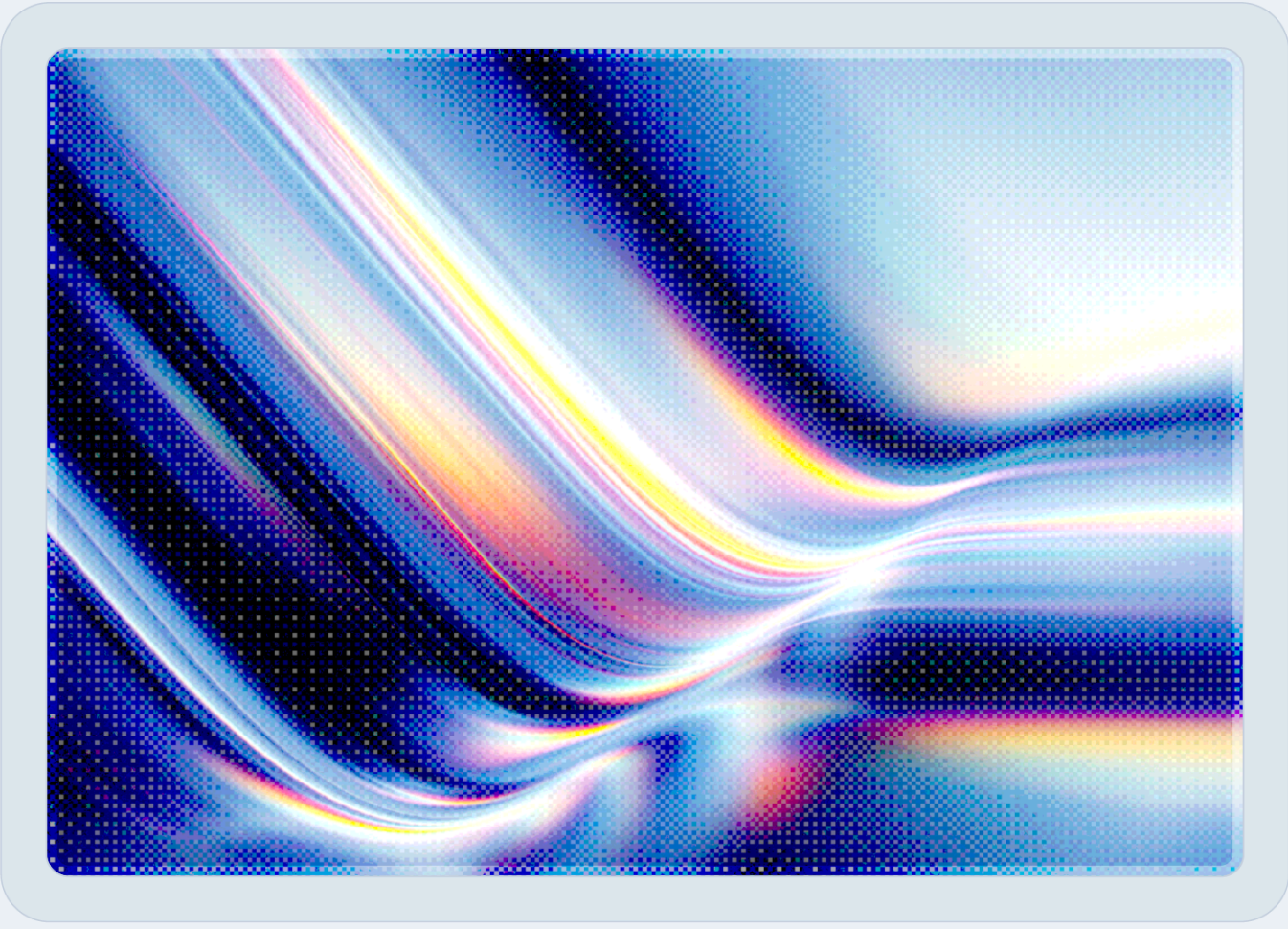
CSAI Foundation | Cloud Security Alliance

AI Adoption Is Flooding the SOC With Noise

Finding the 0.02% of Alerts That Are Real Attacks

2026-09-14

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

A large-scale analysis of enterprise Security Operations Center (SOC) telemetry published in September 2026 found that AI-related alerts grew 685% between February and June 2026, yet only 0.02% of those alerts corresponded to a confirmed real attack [1]. The findings reframe the security risk of enterprise-wide AI adoption: the danger is not that organizations' own AI agents are being hijacked at scale, but that legacy detection logic cannot distinguish routine AI tool use from malicious activity, obscuring genuine signal within a rapidly growing volume of noise. Several points stand out for SOC leaders.

- AI-related alerts remain a small fraction of total SOC volume today (0.43%), but the growth rate means that fraction will not stay small, and teams that size their AI-alert handling to current volume will be under-provisioned within a quarter [1].
- Of all AI-related alerts examined, 94.1% were noise generated by legitimate developer and business use of tools such as Claude, ChatGPT, Cursor, and Codex CLI, while 5.8% represented genuine policy or data-handling risk and just 0.02% were confirmed attacks [1].
- None of the confirmed attacks in the dataset involved an organization's own AI agents malfunctioning or being compromised; instead, attackers exploited employees' trust in AI brand names as phishing lures [1].
- More than 80% of AI-related alerts were auto-suppressed without analyst review [1]. Given current volume, some form of automated suppression is likely unavoidable, though it carries the risk of discarding rare true positives alongside the noise.
- The underlying dynamic is consistent with longstanding SOC alert-fatigue research: independent industry surveys already show that roughly half of all security alerts are false positives and that false positives remain security teams' top operational complaint, even before AI-generated alert volume is added to the mix [2][3].

Background

Enterprise adoption of generative AI and AI coding agents appears to be outpacing many organizations' ability to adapt detection engineering, a dynamic the data in this note illustrates directly. Developers now routinely delegate multi-step tasks to coding agents such as Cursor and Codex CLI, non-technical staff

sign into consumer AI assistants with corporate credentials, and AI features are increasingly embedded directly inside business applications that SOCs already monitor. Each of these behaviors generates telemetry, and because much of it involves activity that legacy detection rules associate with compromise, such as automated multi-step actions, unfamiliar process trees, outbound connections to new API endpoints, and credential use from unfamiliar user agents, detection systems have started firing on AI adoption itself rather than on any underlying attack.

This pattern is an extension of a problem SOCs have struggled with for years. Independent research on alert fatigue consistently finds that a substantial share of daily SOC alert volume never gets meaningfully investigated. A Microsoft and Omdia study of enterprise SOC operations found that 46% of all alerts prove to be false positives and 42% are never investigated at all, driven largely by tool fragmentation and manual triage processes that cannot keep pace with volume [2]. The SANS Institute's 2025 Detection and Response Survey similarly found that 73% of organizations rank false positives as their top detection and response challenge, a figure SANS characterized as a dramatic rise over the prior year [3]. AI adoption did not create alert fatigue, but it is compounding an already-strained baseline with a new and fast-growing category of activity that legacy rules were never designed to interpret.

The September 2026 analysis referenced throughout this note quantifies that compounding effect directly. Drawing on approximately 16.9 million SOC alerts across enterprise environments, researchers identified roughly 73,000 AI-related alerts, or 0.43% of total volume, but found that this category grew 685% between February and June 2026 alone [1]. Because this figure is drawn from one dataset of aggregated enterprise telemetry rather than a cross-industry sample, the precise percentages may not generalize uniformly to every SOC environment, though the underlying dynamic, AI adoption generating disproportionate noise relative to genuine risk, is consistent with the broader alert-fatigue research summarized above. The researchers classified every AI-related alert into one of three buckets: noise generated by legitimate use, genuine risk requiring policy attention, and confirmed attacks. The resulting breakdown, and what it implies for SOC operations, is the focus of the analysis below.

Security Analysis

The composition of AI-related alerts is the central finding of this research, and it has direct implications for how security leaders prioritize AI-related risk in the SOC. Of the AI-related alerts analyzed, 94.1% were classified as noise: legitimate developer and business activity, such as a coding agent committing code, querying a package registry, or reading a repository, that happened to trip detection logic built for a pre-AI threat model. A further 5.8% represented genuine risk in the sense that matters to a data-protection or governance program rather than an intrusion-detection one, cases such as an employee pasting sensitive data into a consumer AI chat interface or an unsanctioned AI tool gaining broad OAuth

access to corporate SaaS data. Only 0.02% of AI-related alerts corresponded to a confirmed attack [1]. Put in absolute terms against a base of 73,000 AI-related alerts, that 0.02% represents a small number of incidents, on the order of 15, buried inside tens of thousands of alerts that individually look similar enough to warrant investigation.

The nature of that 0.02% is worth examining as closely as its size. The researchers found that none of the confirmed attacks resulted from an organization's own AI agents acting maliciously, being compromised, or malfunctioning. Instead, every confirmed attack in the dataset involved a threat actor exploiting the trust employees now place in AI brand names, using Anthropic, OpenAI, and Google Gemini branding as phishing lures to harvest credentials or induce risky actions [1]. This finding is consistent with a broader pattern in which attackers impersonate the routine business processes that AI vendors and their workflows have become part of, such as helpdesk resets, identity verification calls, and software update prompts, rather than attacking the underlying model or agent directly. Based on this dataset, the exploitable surface created by AI adoption in the confirmed-attack category looked social and procedural rather than technical, though a single study is not sufficient to establish that pattern as a general rule across all AI-related intrusions.

The operational consequence of a 94.1%/5.8%/0.02% split is that automated suppression becomes increasingly difficult to avoid, and that suppression carries its own risk. The analysis found that 81.7% of AI-related alerts were automatically suppressed without any analyst review, with only 5.4% escalated to a human [1]. Given the volume involved, blanket human review of every AI-related alert is not realistic for most SOC teams, and automated suppression of well-understood noise is a reasonable response. But suppression logic built primarily to reduce volume, rather than to actively distinguish the 0.02% from the 94.1%, risks discarding the rare true positive along with the noise it resembles. This is a familiar tension in autonomous security systems more broadly: systems tuned aggressively for speed and volume reduction can over-trigger or under-triage in ways that are invisible until an incident surfaces the gap.

A further complication, evident in the tool-level detail underlying this research, is that the tools generating the most noise are not uniform in behavior. Coding agents such as Cursor and Codex CLI perform actions, like autonomous multi-step code changes, dependency installation, and network calls, that closely resemble the kill-chain behaviors detection rules were written to catch. Consumer-facing assistants such as ChatGPT and Claude generate a different noise profile tied to data exposure and shadow IT rather than execution behavior. Tools such as DeepSeek introduce data-residency and vendor-trust questions layered on top of the detection problem, and infrastructure utilities such as ngrok, when used to expose local AI agent endpoints or expedite development workflows, generate alerts indistinguishable from a reverse-tunnel technique used in active intrusions [1]. Treating "AI-related alert" as a single category, rather than building detection logic specific to each tool's legitimate use pattern, is a significant part of why the noise share is so high, and it is the first place SOC teams can make measurable progress.

Finally, the growth trajectory matters more than the current volume. At 0.43% of total alert volume today, AI-related alerts are a manageable, if inconvenient, category. At 685% growth over four months, that share will not stay small, and the researchers explicitly frame today's figure as a floor rather than a ceiling [1]. A SOC that tunes its detection and staffing to current AI-alert volume, rather than to its trajectory, is planning to be under-resourced.

Recommendations

Immediate Actions

Security teams should begin by tuning the specific detections responsible for the largest share of AI-related noise, prioritizing rules that fire on routine developer-agent activity such as automated commits, dependency installs, and API calls from coding assistants, since these represent the highest-volume, lowest-risk category and the most immediate opportunity to reduce fatigue while preserving coverage of the indicators most associated with genuine risk, provided tuning is validated against the confirmed-attack indicators described below. In parallel, SOC and threat-hunting teams should add proactive hunts, rather than waiting on alerts, for the indicators most associated with the confirmed-attack category: permission-bypass attempts, unauthorized reverse tunnels such as ngrok, and unusually broad OAuth grants requested by newly connected applications, since these patterns correlate with the 0.02% of activity that matters most [1].

Short-Term Mitigations

Over the next one to two quarters, organizations should establish behavioral baselines for sanctioned AI tools so that detection logic can distinguish an agent's normal operating pattern from a deviation, rather than treating all agent activity as equally suspect. This should be paired with a written data-sharing policy for third-party AI platforms that specifies what categories of corporate data may be shared with which tools. A meaningful share of that "genuine risk" category likely reflects ambiguity about acceptable use rather than malicious intent, though the source data classifies these alerts by activity rather than by underlying intent. Where feasible, isolating AI tools in dedicated containers or virtual machines helps constrain what an agent can reach and makes it easier to attribute observed behavior specifically to the tool rather than to the underlying user account. Finally, teams relying heavily on automated suppression should build periodic audit sampling into the process, reviewing a statistically meaningful slice of suppressed AI-related alerts on a recurring basis to check that the suppression logic is not quietly discarding true positives.

Strategic Considerations

At a structural level, the volume and growth trajectory documented in this research support the case for redesigning security operations around AI-native principles rather than layering AI-specific rules onto a conventional SOC. Detection, triage, and response processes built for a pre-AI threat model do not scale gracefully to the volume and speed that AI-driven alert growth now demands, and organizations planning multi-year SOC modernization should treat that growth as one of the concrete drivers justifying investment. That case is reinforced by the broader cybersecurity workforce gap, which ISC2's most recent workforce study estimates at approximately 4.8 million unfilled positions globally, a shortfall that leaves most SOC teams with little slack to absorb a fast-growing new alert category through headcount alone [5].

Because analyst capacity is the binding constraint under rising AI-alert volume, organizations should also weigh controlled evidence on AI-assisted investigation tools when deciding where to invest first. CSA's benchmark study of AI agents in the SOC found that AI-assisted analysts completed escalated-alert investigations 45–61% faster than analysts working manually, with better resistance to the accuracy and completeness decline that alert fatigue otherwise produces over sequential investigations [4]. That evidence suggests the highest-value near-term investment is not necessarily new detection content, but tooling that helps a fixed analyst population absorb rising volume without a proportional decline in investigation quality.

CSA Resource Alignment

This research note connects most directly to two published CSA resources. CSA's Benchmark Study of AI Agents in the SOC supplies controlled evidence that AI-assisted investigation measurably reduces the accuracy and completeness decline that alert fatigue causes, directly informing the recommendation above to prioritize investigation tooling as analyst capacity becomes the binding constraint [4]. Organizations without a directly on-point maturity framework for their own AI-tool governance should also consult CSA's AI Controls Matrix (AICM), whose identity and access, model security, and operational governance domains provide control-level detail for the AI tool inventory and least-privilege practices recommended in this note.

References

[1] The Hacker News. "When the Whole Company Adopts AI: What It Does to Your SOC." The Hacker News, September 2026.

[2] Microsoft Security. "Unify Now or Pay Later: New Research Exposes the Operational Cost of a Fragmented SOC." Microsoft Security Blog, February 17, 2026.

[3] SANS Institute. "2025 SANS Detection and Response Survey." SANS Institute, 2025.

[4] Cloud Security Alliance. "A Benchmark Study of AI Agents in the SOC." Cloud Security Alliance, 2025.

[5] ISC2. "2024 ISC2 Cybersecurity Workforce Study." ISC2, October 2024.

