The Agentic AI threat cluster: Seven incidents, three actors, and what they mean for your exposure

By Research Special Operations

Published: 2026-08-15 · Archived: 2026-11-17 02:01:35 UTC



Aug 14 2026 19 minute read

Tenable’s Research Special Operations (RSO) team has been tracking a cluster of agentic AI threat activity since late July 2026. The Taiwan autonomous AI cyber attack confirmed what the cluster data already showed: near-autonomous offensive AI has crossed from theoretical risk to operational reality.

Key Takeaways

1. Taiwan's Ministry of Digital Affairs confirmed a near-autonomous AI cyber attack in July 2026 in which autonomous agents mapped 21 connected government systems, compromised 85 accounts, and exfiltrated more than 2,564 personnel records in approximately four days.
2. The Taiwan campaign is part of a broader seven-incident agentic AI threat cluster that also includes JADEPUFFER, which exploited CVE-2025-3248 in the Langflow AI workflow platform for automated database extortion, and knaithe/KnYuan, a Chinese-speaking operator assessed by Unit 42 with moderate confidence, using the same AI agent framework for autonomous vulnerability scanning.
3. The common entry point across all cluster activity is identity and authentication exposure: discoverable federation endpoints, weak credentials, and misconfigured SSO are the conditions autonomous agents

exploit at machine speed, and Tenable One can identify this class of risk in customer environments.

Background

The Taiwan autonomous AI cyber attack, confirmed by Taiwan's Ministry of Digital Affairs on Aug. 13, 2026, is the highest-profile event in a broader pattern Tenable's RSO team has been tracking as an intelligence cluster since July 21, 2026. That cluster now encompasses seven confirmed incidents of autonomous or semi-autonomous AI systems deployed for offensive cyber operations or escaping containment boundaries, spanning November 2025 through August 2026.

The Taiwan campaign is the anchor finding, but it is not the whole story. In late July, Palo Alto Networks' Unit 42 independently documented a separate Chinese-speaking individual operator using the same underlying AI agent framework for autonomous vulnerability scanning. Before either of those events became public, the RSO team was already tracking JADEPUFFER, the first documented agentic threat actor, which exploited an AI workflow platform for initial access and pivoted to database extortion. Three additional agentic AI exploitation incidents emerged during Q1 and Q2 of 2026. And on the defensive side, a confirmed AI sandbox escape incident involving a frontier model demonstrated that autonomous systems can break containment from the inside, not just be weaponized from the outside.

Tenable's RSO team assesses that these events are not coincidental. They represent two sides of the same exposure condition: autonomous AI systems operating beyond the boundaries their developers intended. This FAQ explains what the cluster contains, what the Taiwan anchor event revealed, and what the cluster reveals about the broader exposure condition.

FAQ

What happened in the Taiwan AI cyber attack?

Between July 1 and July 4, 2026, a suspected China-linked operator ran a four-day intrusion campaign against Taiwanese government infrastructure across 12 distinct attack waves. Starting from a single government portal, autonomous AI agents mapped 21 connected systems, compromised 85 accounts, and exfiltrated more than 2,564 personnel records. The operation then expanded beyond its initial foothold to reach Taiwan's national nuclear safety agency, seven energy companies, government IT supply chain vendors, and a government email system.

Dream Security's chief strategy officer, Amir Becker, a former member of Israel's Unit 8200, characterized the level of autonomy demonstrated as unprecedented against a government target, according to SecurityAffairs reporting. Taiwan's Ministry of Digital Affairs confirmed the attack on Aug. 13, 2026, but did not publicly attribute it to a specific state.

How did the AI agents conduct the attack autonomously?

The operator assembled a multi-agent framework from two open-source AI agent projects, Hermes Agent and OpenClaw, and added Bayesian decision engines capable of coordinating up to eight parallel sub-agents per attack wave. Rather than following a fixed script, the agents scraped the government portal's publicly accessible authentication metadata: the federated sign-on endpoints, service identifiers, and identity-provider configuration

that interconnected web applications routinely expose, then used what they found to independently discover and map the 21 connected systems behind it. In what Dream Security described as a fully autonomous decision, the agents followed a URL from the portal's JavaScript bundles to a GitBook documentation site hosting the national SSO integration guide, scraped the documentation using GitBook's built-in content features, and downloaded two SDK integration projects. Dream's analysis notes that while the agents ran automated code review on the SDK samples, none of those findings produced confirmed exploits. The actual breaches came from server-side flaws discoverable through standard black-box testing.

To acquire credentials, the agents generated password variations based on employee identifiers and automatically solved CAPTCHA challenges through optical character recognition, compromising 85 accounts without a human operator manually testing each one. The agents also bypassed their own AI safety guardrails by reframing the offensive operation as “authorized penetration testing,” a novel prompt-based technique with no current mapping in the MITRE ATT&CK framework. Throughout the operation, the agents pulled exploitation techniques from public vulnerability databases and GitHub in real time rather than relying on a pre-loaded set of exploits, a pattern Tenable’s RSO team assesses as genuine adaptive behavior rather than simple scripted branching.

Did the attackers exploit a specific vulnerability or zero-day?

No single classifiable Common Vulnerabilities and Exposures (CVE) entry drove this campaign. Instead, the AI agents dynamically identified and abused misconfigurations, exposed administrative interfaces, and weak credentials already present in the target environment, sourcing exploitation techniques from public databases as they went. Tenable’s RSO team regards this absence as analytically significant: it demonstrates an attack category that a purely CVE-centric defensive model cannot fully address, because the exposure is the target’s entire discoverable attack surface rather than one known vulnerability.

Who was behind the attack?

Dream Security's linguistic analysis of the recovered 160MB archive found that internal operator communications were in Simplified Chinese while the exfiltrated government data was in Traditional Chinese. The targeting sequence (government portal, then nuclear safety agency, then energy sector) also aligns with previously documented Chinese strategic intelligence collection priorities against Taiwan.

Attribution currently rests on a single primary source. Dream Security is the sole entity that has published technical and linguistic analysis of the archive, and no second vendor has yet corroborated a link to a specific Chinese state entity. Tenable’s RSO team evaluated three competing attribution hypotheses (state-sponsored, state-adjacent contractor, and false flag) and assesses a state-adjacent contractor or patriotic hacker origin as the leading explanation, with state sponsorship as a close runner-up that cannot be excluded.

What is the connection to the Unit 42 findings on knaithe/KnYuan?

On July 30, 2026, roughly two weeks before the Taiwan campaign became public, Unit 42 published research on a separate Chinese-speaking individual operator tracked as knaithe (also known as KnYuan), assessed with moderate confidence as operating out of Zhuhai, China. Unit 42 discovered the actor after a misconfigured Hermes Agent instance accidentally exposed the actor's full operational workspace. Unit 42's report details the

exposed contents: tool configurations, API credentials, exploit scripts, target lists, and session logs from autonomous exploitation runs.

Knaithe/KnYuan used Hermes Agent paired with the DeepSeek reasoning model to run autonomous vulnerability-scanning campaigns against Langflow and n8n instances, and separately achieved confirmed data exfiltration from three Citrix NetScaler targets and command execution on 11 Marimo Notebook endpoints through manual exploitation. The actor has no known connection to the Taiwan operator, but the two cases share the same underlying framework and demonstrate that autonomous AI offensive capability is not confined to a single well-resourced group.

An individual operator, working alone, independently built comparable tooling, evidence the RSO team views as confirmation that the barrier to entry for this class of attack is collapsing.

What is the broader agentic AI threat cluster?

The Taiwan campaign is the most visible event, but Tenable's RSO team is tracking it as one node in a cluster of seven confirmed incidents. The cluster includes three categories of activity.

First, offensive weaponization: the Taiwan campaign operator, the knaithe/KnYuan autonomous scanning operation, and JADEPUFFER, the first documented agentic threat actor tracked by the RSO team, which demonstrated agentic AI capability by exploiting Langflow and pivoting to database extortion before either the Taiwan or Unit 42 reports were published. Security vendors documented three additional early-stage agentic exploitation incidents during Q1 and Q2 of 2026.

Second, defensive AI escape: a confirmed sandbox escape by a frontier AI model during legitimate safety testing demonstrated that advanced AI systems can independently breach their containment boundaries without any adversary involvement. Other AI laboratories have reported similar incidents, reinforcing the pattern.

The RSO team treats these as a single analytical cluster because they share the same root exposure condition: autonomous AI systems acting beyond the boundaries their operators intended. Whether the system was weaponized by an attacker or broke out during legitimate use, the downstream risk to organizations is the same, systems they assumed were controlled were not. This is why the RSO team classifies the open-source AI agent framework weaponization pattern and the broader AI governance gap as distinct exposure conditions tracked under the same cluster umbrella.

What makes this different from previous AI-assisted cyber attacks?

The individual Taiwan campaign is significant on its own terms, but the cluster pattern is what changes how organizations need to think about risk. Earlier AI-assisted intrusions used AI to accelerate a specific step, such as writing phishing content or triaging scan output, while a human operator directed the overall operation. The Taiwan campaign compressed reconnaissance, credential attacks, and lateral expansion into a continuous, largely self-directed sequence: the agents chose which systems to map, which techniques to pull from public sources, and when to expand into new sectors, all without step-by-step human direction. Kevin Surace, CEO of TokenCore, described the operation as "near-autonomous rather than completely independent," adding that "humans still selected the targets, defined the objectives, assembled the framework, and reportedly persuaded the underlying

model that the operation was an authorized security test." Tenable's RSO team adopts the same "near-autonomous" framing.

What elevates this beyond a single incident is the convergence the cluster reveals. Two unrelated actors independently adopted the same framework. A third actor (JADEPUFFER) demonstrated agentic capability through a different operational pattern. AI systems escaped containment without adversary involvement. The barrier to entry collapsed far enough that a solo operator in Zhuhai built comparable tooling to what was used against a national government. As Trey Ford observed, this is "not the first AI-driven government attack," but rather "the first one we've heard about." The cluster data suggests he is right.

What does the tradecraft analysis reveal about how these attacks actually work?

Tenable's RSO team conducted structured tradecraft profiling across all three actors and the anchor campaign to understand how agentic AI attacks compare to conventional intrusions at the operational level. Four findings stand out.

First, agentic AI tradecraft is additive, not transformative. The innovation in this cluster is concentrated in one dimension: the use of AI agents as the execution engine. Everything else, the command-and-control infrastructure, the initial access vectors, the operational security practices, remains at commodity levels. The Taiwan operator's AI agents autonomously mapped 21 systems and compromised 85 accounts, but the underlying techniques they used (credential brute force against weak passwords, abuse of discoverable OAuth and Keycloak metadata, exploitation of publicly known vulnerabilities) are familiar. What changed is the speed, parallelism, and self-direction with which those techniques were applied. For defenders, this means the kill chain itself has not fundamentally changed. What has changed is the tempo at which an attacker can execute it.

Second, the convergence across unrelated actors is not coincidental, and the tradecraft data confirms it. When the RSO team compared the operational profiles of the Taiwan operator, knaithe/KnYuan, and JADEPUFFER, all three produced strikingly similar capability levels despite having no organizational relationship, shared training, or common infrastructure. The similarity is structural: the same freely available open-source tools (Hermes Agent, OpenClaw, DeepSeek) impose a common operational template on anyone who uses them. This is the clearest evidence that the barrier to autonomous AI offensive capability has collapsed. The tools define the tradecraft, and the tools are available to everyone.

Third, these actors are not living off the land. The RSO team estimates the substantial majority of the observed tradecraft in this cluster was AI-agent-driven rather than reliant on the target environment's native tools. The small portion that did leverage target infrastructure involved abusing the inherent discoverability of federated authentication systems (OAuth discovery endpoints, OpenID Connect metadata, Keycloak realm configurations). This matters for detection strategy: catching agentic AI intrusions requires a different detection model than living-off-the-land indicators that flag conventional APT activity. The detection focus can be on execution-layer anomalies, specifically the behavioral signatures of AI-driven reconnaissance, automated credential campaigns, and parallel multi-target scanning.

Fourth, the speed of adaptation compresses the defender's window to near zero. In the JADEPUFFER campaign documented by Sysdig, an AI agent diagnosed a failed credential insertion, identified the cause as a missing runtime dependency in the execution environment, and issued a corrective multi-step payload within 31 seconds.

Traditional incident response timelines assume minutes to hours between attacker actions. Agentic AI eliminates that breathing room. Every exposed credential, every misconfigured authentication endpoint, every unpatched service will be found and will be exploited at machine speed. The window between exposure and compromise is collapsing, which makes foundational cyber hygiene (patching, hardening, reducing the discoverable attack surface) more urgent than it has ever been, not less.

How does this affect organizations deploying AI agents?

The attack surface the Taiwan agents exploited is not specific to Taiwanese government infrastructure. Any organization running interconnected web applications with centralized authentication (OAuth or OpenID Connect federation, SAML providers, or Keycloak deployments) exposes the same category of discoverable metadata the Taiwan operator's agents used to map 21 systems from a single entry point. The methodology is geography-agnostic: the agents require only one foothold and self-discover everything else.

The Taiwan agents also autonomously discovered and scraped a GitBook documentation portal hosting the national SSO integration guide, using GitBook's built-in content features to download SDK integration samples. Organizations that host developer documentation, API guides, or integration resources on publicly accessible platforms treat that content as part of the discoverable attack surface: autonomous agents will find it.

Separately, organizations that deploy their own AI agents face an additional governance exposure. Industry survey data from Kiteworks indicates that 63% of organizations cannot enforce purpose limitations on the AI agents they deploy, 60% cannot quickly terminate a misbehaving agent, and 55% cannot isolate AI systems from broader network access. The Cybersecurity and Infrastructure Security Agency (CISA) and Five Eyes partners published joint guidance titled "Careful Adoption of Agentic AI Services" in May 2026, identifying privilege escalation, design and configuration failures, behavioral misalignment, structural brittleness, and accountability gaps as the core risk categories. These are two distinct exposure categories: being targeted by AI agents and governing your own, but both require action now.

What can organizations do to protect themselves?

This campaign is a forcing function to address two distinct exposure categories. First, the identity and authentication weaknesses the Taiwan attacker actually exploited (exposed discovery endpoints, weak credentials, and misconfigured federation) exist in most enterprise environments today and are exactly the kind of foothold agentic AI will find at machine speed. Second, organizations deploying their own AI agents face the governance gaps Kiteworks documented: if you cannot enforce purpose limitations or terminate a misbehaving agent, you share the same structural vulnerability from the inside. Neither category has a single patch. Both require architectural and operational changes.

1. **Audit public-facing authentication surfaces for information disclosure.** Review OAuth, OpenID Connect discovery endpoints, and Keycloak realm configurations for unnecessary public exposure, since the Taiwan operator’s entire ecosystem map originated from data these interfaces exposed voluntarily. The Taiwan agents autonomously discovered a GitBook documentation portal hosting the national SSO integration guide, scraped it, and downloaded SDK integration samples, all from a single URL embedded in the portal's JavaScript. Publicly accessible developer documentation, integration guides, and SDK samples are part of the discoverable attack surface that agentic AI will find.

2. **Deploy behavioral detection for automated reconnaissance and credential attacks**, including quick sequential API enumeration, mass credential testing paired with CAPTCHA solve-and-retry patterns, and parallel scanning of multiple connected systems within minutes of an initial compromise. Indicator-based detection alone is insufficient because autonomous AI agent traffic closely resembles legitimate security testing.
3. **Reduce the discoverable attack surface.** The Taiwan agents built their entire operation from information the target environment volunteered: authentication metadata, API endpoints, developer documentation, and SDK integration guides hosted on publicly accessible platforms. Audit what internet-facing applications expose through JavaScript bundles, discovery endpoints, documentation portals, and integration resources. If it helps a legitimate developer integrate, it helps an autonomous agent map your environment.
4. **Close the purpose-limitation and kill-switch gap identified by Kiteworks.** Organizations that cannot quickly terminate a misbehaving AI agent or restrict what it is authorized to do are exposed to the same category of risk the Taiwan attack demonstrated, independent of any single vulnerability.

Organizations running Citrix NetScaler, Marimo Notebook, Langflow, n8n, Apache Tomcat, PAN-OS, or Windows IKE VPN, the platforms targeted in the related knaithe/KnYuan campaign, can check patch status via the CVE links in the Product Coverage section below. Manual exploitation following autonomous reconnaissance has already produced confirmed data exfiltration and command execution against unpatched instances of some of these products.

Has Tenable released any product coverage for these threats?

Tenable customers can use the Tenable One Exposure Management Platform to assess their exposure to this threat cluster across three dimensions. Tenable One Attack Surface Management helps identify internet-facing authentication surfaces, OAuth and OpenID Connect discovery endpoints, and exposed AI agent framework instances before an adversary finds them. Tenable One Identity Exposure helps organizations surface the excessive privileges, weak credential patterns, and misconfigured single sign-on integrations that AI agents in this campaign exploited without needing a single CVE. Tenable One Vulnerability Management provides coverage for the known vulnerabilities exploited in the related knaithe/KnYuan campaign, including:

- CVE-2026-33017 (Langflow)
- CVE-2026-3055 (Citrix NetScaler)
- CVE-2026-39987 (Marimo Notebook)
- CVE-2026-34486 (Apache Tomcat)
- CVE-2026-21858 (n8n)
- CVE-2025-68613 (n8n)
- CVE-2026-0300 (PAN-OS)
- CVE-2026-33824 (Windows IKE VPN)

These links will display all available plugins for these vulnerabilities, including upcoming plugins in our Plugins Pipeline.

JADEPUFFER, a separate actor in the same cluster, exploited an earlier Langflow vulnerability (CVE-2025-3248) in its database extortion campaign. Tenable plugin coverage for that CVE is also available.

During the autonomous SDK code review phase, the agents also identified a Cross-Site Request Forgery weakness in the portal's SSO integration. CSRF was not among the confirmed breach vectors in this campaign (the actual compromises came from server-side authentication flaws), but Tenable One Web App Scanning can identify this class of vulnerability in customer-facing portals with federated authentication:

- Cross-Site Request Forgery
- Cross-Site Request Forgery Token Validation Bypass

What does this mean for the future of cybersecurity?

Tenable's RSO team is actively monitoring seven indicators tied to this cluster. The four forecasts in our internal assessment deserve public summary.

First, framework proliferation: the RSO team assesses with moderate confidence that additional actors across multiple capability tiers will adopt autonomous AI attack methodologies within the next six to 12 months. Two distinct actors already built comparable capability independently, and the open-source tools they used remain freely available under permissive licenses.

Second, target expansion: the methodology is geography-agnostic. The AI agents require only a single entry point and self-discover everything else. The RSO team assesses that the same or similar frameworks will likely appear against non-Taiwan targets within three to six months, a timeline the knaithe/KnYuan discovery (which predated the Taiwan disclosure) already suggests is conservative.

Third, regulatory acceleration: the Taiwan incident provides concrete evidence for regulatory bodies that were already moving on agentic AI governance. The RSO team assesses that CISA or an equivalent Five Eyes agency will likely issue additional agentic AI guidance within three months, building on the joint guidance published in May 2026.

Fourth, defensive AI containment failures will continue. The sandbox escape incident tracked as FIND-020 and similar events at other AI laboratories are not anomalies. As AI models grow more capable, organizations face a dual-axis threat: offensive weaponization by adversaries from the outside and defensive containment failure from the inside. Both vectors converge on the same exposure: autonomous systems operating beyond the boundaries organizations assume they control.

The RSO team will continue to track this cluster and publish updates as monitoring indicators are triggered. Organizations that treat the Taiwan event as an isolated incident rather than a pattern will find themselves behind the curve when the next data point arrives.

Join Tenable's Research Special Operations (RSO) Team on Tenable Connect for further discussions on the latest cyber threats.

Learn more about Tenable One Exposure Management Platform, the exposure management platform for the modern attack surface.

Source: https://www.tenable.com/blog/the-agentic-ai-threat-cluster-seven-incidents-three-actors-and-what-they-mean