Open-source AI pentesting framework Cairn now seen in attacks documented by four security firms



Amazon, Gambit, NorForge and Unit 42 have each documented attackers using the open-source Cairn framework to run autonomous AI agents against real targets. Defenders are advised to automate response and prove recovery.

- Amazon Threat Intelligence describes a Chinese-nexus actor whose Cairn-based agents, given only IP addresses and a goal, targeted East Asian universities from September 8, 2026.
- After about a day of failed attempts, the agents exploited a previously unknown path traversal flaw in an open-source medical imaging AI application and obtained a root reverse shell.
- Amazon is the fourth security company in about two months to document Cairn used against real targets, following Gambit Security, NorForge and Unit 42.
- Other open-source frameworks, including ARTEX, PentAGI, HexStrike AI and CyberStrikeAI, are also

> in use by criminal and espionage groups.
>
> - Advice centers on automated high-confidence response, identity hygiene, a smaller attack surface, patching, and proven restoration of revenue-critical systems.

Several security firms have now reported attackers using Cairn, an open-source framework that its developer describes as a general-purpose engine for finding a path to a goal through an unknown state space. It was not built specifically for penetration testing but has proven effective at it.

What Amazon observed

Amazon Threat Intelligence found a publicly hosted Cairn instance running without authentication and, from its configuration, prompts and output, assessed the operator as a Chinese-nexus actor. The operator supplied target IP addresses and the goal of obtaining shell access. Starting on September 8, 2026, the agents worked against East Asian universities, spending about 16 hours on an unauthenticated Redis server and then more than six hours trying over 30,000 credential combinations built from roughly 80 usernames gathered from public sources. None of this succeeded.

The breakthrough came through a previously unknown path traversal flaw in an upload endpoint of an open-source medical imaging AI application. The agents went from mapping the API to writing arbitrary files in about 11 minutes, and later obtained a reverse shell running as root. A human operator stepped in once, to give the agents a tunnel into a second university network running the same software.

A wider pattern

Gambit Security, NorForge and Unit 42 have reported similar use, by both financially motivated and suspected China-nexus actors, against retailers, semiconductor, defense and civil society targets. Unit 42 noted that the flaws exploited were mostly weak credentials, authentication bypasses and old CVEs. Gambit argues that fast patching alone is no longer enough.

Defensive direction

Amazon recommends high-confidence detections that trigger automated response, with humans overseeing outcomes, because agents can outpace SOC triage. Defenders should expect persistent attacks that do not tire. Gambit advises identifying the systems that keep revenue flowing and proving they can be restored under attack.

What to do now

1. Review SOC playbooks so high-confidence detections trigger automated containment, with analysts overseeing outcomes rather than reviewing every action.
2. Inventory internet-facing services such as Redis, vector databases, admin dashboards and VPNs, and

remove unauthenticated access or unneeded exposure.

3. Strengthen identity hygiene: enforce MFA, remove weak and default credentials, and monitor for large-scale password spraying.
4. Check open-source AI and imaging applications in your estate for exposed upload endpoints and apply vendor patches promptly.
5. Identify revenue-critical systems and test restoration under simulated attack, recording results for ECC and SAMA CSF assurance.