browser/keychain theft, terminal history collection	Interactive shell, file manipulation, clipboard access, lateral movement, persistence
Defense evasion	Removes `com.apple.quarantine` from ROOFDECK	Launch Agent persistence disguised as renderer process

By April 20, 2026, the attackers redeployed an updated ROOFDECK variant with symbols and debug information stripped and deleted the original binaries. The timing — the day after LayerZero publicly acknowledged the KelpDAO hack — suggests the group was actively managing its exposure as attribution reporting began to circulate [1]. SentinelOne's telemetry shows the last confirmed command-and-control communication from this intrusion on June 1, 2026, though the absence of further beaconing does not by itself confirm eradication, since ROOFDECK's dead-drop C2 design is specifically intended to make that determination difficult from network monitoring alone [3].

Recommendations

The Jade Sleet campaign points to a small number of concrete controls organizations can apply immediately, alongside longer-term shifts in how AI-assisted developer tooling is governed. The recommendations below are organized by time horizon, from actions security teams can take this week to strategic questions that belong in a broader AI governance program.

Immediate Actions

Security teams should treat any `.terraform.lock.hcl`, `package-lock.json`, or comparable dependency lockfile arriving from an unsolicited "coding assessment," recruiter-provided repository, or unfamiliar external contributor as untrusted until the referenced registries and provider sources have been manually verified against known-good domains. Endpoint telemetry should be searched for Launch Agent plists created outside standard software installation workflows, particularly

entries in `~/Library/LaunchAgents` that invoke binaries with `--type=renderer` or similar browser-process arguments not tied to an installed browser. Outbound network activity from developer workstations also warrants review for unexpected Telegram Bot API calls and connections to Nostr relay infrastructure, both of which are unusual in typical enterprise egress patterns.

Short-Term Mitigations

Organizations should require that infrastructure-as-code tooling such as Terraform, along with package managers, run inside isolated, ephemeral build environments rather than directly on engineer workstations that hold cloud credentials, SSH keys, or session tokens. AI coding assistants, including Cursor, should be configured to require explicit human approval before executing shell commands or launching background processes tied to opening a new project workspace, with any "auto-run" or unattended task execution disabled for repositories that have not been reviewed. Endpoint detection rules should also be extended to flag child processes spawned by AI coding assistant applications, since this campaign demonstrates that an editor's own process tree can become a malware launch point that traditional "suspicious parent process" heuristics may not cover.

Strategic Considerations

Security teams should incorporate AI-assisted developer tools explicitly into supply chain and endpoint threat models rather than treating them as extensions of a general-purpose code editor, since their capacity to autonomously execute commands changes the risk calculus for what a compromised or socially engineered developer workflow can trigger. The same scrutiny organizations already apply to phishing email in security awareness training should extend to job-interview and recruitment-themed repository lures, given the persistence of this technique across multiple unrelated DPRK campaigns over the past two years. Organizations should also track DPRK-attributed operations against AI developer environments as a distinct and recurring pattern rather than isolated incidents: overlapping North Korean clusters — including Jade Sleet/TraderTraitor, responsible for the interview-lure compromise described in this campaign, and Sapphire Sleet, responsible for npm supply chain poisoning of AI framework packages — are independently converging on the same class of high-value target, namely AI-assisted developer workstations and the trust gaps in the tools that run on them.

CSA Resource Alignment

This campaign sits at the intersection of two threat patterns CSA has already examined in its AI-assisted rapid research program, and both should inform how organizations interpret this incident.

CSA's "Sapphire Sleet Poisons Mastra AI npm Supply Chain" [6] documents a related June 2026 operation in which a different DPRK-attributed cluster, Sapphire Sleet, compromised roughly 145 npm packages in the Mastra AI agent framework through a typosquatted transitive dependency. That report's core finding — that AI developer workstations concentrate unusually high-value credentials (LLM API keys, cloud tokens, CI/CD secrets) and have consequently become a priority target for North Korean operators — is directly reinforced by the Jade Sleet campaign described here, which targeted the same class of engineer using a different delivery mechanism. Organizations that implemented the lockfile discipline, `--ignore-scripts` enforcement, and credential-scoping recommendations from that note should extend the same controls to infrastructure-as-code tooling like Terraform, not just npm.

CSA's "GhostApproval: A Shared Symlink Trust-Boundary Flaw in AI Coding Assistants" [7] is the more structurally relevant companion piece, and it names Cursor specifically as one of six affected tools. Where GhostApproval examined how a malicious repository can manipulate an AI coding assistant's file-write behavior through symlink trickery that defeats human-approval dialogs, the Jade Sleet campaign shows the adjacent risk: an AI assistant's willingness to execute workspace-configured commands and background tasks without an equivalent approval gate. Both cases share a root cause CSA's research identified as an industry-wide "identity and access maturity gap" in agentic developer tools — these assistants commonly inherit the full file-system and execution permissions of the developer account they run under, and CSA's research to date has not identified a widely adopted agent-scoped identity boundary that would contain the blast radius of a single compromised repository.

Both incidents map to the AI Controls Matrix (AICM v1.1) [8], particularly its supply chain security and identity and access management domains, and to the MAESTRO agentic AI threat modeling framework's development-infrastructure and agent-ecosystem layers. Security teams evaluating AI coding assistant deployments should treat vendor responses to symlink resolution and unattended command execution — the specific gaps GhostApproval catalogued — as a leading indicator of how well a given tool will resist exactly the kind of workspace-triggered activation observed in this campaign.

References

[1] The Hacker News. "Jade Sleet Linked to Indian IT Provider Breach With FLATROOF and ROOFDECK Backdoors." The Hacker News, September 2026.

[2] GBHackers. "Hackers Weaponize Terraform Lock Files to Infect DevOps Engineers With macOS Backdoors." GBHackers, September 2026.

[3] SentinelOne. "Don't Call Us, We'll Call Your APIs | TraderTraitor Backdoors Resurface on Victim With No Crypto Ties." SentinelOne Labs, September 2026.

[4] Decrypt. "LayerZero Pins $292M KelpDAO Bridge Hack on North Korea's Lazarus Group." Decrypt, April 2026.

[5] LayerZero Labs. "LayerZero Labs KelpDAO Incident Report." LayerZero, May 2026.

[6] Cloud Security Alliance. "Sapphire Sleet Poisons Mastra AI npm Supply Chain." Cloud Security Alliance, June 2026.

[7] Cloud Security Alliance. "GhostApproval: A Shared Symlink Trust-Boundary Flaw in AI Coding Assistants." Cloud Security Alliance, July 2026.

[8] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2026.

