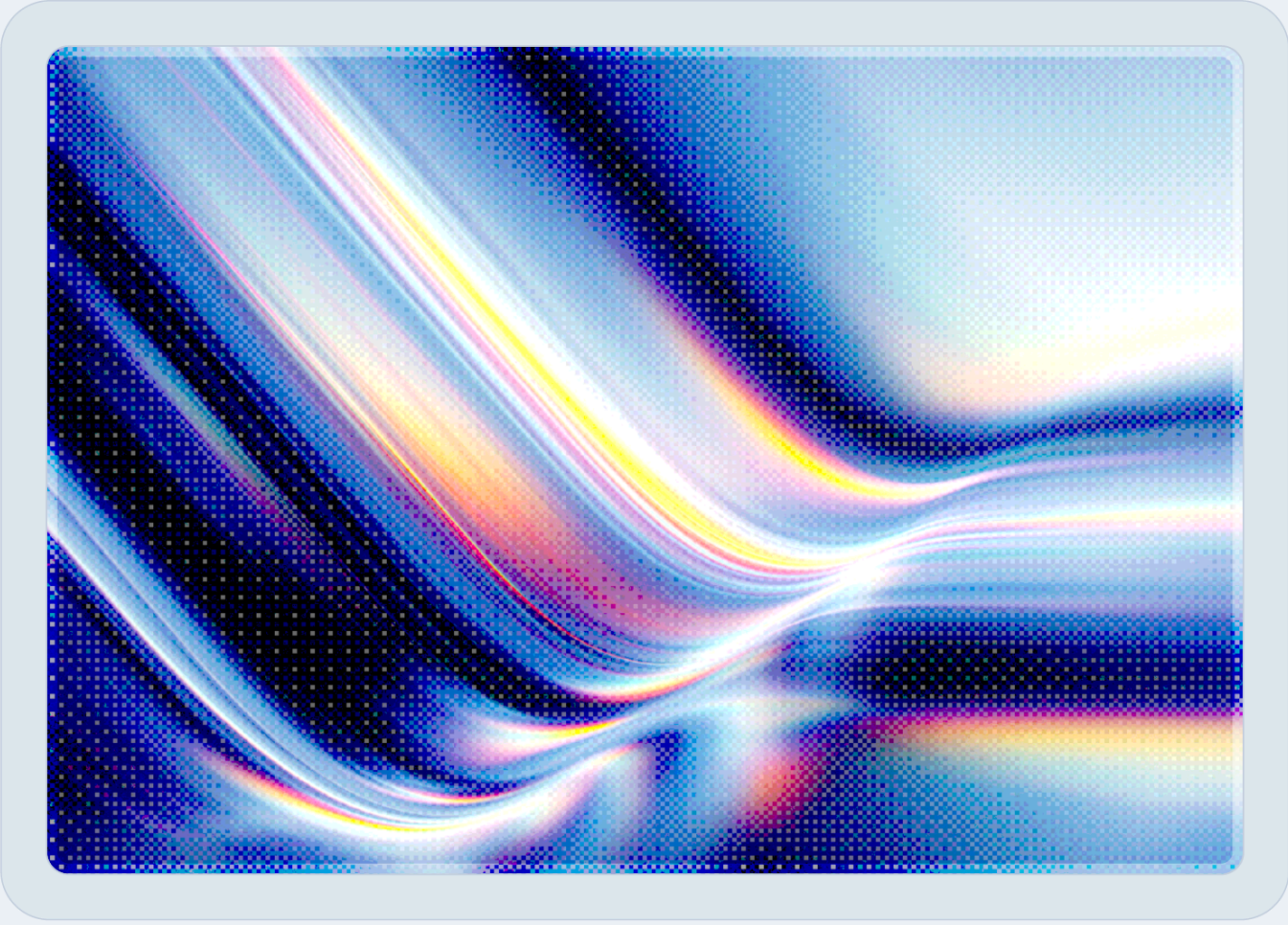
CSAI Foundation | Cloud Security Alliance

When Trusted AI Platforms Become Malware Delivery Infrastructure

2026-09-14

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

Over the summer of 2026, threat actors repeatedly turned the public-facing, user-generated-content features of major AI platforms into malware delivery infrastructure, abusing the inherent trust users place in domains such as claude.ai and chatgpt.com. In the FakeAgent campaign, a malicious Claude Artifact hosted on Anthropic's own domain impersonated the Claude Desktop download page and infected at least 29 organizations with the SectopRAT information stealer before it was taken down [1][2]. A related campaign, dubbed ClaudeFix, weaponized shared Claude.ai conversation links posing as Apple Support guidance to trick macOS users into pasting a malicious terminal command that deployed the MacSync stealer [3][4]. A third and earlier pattern, first documented in December 2025, showed attackers seeding shared ChatGPT and Grok conversations with fake troubleshooting advice that led victims to install the Atomic macOS Stealer (AMOS) [5][6][7]. Across all three cases, the underlying exploit is the same: the malicious content lives on a domain the victim, their browser security controls, and their organization's web filtering all treat as inherently safe, because it genuinely is the AI vendor's own domain. Security teams should treat AI platform artifacts, shared conversation links, and chatbot-generated instructions as untrusted, externally influenced content, and should update endpoint, web-filtering, and awareness-training controls accordingly, since domain reputation alone can no longer be relied on to distinguish safe AI platform content from attacker-controlled payloads.

Background

The rapid mainstream adoption of generative AI assistants has given Anthropic's Claude, OpenAI's ChatGPT, and xAI's Grok enormous installed bases and correspondingly high domain trust. Each of these platforms has also built features that let users generate and publicly share content: Claude's "Artifacts" allow a user to build and host interactive mini-applications directly on claude.ai, while both Claude and ChatGPT support shareable, publicly viewable conversation links that render on the vendor's own domain. The incidents examined in this note suggest these features were not built with the kind of malware-scanning or content-moderation pipeline that CDNs and app stores apply to third-party payloads, even though they let any user publish content that renders on the vendor's own trusted domain. That gap became exploitable the moment attackers realized a link ending in claude.ai or chatgpt.com would sail past reputation-based filtering, security-conscious users' instincts, and even some automated URL-scanning tools that treat major AI vendor domains as inherently benign.

Security researchers at Huntress, corroborated by reporting from BleepingComputer, Zscaler, and Malwarebytes, documented three distinct but related campaigns between December 2025 and August 2026 [1][2][3][4][5]. The first, active between July 21 and July 22, 2026, involved a sponsored Bing advertisement for "Claude Desktop app" that, unusually for malvertising, pointed to the legitimate claude.ai domain rather than a lookalike. Clicking through led not to Anthropic's real download page but to a public Claude Artifact built to imitate it; clicking "Download" on the spoofed artifact then redirected victims through a chain of attacker-controlled domains before serving a trojanized ClaudeDesktop.exe. Huntress named this campaign FakeAgent and reported it compromised at least 29 organizations in roughly 48 hours before Anthropic removed the offending artifact [1][2].

The second campaign, which Zscaler tracked and named ClaudeFix, ran between June 12 and June 19, 2026, and used paid Google ads to route macOS users searching for "claude download" or "claude mac" to a shared Claude.ai conversation link [3]. The attackers had set their Claude account's display name to "Apple Support," so the resulting shared chat appeared to victims as an authoritative Apple troubleshooting guide hosted on Anthropic's own domain [3]. This impersonation-and-shared-chat technique was not new in June: Malwarebytes had already documented fraudulent search results directing macOS users to Claude.ai shared chats a month earlier, and CSA's own research had separately tracked an earlier wave of the same "Apple Support" impersonation delivering the MacSync stealer through Claude.ai [4][12]. The third and earliest-documented pattern, reported by Malwarebytes in December 2025 and later summarized by Huntress and Dark Reading, involved attackers using prompt engineering to induce ChatGPT and Grok to generate polished "fix" or "cleanup" instructions for macOS, then publishing the resulting shared conversation links and promoting them through SEO and paid placement so they ranked highly for common troubleshooting searches such as "clear disk space on macOS" [5][6][7].

Security Analysis

All three campaigns rely on what researchers increasingly describe as trust transfer: an AI platform's institutional reputation, established through legitimate use, is borrowed by attacker-controlled content that happens to be hosted on the platform's own infrastructure. This is distinct from classic phishing or typosquatting, where a lookalike domain can at least in principle be flagged by domain-reputation systems; here the domain is not fake, only the content is. Security controls that rely on allowlisting major SaaS and AI vendor domains, a common practice for reducing alert fatigue in secure web gateways, therefore fail by design against this technique, since the traffic destination is legitimately claude.ai or chatgpt.com throughout the entire redirect chain until the final malware-hosting hop.

The delivery mechanics also converge on a small set of well-understood techniques repackaged for the AI context. FakeAgent's payload was not a novel malware family; SectopRAT is an established .NET-based remote access trojan with credential- and session-theft capabilities, delivered here via DLL sideloading, in which a legitimate, digitally signed JetBrains Chromium component was repackaged to load a malicious libcef.dll at runtime [1][2]. A comparable DLL-sideloading delivery mechanism appears in a separate AI-chatbot-driven cryptojacking campaign CSA has documented, in which a signed, trusted binary was similarly repurposed to load a malicious payload [8]. FakeAgent's command-and-control infrastructure used EtherHiding, a technique that embeds C2 configuration inside Ethereum blockchain transactions, letting operators rotate infrastructure by publishing a new transaction rather than standing up a new server that defenders can eventually enumerate and block [1][2].

The ClaudeFix and AMOS-distribution campaigns instead rely on ClickFix, a social-engineering technique first observed broadly in 2024 in which a victim is walked through copying a command into their clipboard and pasting it into a Terminal or Run dialog to "fix" a manufactured problem [3][4][5]. ClickFix succeeds because it asks the victim to perform the technically risky action themselves, sidestepping browser download warnings, Gatekeeper prompts, and email attachment scanning entirely; the victim, believing they are following legitimate troubleshooting guidance, executes the payload under their own credentials. Pairing ClickFix with content hosted on an AI vendor's domain, and in the ClaudeFix case with a spoofed "Apple Support" display name, compounds the technique's existing effectiveness with borrowed institutional trust, a combination CSA has flagged as a distinguishing feature of the broader shift toward AI-mediated malware distribution [8]. The resulting malware, MacSync and AMOS respectively, targets browser-stored credentials, keychain secrets, session tokens, and cryptocurrency wallets, giving attackers both immediate financial upside and durable access to victim accounts through stolen session material that can survive a password reset [3][4][5].

A further complication is content lifecycle asymmetry. AI vendors can and did remove the offending artifacts and shared links once notified, Anthropic took down the FakeAgent artifact and the ClaudeFix conversations after external researchers flagged them, but there was a meaningful window, in some cases days, during which the malicious content remained live and continued to convert victims [1][3]. Because neither Claude's Artifacts feature nor ChatGPT's and Grok's shared-conversation features were designed with adversarial publishing in mind, detection currently depends on third-party security researchers noticing abuse and reporting it, rather than on proactive content screening comparable to what app stores or CDNs perform on hosted binaries.

Recommendations

Immediate Actions

Security teams should treat links to AI platform artifacts and shared conversations, including those on claude.ai, chatgpt.com, and grok.com, as untrusted external content for the purposes of secure web gateway and email filtering policy, rather than extending blanket allowlist trust to these domains. Endpoint controls should restrict clipboard-driven execution in Terminal, PowerShell, and Run dialogs, since ClickFix depends entirely on a user completing that paste-and-run step voluntarily. Security awareness communications should specifically warn staff that a legitimate-looking AI chatbot conversation or artifact is not proof that its instructions are safe, and that troubleshooting guidance calling for a pasted terminal command should be independently verified against the vendor's own official support documentation before execution.

Short-Term Mitigations

Organizations should extend application allowlisting (Windows Defender Application Control, AppLocker, or macOS Gatekeeper policies) to account for DLL sideloading against signed binaries, since FakeAgent demonstrated that allowlisting a trusted vendor's executable is not sufficient if that executable can load an attacker-supplied library at runtime. Endpoint detection should be tuned to flag newly created scheduled tasks, antivirus exclusion changes made outside change management, and unexpected child processes spawned from Terminal or PowerShell shortly after a browser session, all of which appeared as post-compromise indicators across these campaigns. Security teams should also establish a direct reporting channel to their organization's AI platform vendors for suspected malicious artifacts or shared content, since the FakeAgent and ClaudeFix campaigns were only remediated after external researchers escalated directly to Anthropic.

Strategic Considerations

Longer term, enterprises should incorporate AI platform sharing features, artifacts, shared conversations, and similar user-generated-content surfaces, into their third-party and SaaS risk assessments, recognizing that a vendor's core product security posture does not automatically extend to features that let any user publish content on that vendor's domain. CSA encourages AI platform providers to adopt content-moderation and malware-scanning pipelines for publicly shareable artifacts and conversations comparable to those used by app stores and CDNs, particularly for artifacts that request downloads or display executable instructions. Security architects designing Zero Trust programs

should extend verification requirements to AI-mediated content itself, treating an AI platform's domain reputation as necessary but not sufficient evidence of content safety, and building detection logic that inspects the actual payload or instruction a link resolves to rather than the hosting domain alone.

CSA Resource Alignment

This research note extends CSA's existing threat intelligence line on AI platforms being repurposed as malware distribution channels. CSA's own "AI Chat Trust Weaponized in Mac Malvertising Campaign" documented an earlier wave of the same technique described here, threat actors abusing Claude.ai's shared-chat feature and an "Apple Support" display name to deliver the MacSync stealer to macOS users via Google Ads, and framed that pattern explicitly as an evolution of the December 2025 ChatGPT/Grok-AMOS campaign; the ClaudeFix campaign Zscaler tracked and named in June 2026 appears to be a continuation of the same technique [12]. CSA's "Poisoned AI Recommendations: Chatbots as Malware Delivery Vectors" documented the closely related pattern of SEO poisoning migrating into AI chatbot recommendation channels via a Microsoft-disclosed cryptojacking campaign, and its core finding, that AI recommendation surfaces carry disproportionate institutional trust and should be treated as untrusted, externally influenced input, applies directly to the FakeAgent, ClaudeFix, and AMOS-distribution campaigns described here [8].

CSA's "Phantom Squatting: AI Hallucinated Domains as Phishing Infrastructure" analyzed a complementary abuse pattern in which adversaries register domains that AI models hallucinate, rather than compromising the AI vendor's own domain outright [10]; together the two reports show attackers pursuing both ends of the trust chain, spoofing what an AI system recommends and hijacking the platform it runs on. CSA's "Using Zero Trust to Counter Identity Spoofing & Abuse" provides the applicable control framework for the identity-spoofing element of ClaudeFix, in which attackers set a Claude account's display name to "Apple Support" to borrow a second layer of institutional credibility, and its Zero Trust-aligned detection and mitigation guidance for identity abuse maps directly onto verifying the provenance of AI-hosted conversational content [11]. Finally, organizations formalizing controls in response to this threat category should reference the AI Controls Matrix (AICM) v1.1, whose threat and vulnerability management and universal endpoint management domains cover the application allowlisting, DLL sideloading resistance, and endpoint monitoring controls recommended above [9].

References

[1] Huntress. "Inside FakeAgent: How a Claude Desktop Malvertising Campaign Hit 29 Organizations with SectopRAT." Huntress, July 2026.

[2] BleepingComputer. "Fake Claude app promoted by Bing ads pushes SectopRAT malware." BleepingComputer, July 2026.

[3] Zscaler. "ClaudeFix: Shared Claude Chats Meet ClickFix." Zscaler ThreatLabz, July 2026.

[4] Malwarebytes. "Fake Claude search results lure Mac users into ClickFix attack." Malwarebytes Labs, 2026.

[5] Malwarebytes. "Google ads funnel Mac users to poisoned AI chats that spread the AMOS infostealer." Malwarebytes Labs, December 2025.

[6] Huntress. "AI-Poisoning & AMOS Stealer: The Biggest Mac Threat." Huntress, 2026.

[7] Dark Reading. "ClickFix Style Attack Uses Grok, ChatGPT for Malware Delivery." Dark Reading, December 2025.

[8] Cloud Security Alliance. "Poisoned AI Recommendations: Chatbots as Malware Delivery Vectors." Cloud Security Alliance AI Safety Initiative, May 2026.

[9] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2025.

[10] Cloud Security Alliance. "Phantom Squatting: AI Hallucinated Domains as Phishing Infrastructure." Cloud Security Alliance AI Safety Initiative, July 2026.

[11] Cloud Security Alliance. "Using Zero Trust to Counter Identity Spoofing & Abuse." Cloud Security Alliance, 2026.

[12] Cloud Security Alliance. "AI Chat Trust Weaponized in Mac Malvertising Campaign." Cloud Security Alliance AI Safety Initiative, May 2026.

