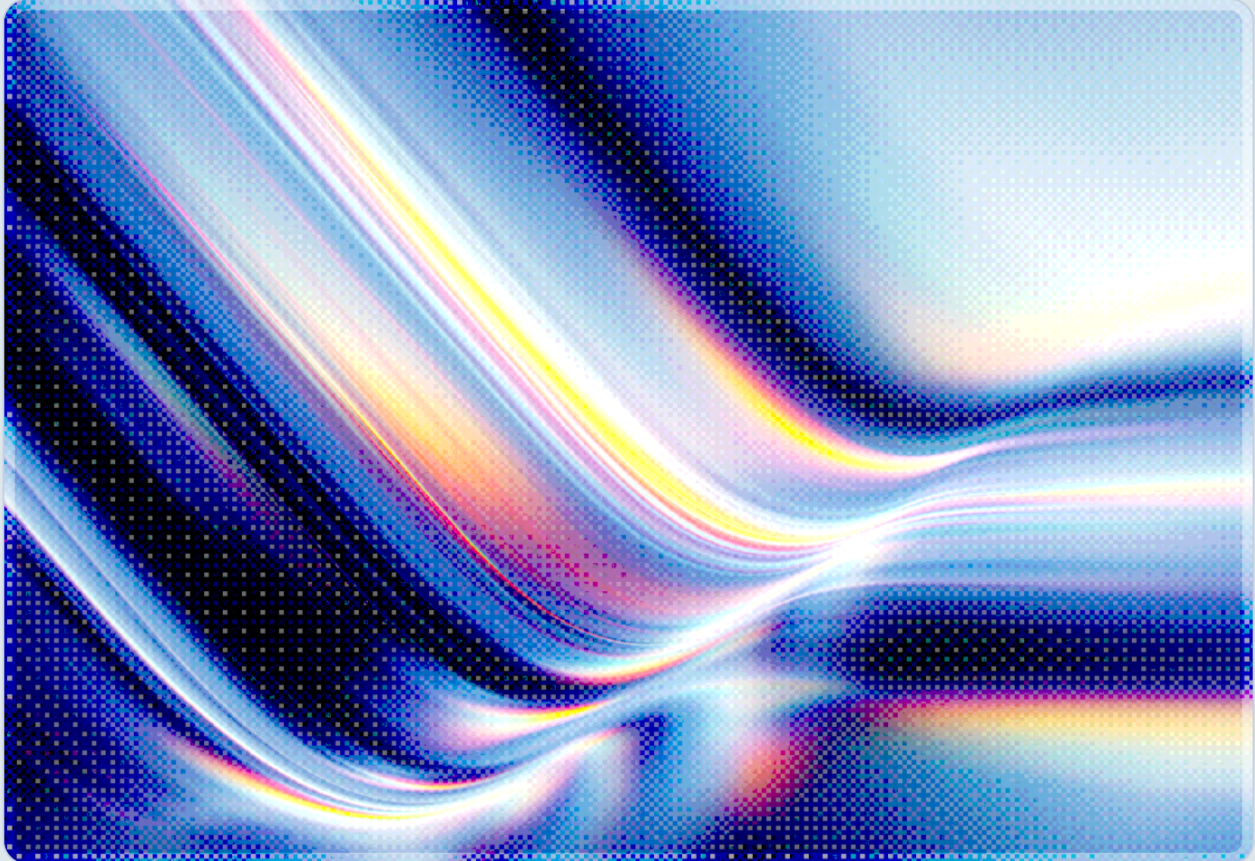


Multi-Agent Offensive Frameworks in the Wild: CyberStrikeAI and ARTEX

Security Implications of the Japan and Korea Campaigns

2026-10-28

 AI-assisted Rapid Research



CSAI

cloud
security
alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

- Over the three weeks ending October 27, 2026, CrowdStrike, GMO Cybersecurity by Ierae, and TrendAI, together with Seoul Metropolitan Police findings reported by BleepingComputer, published evidence that open-source, multi-agent penetration-testing frameworks are being used in real intrusions against organizations in Japan and South Korea [1][3][6][10]. In the same period, Microsoft reported the same tooling in a global cloud supply-chain campaign, and Recorded Future added analysis of the Korean banking activity [7][8].
- Two frameworks dominate the reporting. ARTEX has been linked to a Chinese-speaking, financially motivated operator targeting Korean banks [6][7], to a China-state-aligned actor Microsoft tracks as Storm-0940 [8], to attack infrastructure aimed at about 20 Japanese companies [3], and to a domestic Korean criminal group that police say breached more than 90 organizations [10]. CyberStrikeAI, which Team Cymru linked to the early-2026 FortiGate campaign [11][12], was found running 44 coordinated agents against at least 19 organizations, at least 16 of them in Japan [1].
- The tools are the common thread. The operators are not: they range from state-aligned groups to individual criminals, and as of October 28, 2026, no published source has linked the separate campaigns to a single actor [1][2].
- The frameworks mostly add scale, speed, and repeatability rather than new ways in. In the reports that disclosed an entry point, the vector was an exposed administration interface, default or reused credentials, a legacy web system, or an over-privileged third-party application identity [1][8][12]. This suggests that exposure management and non-human identity governance remain among the highest-value controls against this class of activity.
- Much of what is publicly known about the frameworks comes from attacker mistakes, such as open directories and exposed consoles. Defenders should treat the published picture as a lower bound and not as the full extent of the activity [1][6][7].

Background

When CSA published *Autonomous Agentic AI Adversaries* in June 2026, it argued that AI-driven offensive capability had moved from theory to a present operational reality [14]. The evidence available then was mostly campaigns understood as single operators using commercial models as assistants. The best-

documented example was the January–February 2026 compromise of more than 600 FortiGate devices in over 55 countries, which Amazon Threat Intelligence described as "an AI-powered assembly line for cybercrime" [12]. Team Cymru later tied infrastructure from that campaign to CyberStrikeAI. It observed 21 unique IP addresses running the tool between January 20 and February 26, 2026, and found a CyberStrikeAI service banner on a server that Amazon had linked to the FortiGate activity [11]. In other words, multi-agent tooling was already in use earlier in the year than the public record first made clear.

Freely available frameworks that package the whole intrusion lifecycle as a multi-agent pipeline are now appearing in real intrusions, and two of them, CyberStrikeAI and ARTEX, dominate current public reporting. CyberStrikeAI is written in Go and released under Apache-2.0. Its public repository describes single-agent, plan-execute, and supervisor multi-agent modes built on the Eino orchestration library. It also offers Model Context Protocol (MCP)-native tool integration, more than 100 tool recipes that span reconnaissance through post-exploitation, more than a dozen predefined testing roles, and support for any OpenAI-compatible model endpoint [13]. The repository presents the project as being for authorized testing only, and it includes approval gates, tool allowlists, and audit logs. Nothing in that design, however, stops an operator who controls the deployment from disabling those safeguards.

ARTEX follows the same pattern. Microsoft describes it as an open-source, multi-agent LLM system written in Go with a Next.js front end. It can use any AI model, it records discovered assets and vulnerabilities in a graph, and it runs "an autonomous penetration test from reconnaissance through data exfiltration" with little human interaction [8]. The interface can switch between English, Korean, Mandarin, Spanish, and Japanese [8]. CSA's earlier note on the Korean banking campaign recorded that the developer, who uses the handle Autumn-27, reportedly took the project closed-source after the abuse became public [15]. Copies already in circulation, however, cannot be recalled.

October 2026 brought a cluster of disclosures. On October 9, Japan's National Cybersecurity Office issued an alert to businesses [1]. Over the following weeks, separate vendors and a police agency each published a different part of the picture, and the analysis below brings those reports together.

Security Analysis

Five Campaigns, Two Frameworks, Many Operators

The table below summarizes the major public reports. The figures come from different methods. "Targeted," "attempted," and "confirmed compromised" are different measures and should not be added together or treated as interchangeable. Where a model or assistant tool was observed alongside the framework, it is noted separately.

Report (date)	Framework (models/tools observed)	Operator assessment	Targets and reported scale	Reported entry point
CrowdStrike (Oct 7) [6]; Recorded Future [7]	ARTEX (DeepSeek, GLM, Grok; separate Claude Code sessions)	Unnamed by CrowdStrike; likely Chinese-speaking and financially motivated (moderate confidence); Recorded Future tracks the actor as TAG-217	South Korean financial institutions; at least seven affected, personal data of at least 68,000 people reportedly exposed [7]	Not detailed; a broker loan-inquiry service and an employee mobile work-support system were compromised [6]
GMO Cybersecurity by Ierae (Oct 19) [3] [4][5]	ARTEX	Unattributed; country- and industry-specific campaigns	One attack server targeting about 20 Japanese companies (nine in one industry); six more AI-driven attack servers targeting Japan; eight targets disclosed leaks in Sept–Oct	Not disclosed
Microsoft (Oct 21) [8][9]	ARTEX	Storm-0940, China-state-aligned, active since 2021	Global, no specific countries named; sign-ins against more than 2,250	Compromised OAuth application identities from ConnectWise Cloud Backup

Report (date)	Framework (models/tools observed)	Operator assessment	Targets and reported scale	Reported entry point
			Microsoft 365 tenants via compromised application identities; privileged cloud access in under three hours in several cases	(how they were obtained is not disclosed)
Seoul Metropolitan Police via BleepingComputer (Oct 26) [10]	ARTEX in most attacks	Domestic group of seven suspects led by a former penetration tester	More than 90 Korean organizations; about 8.2 million records; extortion demands in 38 incidents	Not disclosed
TrendAI (Oct 27) [1][2]	CyberStrikeAI with 44 agents (Case A)	Unattributed; Simplified Chinese working language	At least 19 organizations, at least 16 in Japan, plus Taiwan and Hong Kong	Unauthenticated administration interface on a legacy web system

The table shows how widely the operators differ. The same framework, ARTEX, appears with a state-aligned espionage actor running a supply-chain operation, with a financially motivated foreign operator going after bank data, and with a domestic criminal group whose leader police describe as a self-taught former security professional [6][8][10]. Microsoft says plainly that actors in multiple regions are using ARTEX [9]. TrendAI likewise reports no evidence that its three Japanese cases share an operator, and it describes its findings as distinct from the ARTEX activity [1]. For defenders, the conclusion is that attribution to a framework says little about who is behind an attack. Signals such as a tool with Chinese-

language origins or Simplified Chinese prompts are weak evidence of who is at the keyboard. TrendAI cautions explicitly that the working language "is only a weak clue to location or affiliation and cannot establish either" [1], while CrowdStrike rests its assessment partly on such signals and accordingly assigns it only moderate confidence [6].

What "Multi-Agent" Means in Practice

TrendAI's Case A is the clearest public view of an autonomous offensive framework at work. A record reading "credential-search agent 8/44" shows that 44 agents divided the work, with roles that included reconnaissance, vulnerability search, exploitation, and exfiltration [1]. Results were written automatically to a database and turned into structured reports with an evidence index. The dashboard tracked tool execution success rates, token usage, and batch creation [1]. The platform throttled traffic to each target to stay under detection thresholds, and it generated Japanese-language forms and fake login pages [1]. TrendAI classed the case as "AI-driven rather than AI-assisted" because of the scale and consistency of the output. It also noted that "AI changed the scale and repeatability of the attack, but it still depended on an existing entry point" [1].

Microsoft's account of Storm-0940 shows what this looks like from the target's side, and it describes two distinct phases. Between October 10 and 14, a password-spray phase generated more than 151,000 unsuccessful logon events against 327 accounts at eight organizations, peaking at 2,900 failures per minute. None of those attempts succeeded [8]. The more consequential path, which reached the more than 2,250 tenants, ran through compromised identities for ConnectWise Cloud Backup applications; Microsoft does not say how the actor obtained them [8]. These were pre-authorized, multi-tenant OAuth applications, so authenticating with them required no interactive user sign-in [8][9]. In one subset of activity, the actor made more than 200 sign-ins in under four minutes using four application credentials. Agents then ran Microsoft Graph reconnaissance, added application credentials for persistence, and changed identities, groups, and privileged roles [8]. Microsoft warns that these changes can persist as a backdoor after the original application has been remediated [8][9].

These cases suggest that the main operational change is economic rather than technical. Each individual technique, from credential spraying to Graph enumeration, is well known. What the frameworks change is how much a small team can attempt in parallel and how little analyst time each additional target requires. Korean investigators drew a broader conclusion, telling Yonhap, as reported by BleepingComputer, that "even individuals who are not professional hacker organizations can now hack on a large scale with the help of AI" [10]. That case is imperfect evidence for the point, however. The group's leader had professional penetration-testing experience, and the report says the leader adopted AI utilities in late

August 2026 to expand a campaign that began in September 2025. Police describe ARTEX as used in most of the attacks, but the report does not say how many of the more than 90 intrusions came after that adoption, so the totals cannot be credited to AI alone [10].

Model Supply Chain and Guardrail Evasion

The model backends are a second point of interest. CrowdStrike found that the Korean banking operator ran ARTEX primarily on DeepSeek v4.1-flash, likely accessed through an API reseller. The operator added GLM-5.3 and Grok 4.6 in separate Claude Code sessions, and also asked Claude where Korean breach data is typically sold [6]. Because these frameworks accept any OpenAI-compatible endpoint [8][13], a single provider's safety controls cannot reliably stop them. An operator blocked by one provider can switch to another, to a reseller, or to stolen credentials. CSA's LLMjacking research has documented that last path as an emerging feature of offensive agent infrastructure [19]. TrendAI's Case B shows the guardrail-evasion side: the operator framed requests as an authorized assessment, which TrendAI describes as "a false premise intended to circumvent the model's safety restrictions" [1].

Case B also shows that AI infrastructure is now a target in its own right. That operator's lists included AI workflow platforms, LLM preprocessing services, and inference endpoints, apparently in order to obtain large amounts of inference capacity cheaply [1]. This points to a potential feedback loop, in which inference services left exposed by defenders become fuel for the next round of agent-driven attacks.

Entry Points Remain Familiar

Across the reports that disclosed an entry vector, most of the weaknesses were ordinary. Amazon observed no FortiGate vulnerability exploitation in the earlier campaign, only exposed management ports and weak single-factor credentials [12]. TrendAI's three Japanese cases relied on an unauthenticated administration interface, a GIS server with default credentials, and Yamaha RTX1210 router credentials recovered from configuration files [1]. Storm-0940 is different in kind. It abused a trusted third-party application's identities, compromised upstream at the supplier, which victim tenants could not see or close themselves [8]. The enabling condition on the victim side, broad pre-authorized permissions for that application, is nonetheless a familiar governance gap. The CrowdStrike, GMO, and Korean police reports disclose no entry vector at all, so these observations rest on a partial sample.

Recorded Future similarly observed that the evidence "does not establish the extent to which AI accelerated these attacks," and that each Korean victim had different systems compromised, a pattern consistent with testing many targets until some yielded value [7]. TrendAI summarized the situation this way: "AI has increased the scale and speed of these attacks, but the attackers are still looking for an unlocked door" [1].

Visibility Comes from Attacker Mistakes

Much of the framework-level detail above, including agent counts, configuration files, model choices, and session histories, comes from operator mistakes rather than from defensive detection. CrowdStrike and Recorded Future reconstructed the Korean banking campaign from open directories that exposed Claude Code session histories, ARTEX configuration files, and memory files [6][7]. TrendAI rebuilt all three of its Japanese cases from files left in publicly accessible directories on attacker servers [1]. Team Cymru identified CyberStrikeAI deployments through service banners [11]. Microsoft's Storm-0940 analysis is the main exception, drawn from its own telemetry [8], and the Korean case comes from a law-enforcement investigation [10]. More careful operators using the same frameworks would leave far fewer traces. The published campaign counts are therefore best read as a floor, and defenders should not assume that an absence of public reporting on their sector means an absence of activity.

Recommendations

Immediate Actions

Organizations with operations in Japan, South Korea, Taiwan, or Hong Kong, and any Microsoft 365 tenant that has authorized ConnectWise Cloud Backup, should act first. Because Microsoft reports ARTEX use by actors in multiple regions [9], the same controls apply to organizations elsewhere. The steps below follow the defensive guidance in the primary reports [1][8].

The first priority is the external attack surface. Organizations should inventory internet-facing web systems, administration interfaces, VPN appliances, and edge devices, remove default credentials, require MFA on all management access, and isolate or retire legacy platforms that cannot be patched. Credentials stored in device configuration backups should be rotated, since one of TrendAI's Japanese cases relied on router credentials recovered from configuration files [1].

Tenants that authorized ConnectWise Cloud Backup should follow Microsoft's sequence. They should keep the affected connections disabled, preserve Entra, Purview, and SIEM evidence before making changes, and search by application ID rather than display name. Microsoft advises reviewing activity from at least October 15, 2026 and, where records are available, extending service-principal sign-in review to October 1, including failed authentications and volume anomalies [8]. Because the password-spray phase ran from October 10 to 14, any lookback should cover those dates as well. More broadly, all service principals should be reviewed for recent credential additions, new owners, consent grants, and privileged role assignments. Finally, published indicators from CrowdStrike, Microsoft, and TrendAI should be added to blocking and retrospective hunting, recognizing that attacker infrastructure rotates quickly [1][6][8].

Short-Term Mitigations

Detection logic needs adjusting for agent-driven traffic. TrendAI observed per-target throttling designed to evade threshold alerts, so monitoring should correlate low-frequency, distributed attempts across assets and time rather than relying on per-source rate limits [1]. Recorded Future recommends connecting external infrastructure intelligence with internal signals such as unusual sign-ins and abnormal data transfers, and pre-approving proportionate automated containment for well-understood threats [7].

Non-human identities deserve the same scrutiny as privileged human accounts. Multi-tenant SaaS integrations, backup tools, and OAuth applications should be held to least privilege and reviewed on a regular schedule, and organizations should test that they can revoke them quickly [8][9]. Organizations running internal AI services should confirm that inference endpoints, workflow platforms, and GPU clusters are not exposed to the internet. They should also monitor egress for calls to unapproved LLM APIs and reseller endpoints, which can indicate a compromised host being used as agent infrastructure [1][19].

Strategic Considerations

The cost of finding and exploiting ordinary weaknesses is falling. As a result, the time between exposure and exploitation is likely to shrink for organizations of every size, including small businesses and local public bodies, which TrendAI found among Case A's targets [1]. CSA's assessment is that continuous automated external attack-surface testing should now be treated as a baseline control rather than an advanced one. Running authorized agentic testing tools against one's own environment, under written scope, is one practical way to see what these frameworks see.

Dual-use release governance also needs attention from the security community. Both frameworks were published as security-testing tools, CyberStrikeAI with optional safeguards that operators can disable [13], and both have since appeared in the hands of multiple hostile operators [1][6][8][10][11]. Taking a repository closed-source after abuse becomes public does not recall existing copies [15]. Model providers, open-source maintainers, and code-hosting platforms should develop shared norms for staged release, abuse reporting, and provenance of offensive agent tooling. Enterprises, for their part, should establish a channel for reporting suspected abuse to the AI providers whose models appear in their incident evidence.

CSA Resource Alignment

CSA's [Autonomous Agentic AI Adversaries](#) (June 2026) is the most directly relevant prior work [14]. It argued that autonomous agentic adversaries were already operational and mapped defensive guidance to MAESTRO, AICM, and Zero Trust. The October campaigns are consistent with that thesis, though the degree of autonomy varies across cases and Recorded Future notes that the extent of AI acceleration is not yet established [7]. They also add something the June paper could only anticipate: the same packaged framework reused across state-aligned, foreign criminal, and domestic criminal operators.

This note builds on two earlier CSA rapid-research notes. [ARTEX: Open-Source Agentic Pentest Tool Turned on Korean Banks](#) (October 10, 2026) covered the first ARTEX disclosure, the tool's architecture, and the developer's move to closed source [15]. [AI-Assisted Mass Network Infrastructure Exploitation: The 600+ FortiGate Campaign](#) analyzed the campaign later linked to CyberStrikeAI [16]. Read together, the three notes trace how public evidence evolved in 2026: from campaigns first understood as single-operator AI assistance, some of which were later linked to multi-agent tooling, to direct observation of packaged multi-agent frameworks used by many unrelated actors. CSA's [LLMjacking research](#) complements these notes by examining how stolen AI compute and API access become offensive agent infrastructure [19].

For threat modeling, CSA's [MAESTRO framework](#) provides a layered model for analyzing agent ecosystems [17]. Its agent-ecosystem and deployment-infrastructure layers provide a useful lens for assessing how an organization's own AI services could be abused as attacker inference capacity, as in TrendAI's Case B. The [AI Controls Matrix \(AICM\) v1.1](#) supplies the control baseline [18]. Its Threat and Vulnerability Management domain addresses the exposure-management practices relevant to the disclosed entry points in the TrendAI and FortiGate cases, namely unauthenticated administration interfaces, default credentials, and exposed management ports. Its Identity and Access Management domain addresses the third-party application-identity governance that could have limited the impact of the Storm-0940 supply-chain path.

References

- [1] TrendAI (B. Matsukawa and J. C. Chen). "[Not Only ARTEX: The Different Levels AI Is Used in Attacks Targeting Japan](#)." TrendAI Deep Research, October 27, 2026.
- [2] R. Lakshmanan. "[CyberStrikeAI Uses 44 AI Agents to Automate Hacking Operations Against Japan Targets](#)." The Hacker News, October 27, 2026.
- [3] GMO Cybersecurity by Ierae (D. Fukumori). "[Many Sources of the Recent Unauthorized Access Incidents Confirmed to Be AI \(Japanese\)](#)." GMO Cybersecurity by Ierae Blog, October 19, 2026.
- [4] ScanNetSecurity. "[GMO Cybersecurity by Ierae Confirms AI Autonomous Pentest System ARTEX at Source of Unauthorized Access Attacks \(Japanese\)](#)." ScanNetSecurity, October 21, 2026.
- [5] M. Negishi. "[AI Tool That Hacked Korean Banks Was Used in Japan: Security Firm](#)." Insurance Journal (Bloomberg), October 20, 2026.
- [6] CrowdStrike Intelligence. "[Unknown Threat Actor Uses AI-Driven ARTEX to Target South Korean Finance](#)." CrowdStrike Blog, October 7, 2026.
- [7] Recorded Future Insikt Group. "[TAG-217 Targets South Korean Financial Services with ARTEX](#)." Recorded Future, October 2026.
- [8] Microsoft Threat Intelligence. "[AI-Driven, Large-Scale Nation-State Campaign Using ARTEX Enables Supply Chain Compromise](#)." Microsoft Security Blog, October 21, 2026.
- [9] G. Swain. "[China-Linked Hackers Used Agentic AI in Attack on 2,250 Microsoft 365 Tenants](#)." CSO Online, October 22, 2026.
- [10] B. Toulas. "[Former Pentester Used ARTEX AI to Steal 8.2 Million Data Records](#)." BleepingComputer, October 26, 2026.
- [11] W. Thomas, Team Cymru. "[Tracking CyberStrikeAI: AI-Native Offensive Tools & MSS Ties](#)." Team Cymru, March 2, 2026.
- [12] C. J. Moses, Amazon Threat Intelligence. "[AI-Augmented Threat Actor Accesses FortiGate Devices at Scale](#)." AWS Security Blog, February 20, 2026.
- [13] CyberStrikeAI Project. "[CyberStrikeAI Repository](#)." GitHub (formerly EdIsOnZ/CyberStrikeAI), accessed October 28, 2026.

- [14] Cloud Security Alliance. "[Autonomous Agentic AI Adversaries](#)." CSA AI Safety Initiative, June 24, 2026.
- [15] Cloud Security Alliance. "[ARTEX: Open-Source Agentic Pentest Tool Turned on Korean Banks](#)." CSA AI Safety Initiative, October 10, 2026.
- [16] Cloud Security Alliance. "[AI-Assisted Mass Network Infrastructure Exploitation: The 600+ FortiGate Campaign](#)." CSA AI Safety Initiative, March 8, 2026.
- [17] K. Huang, Cloud Security Alliance. "[Agentic AI Threat Modeling Framework: MAESTRO](#)." CSA Blog, February 6, 2025.
- [18] Cloud Security Alliance. "[AI Controls Matrix \(AICM\) v1.1](#)." Cloud Security Alliance, June 22, 2026.
- [19] Cloud Security Alliance. "[LLMjacking Evolution: From Cost Theft to Offensive Agent Infrastructure](#)." Cloud Security Alliance, 2026.