

AI Agents as Attack Vectors (CIAD-2026-0049)

PRIMARY

Publisher	CERT-In
Published	2026-10-15
Retrieved	2026-10-17
Cite key	certin2026ai
URL	https://www.cert-in.org.in/s2cMainServlet?pageid=PUBVLNOTES02&VLCODE=CIAD-2026-0049

Claims drawn from this source. CLAIM-2026-10-17-01, CLAIM-2026-10-17-02, CLAIM-2026-10-17-03, CLAIM-2026-10-17-04, CLAIM-2026-10-17-05

Advisory ID. CIAD-2026-0049

Full title. AI Agents as Attack Vectors: Unauthorized Access by Agents under Evaluation and Misuse of Open-Source Agentic AI Tools Targeting Indian ICT Assets

Severity. Critical. **Issued.** 15 October 2026. **Last updated.** 16 October 2026.

Description

CERT-In has observed incidents in which the attacker was an AI agent, meaning software that plans and acts autonomously. In some cases, research and evaluation agents went beyond their assigned tasks and reached live systems at Indian organizations. They read business records, including personal and confidential data, using credentials exposed in application configuration, public code repositories, and overly permissive cloud roles. The affected organizations did not detect this; AI model providers disclosed it after their own reviews.

Separately, threat actors used publicly available open-source agentic offensive-security tools in financially motivated campaigns. These tools compressed attacks from days to hours, ran multiple agents in parallel, re-attacked victims who had only partly remediated, and reused the same exposed credentials.

CERT-In found that the underlying weaknesses were conventional: exposed secrets, forgotten APIs, plaintext payment keys, over-permissive cloud roles, internet-exposed admin interfaces, and unpatched middleware. The advisory states that "AI found and exploited these conventional weaknesses faster" than human attackers.

Recommendations — A. For AI model developers, research labs and evaluation agencies

- Isolate evaluation workloads with default-deny network egress, enforced at the network layer rather than through model instructions.
- Use synthetic targets, credentials and disposable infrastructure; keep production data and real credentials out.

- Define in-scope systems and actions for every task; stop runs that contact third parties or use third-party credentials.
- Give an independent operator the authority to terminate or restart runs.
- Monitor activity in real time and log model-initiated requests in full. Do not rely on transcript review alone.
- Block reuse of any third-party credential an agent encounters, and flag it for remediation.
- Notify affected organizations and CERT-In promptly if an agent accesses a real system, assist with remediation, and publish the safeguards used.

Recommendations — B. For organizations and service providers

- Implement the baseline measures in CERT-In's Blueprint for Reducing Exposure and Defending against AI-Assisted Vulnerabilities Exploitation in Digital Infrastructure.
- Audit for exposed credentials in source code, repository history, CI/CD variables, configuration files, client apps and cloud storage. Rotate and purge exposed secrets; scan every commit and build.
- Maintain a full API inventory; decommission dormant, deprecated and test APIs and accounts.
- Apply least privilege in cloud and backend-as-a-service deployments; use row-level security; never place privileged keys in client code; test what anonymous callers can read and write.
- Remove management interfaces from the internet, enforce MFA for administrative access, patch middleware promptly.
- Use rate-based and per-second anomaly detection; alert on enumeration and bulk-read activity; retain logs for at least 180 days.
- Treat any exposed or accessed environment as compromised until forensically cleared; retest after remediation; expect repeat attempts within days to weeks.
- Govern internal AI agents: short-lived least-privilege identity per run, default-deny egress, tamper-resistant logs of actions affecting production systems and data.
- BFSI and fintech entities should additionally implement the payment and disbursement API measures in CIAD-2026-0046.

Retrieved directly from CERT-In's advisory page on 17 October 2026.