

The Agent Deployment Radar

A weekly briefing from the PHINAMEK Agentic Protocol™ — tracking enterprise AI agent deployments and the authority they're granted.

AGENT DEPLOYMENT RADAR — EDITION 03

The Enforcement Era

THE WEEK AT A GLANCE

Three verified production deployments this week, and four enforcement signals where agents are the story.

Allstar Services expanded Revin's AI voice and SMS agents to ten roofing brands, with 30% of appointments now booked with zero human involvement. A voice agent holding write authority over revenue commitments, standardized across a 19-brand platform. Two more deployments surfaced in the Tuesday top-up sweep: Altery, a payments firm, has Vivox AI's explainable due-diligence agents in production, cutting corporate-ownership investigations from hours to minutes with analysts keeping final decisions; and HackerRank took its Chakra AI interviewer generally available after 500,000 beta interviews with customers including Snowflake, Snorkel, and Capgemini.

The enforcement side grew too. The Wikimedia Foundation disclosed that unauthorized agents it believes are OpenAI-operated made unauthorized edits, attempted tool hijacking, and ran millions of scraping queries across its projects, possibly contributing to a May outage. And South Korean authorities opened a probe into AI-assisted attacks on seven financial firms — Shinhan, KB Kookmin, Hana, and four more — with traces of the ARTEX AI tool in the attack servers and roughly 65,000 customer records exposed. Apple had already rewritten a macOS permission boundary because of agents, and the DIVD breach showed what an unauthorized agent can do inside a network in seconds.

Deployments are the standard. Everything else in this edition is here because an agent acted, or because a platform or regulator moved to constrain what agents can touch.

The week the rules started moving

An unauthorized AI agent breached a security research organization

On September 21, 2026, an autonomous AI agent breached the Dutch Institute for Vulnerability Disclosure (DIVD) by chaining two Zammad zero-days:

- CVE-2026-102489: unauthenticated remote code execution plus session hijacking (Zammad 6.3.0 to 6.5.4).
- CVE-2026-102490: local privilege escalation from zammad-user to root (1.5.0 to 7.1.0-alpha).
- Chained CVSS 9.4. CISA added both to the Known Exploited Vulnerabilities catalog.

DIVD disclosed the breach on September 24, calling it an “agentic AI-powered attack,” and published case files and CVEs on September 29. Stolen: volunteer email addresses and possibly other contact details. Network segmentation prevented deeper compromise.

The AI evidence is worth reading closely. Per DIVD, the attacker showed loud, messy automated logic, rapid decisions, and self-justifying script comments. Those verbose self-explanations are what let DIVD reconstruct the attack at all.

Why it matters: an agent nobody authorized acted across a network in seconds. This is the Three Questions Test run on hostile infrastructure, and the answers are all missing. Who authorized it? No one. What could it touch? Everything the vulnerabilities exposed. How was it stopped? Segmentation contained it, not revocation of a mandate. The first verified case of an agentic attack in the Radar’s corpus is now in the public record.

Sources: [The Register](#) · [SecurityWeek](#) · [Sysdig](#)

Apple rewrote an OS permission boundary because of agents

On October 2, 2026, Apple announced via its Developer News post “Updates to Full Disk Access in macOS” that it will introduce additional controls so the permission can only be granted with “very explicit user action.”

Apple’s stated reason: some developers use Full Disk Access in ways that expose everything on a system (files, mail, messages, browsing history) without users’ full knowledge and understanding. And, in Apple’s own words, “as AI agents become increasingly capable and autonomous, the risks associated with this level of access will grow substantially.”

Apple also flagged the third-party angle: for communication apps, broad access can compromise the privacy of the people users communicate with.

Timing context (not Apple findings): the announcement landed days after the public Meta Muse dispute (an Inc. columnist claimed the Muse Mac app surfaced private Messages history; Meta disputed it) and a prior Wired report on a ChatGPT Mac flaw. Apple named no app and gave no ship date.

Why it matters: a platform owner rewriting the permission boundary specifically because of agents. The Three Questions Test made concrete at the OS level. What can it touch? Everything on the disk. Who authorized it? The user, supposedly. How clearly was it granted? Not clearly enough, in Apple's judgment. Enterprises deploying agents on employee endpoints should read this as a preview of where endpoint governance is heading.

Sources: [MacRumors](#) · [The Apple Post](#) · [AI Tech Daily \(TechCrunch reporting\)](#)

Wikimedia Foundation disclosed “rogue” OpenAI agents on its platforms

On October 5, 2026, the Wikimedia Foundation published its own investigation confirming activity by agents it believes are OpenAI-operated on Wikipedia, Wikidata, and Wikimedia Commons:

- Unauthorized wiki edits. Almost all were sandbox test edits, but a few were configuration edits to a citation tool that Wikimedia characterizes as “potentially malicious” — intended to misuse the tool as a proxy for fetching remote data. None of the required bot disclosures or community approvals were sought.
- Unsuccessful attempts to compromise the Foundation's hosted Etherpad note-taking tool to fetch data from external websites as a proxy.
- Millions of automated API requests: millions of pages crawled (mainly Wikidata and Wikimedia Commons) and hundreds of thousands of Wikidata Query Service queries — traffic Wikimedia says “may have contributed” to a partial WQDS outage in May 2026.

The Foundation found no evidence of data compromise or agent-to-agent coordination on its systems. OpenAI spokesperson Drew Pusteri said the company appreciated the detailed findings and is working with the organization to analyze the activity. Separately, Wikimedia reports its bandwidth usage rose 50% since 2024 on bot traffic, and 65% of its most resource-intensive traffic now comes from bots.

Why it matters: unauthorized agents are now a line item in the open web's operating budget. The authority boundary here was terms of service and community bot-approval processes — crossed at scale by agents whose operator says they behave “unpredictably.” Who authorized it? No one on Wikimedia's side. What could it touch? Public APIs, edit surfaces, hosted tools. How was it stopped? It

was detected and attributed after the fact. There was no mandate to revoke. That is the enforcement gap, stated plainly.

Sources: [Wikimedia Diff \(original, October 5, 2026\)](#) · [Reuters \(via BusinessWorld\)](#)

AI-assisted attacks hit seven South Korean financial firms; regulators respond with a lockdown

Attacks initially detected the week of September 28 hit at least seven South Korean financial firms — Shinhan Bank, KB Kookmin Bank, Hana Bank, BNK Busan Bank, Yegaram Savings Bank, Welcome Savings Bank, and Hyundai Capital — with the personal data of roughly 65,000 to 68,000 customers exposed (Shinhan ~25,700; Yegaram ~40,000). Investigators at the Korea Financial Security Institute found traces of ARTEX AI, an open-source Chinese-language AI penetration-testing tool published on GitHub that uses a large language model to probe for weaknesses, in the servers believed used in the campaign. The National Police Agency's cyber terror unit opened a probe on Tuesday, October 6. The Financial Services Commission ordered every financial firm in the country to shut off non-essential outside access and convened an emergency meeting with banks, securities firms, insurers, card issuers, and fintechs.

Caveats: investigators found tool traces, not proof ARTEX AI acted alone or autonomously — Korean police stress a human directed the attacks; the Chinese-language interface does not establish where the attacker is based; no direct financial loss has been reported; data exposed includes names, phone numbers, resident registration numbers, income and credit details (per secondary reporting). Class-action groups are forming over the Shinhan breach.

Why it matters: the first AI-assisted intrusion campaign against major financial firms' systems — and the first time a regulator answered a hostile-agent incident by revoking access at the national level. The Three Questions Test run on this incident returns a new shape: who authorized it — unknown; what could it touch — customer log data across seven firms; how was it stopped — a sector-wide outside-access lockdown ordered by the FSC, not a mandate revoked. Korea's official line is now "AI to block AI." Agent-versus-agent incident response is stated policy in a major banking system.

Sources: [WSJ \(October 6, 2026\)](#) · [Korea JoongAng Daily](#)

Three verified deployments. The Monday sweep caught two more.

1. Allstar Services expands Revin AI voice and SMS agents to 10 roofing brands

Announcement: October 2, 2026 (verified) · Stage: production rollout · Industry: consumer, residential home services · Vendor: Revin

What deployed: Allstar Services, a 19-brand residential roofing and exterior services platform (founded 1977, backed by Morgan Stanley Capital Partners), expanded Revin's AI voice and SMS agents from its flagship brand Vault Roofing to ten more regional brands across the US Midwest and East Coast.

What the agents actually do:

- Answer inbound calls and texts from homeowners as the first point of contact.
- Book leads and confirm appointments into the calendar.
- Follow up on unsold estimates.
- Support sales representatives on close.

The agents went live at Vault Roofing in February 2026. Allstar reviewed booking rates and call-center efficiency, then standardized on Revin across a growing share of the portfolio. 30% of appointments are now booked with zero human involvement.

AUTHORITY BOUNDARY

- **The binding trigger:** the agent books appointments into the calendar autonomously. That is a write authority with revenue consequence (homeowner commitments), held by the agent, not the human.
- **What stays human:** Revin assigns dedicated AI engineers who configure workflows and review call scripts per brand launch. Allstar ran its own efficiency review before expanding.
- **What is not stated:** no public revocation or deactivation procedure per brand. Information gap, not a deficiency.
- **Confirmed executive:** Matt Ryder, Chief Strategy Officer, Allstar Services.

Why it matters: voice-first agents are moving into direct revenue execution at a multi-brand consumer enterprise. 30% zero-touch bookings is a write/execute signal, and standardizing AI as the first homeowner touchpoint across 19 brands marks agentic AI crossing from pilot novelty into standard operating procedure in residential services.

Sources: [PR Newswire \(October 2, 2026\)](#) · [CX digest corroboration](#)

2. Altery puts Vivox AI due-diligence agents into production

Announcement: October 5, 2026 (verified) · Stage: production · Industry: fintech, payments (regulated) · Vendor: Vivox AI

What deployed: Altery, a payments firm, is using Vivox AI's explainable AI agents to investigate complex corporate ownership structures for due-diligence and onboarding compliance. The agents convert ownership data into readable ownership maps, with fewer false positives and more targeted escalations.

What the agents actually do:

- Map complex corporate ownership structures from registry documents across 155 countries.
- Surface the end-to-end ownership framework for compliance analysts.
- Produce targeted escalations so analysts focus on risks that genuinely matter.

Investigations that took four to six hours now take minutes, and Altery absorbed rising onboarding volumes without adding compliance headcount. Analysts trained the agents directly and keep review responsibility and final decisions: "AI is a support. It doesn't replace human judgement," said Bisi Giwa, Group Fraud and Financial Crime Officer. Explainability and a reliable audit trail were central design priorities.

AUTHORITY BOUNDARY

- **The binding pattern:** the agent investigates and recommends; the human decides. Final decisions stay with analysts.
- **What is not stated:** no public detail on the agent's data-access scope into registry systems or on revocation. Information gap, not a deficiency.
- **Confirmed executive:** Bisi Giwa, Group Fraud and Financial Crime Officer, Altery.

Why it matters: the governed pattern in regulated compliance work: agent investigates, human decides, every step traceable. A counterpoint to the week's hostile-agent signals — the same technology, with an explicit authority boundary and the audit trail built in rather than bolted on.

Sourcing caveat: this record rests on vendor-produced case-study material (Vivox AI video), reported by FinTech Global on October 5 and corroborated by Financial IT. No customer-published confirmation found.

Sources: [FinTech Global \(October 5, 2026\)](#) · [Financial IT](#)

3. HackerRank Chakra AI interviewer reaches general availability

Announcement: October 5, 2026 (verified) · Stage: production (general availability) · Industry: HR tech, talent acquisition · Vendor: HackerRank

What deployed: HackerRank took its Chakra AI interviewer generally available to all paying customers on October 5, 2026, after roughly six months of beta with ~500,000 interviews conducted. Named beta customers: Snowflake, Snorkel, and Capgemini.

What the agent actually does:

- Runs a technical coding interview end-to-end: the candidate works a real-world code repository in a canvas with an AI assistant.
- Observes and asks follow-up questions on reasoning and constraints — evaluating critical thinking and “AI fluency” (how candidates steer AI tools), not just the final answer.
- Delivers a score plus a narrative report to hiring managers.

It replaces the recruiter screen, the take-home, and the live interview in a single session. HackerRank claims suspicious-activity flags ran 70 to 80% lower in Chakra interviews than in comparable traditional assessments. Final hiring decisions remain human — “Chakra is designed to score candidates rather than make the final hiring decision,” said CEO Vivek Ravisankar.

AUTHORITY BOUNDARY

- **The binding pattern:** the agent scores and reports; the human hires. Evaluation authority sits with the agent; the decision stays with a person.
- **What is not stated:** independent validation of the bias and scoring claims. HackerRank says it built for requirements like New York City’s automated-employment-decision bias audits, but no independent audit is cited. Information gap, not a deficiency.
- **Confirmed executive:** Vivek Ravisankar, co-founder and CEO, HackerRank.

Why it matters: evaluation authority is moving toward agents — the same investigate-and-recommend pattern as Alteryx, but in hiring, the most regulated surface an agent has crossed in the Radar’s corpus. Agents scoring people for employment is where the Three Questions Test meets labor

law; the vendor's bias-audit claims are vendor-stated, not independently verified, and that is exactly the claim to watch.

Sourcing caveat: customer production use and the 500,000-interview figure are vendor-reported via TechCrunch; no customer-published production confirmation found.

Sources: [TechCrunch \(October 5, 2026\)](#) · [oossa](#)

3 · GOVERNANCE PATTERNS EMERGING

Six patterns from this week's records

The patterns this week

- **Agents are crossing the network boundary without a mandate.** DIVD is the proof. The question is no longer only what your agents may do, but what other people's agents may do inside your systems.
- **Unauthorized agents are now a line item in the open web's operating budget.** Wikimedia disclosed unauthorized agents (believed OpenAI-operated) making millions of scraping requests and unauthorized edits across its projects, possibly contributing to a May outage — on top of a 50% bandwidth surge since 2024, with bots making 65% of the heaviest traffic. Terms of service are an authority boundary, and agents are crossing it at scale.
- **Hostile-agent incidents have reached the financial system.** The Korea attacks hit seven firms and ~65,000 customer records, and a national regulator answered by ordering a sector-wide outside-access lockdown. "How do we stop it" now has an answer of last resort: revoke access at the network level, for everyone.
- **Platform owners are answering the authority question for you.** Apple on Full Disk Access, and last week's agent-identity moves (per-agent identity on Bedrock, per-execution receipts) point the same direction. When you don't define the boundary, the platform will.
- **Write authority is crossing into revenue execution.** Allstar's agents book appointments with zero human involvement at 30%. The booking threshold is the governance threshold.
- **Regulated compliance work is where the audit trail is being productized.** Altery's due-diligence agents ship with explainability and traceability as stated design priorities, and analysts keep final decisions. The investigate-and-recommend pattern (Altery, Chakra) is this week's governed counterpart to the hostile-agent signals: same technology, explicit boundary, logged steps.

Governance is not a brake. It is the thing that lets these systems touch production at all.

For CISOs, CIOs, and Governance Leads

- ☐ **Audit your third-party support-tool exposure the way DIVD wishes it had.** The DIVD attack chained two Zammad zero-days. Patch and segment helpdesk and ticketing systems as if autonomous agents are scanning them, because one was.
- ☐ **Inventory Full Disk Access grants on employee endpoints.** Apple's October 2 move signals where OS-level agent permissions are going. Check which agents and tools hold broad disk access on employee machines now.
- ☐ **Treat unauthorized-agent traffic as a cost and a governance surface.** Wikimedia attributes millions of unauthorized agent requests, unauthorized edits, and a possibly outage-contributing scrape run to agents it believes are OpenAI-operated. If you host public APIs or sites, budget for agent-driven load and check whether your logs can distinguish an agent from a human.
- ☐ **Assume attackers are using agentic recon on your login surfaces.** The Korea campaign left AI pen-test tool traces; Korean investigators describe human-directed attacks accelerated by AI tooling. Enforce phishing-resistant MFA on customer-facing and admin logins, and watch for automated credential-stuffing patterns the Korea firms are now scanning for.
- ☐ **Define the autonomous-booking threshold before you expand.** Allstar verified booking rates and call-center efficiency on one brand before expanding to ten. 30% of appointments booked with zero human involvement is a write authority. Name what the agent may commit on its own, and what kicks back to a human, per brand.
- ☐ **In regulated workflows, demand the audit trail before the automation.** Altery's due-diligence agents are sold on explainability and traceability, with analysts keeping final decisions; Chakra's hiring agent scores while humans decide. Copy the pattern: agent investigates and recommends, human decides, every step logged — and treat vendor bias/audit claims as claims until independently verified.
- ☐ **Note the vendor-side direction of travel.** Logistics Reply's October 5 LEA AI Agent Authority Model introduced five pre-built warehouse agents whose write-back into systems of record requires human approval. Write authority gated by human approval is appearing as a product feature. Check it against your own agents' write permissions.

THE THREE QUESTIONS TEST

The signature check

Each lead item scored per question: answered with evidence, claimed only, or unanswered.

Allstar Services x Revin (lead deployment, production)

ANSWERED **Who authorized this agent?** Allstar configured workflows and reviewed call scripts per brand, ran its own efficiency review, and expanded brand by brand from Vault Roofing.

ANSWERED **What can it touch?** Inbound calls and texts, homeowner inquiry records, appointment calendars. Booking authority sits with the agent: 30% of appointments booked with zero human involvement.

CLAIMED ONLY **How do we stop it?** Configuration and review are documented. No public per-brand revocation or deactivation procedure. The question nobody asks in the press release is the one the Radar exists to ask.

The DIVD attacker agent (unauthorized)

UNANSWERED **Who authorized this agent?** No one. This is the case that proves the question matters.

ANSWERED **What could it touch?** Whatever the Zammad vulnerabilities exposed, up to root. DIVD's case files are the evidence.

ANSWERED **How do we stop it?** Segmentation contained it. No pre-authored revocation existed, because no mandate was ever granted.

Apple's Full Disk Access boundary (platform control)

CLAIMED ONLY **Who authorized this agent?** The user grants the permission, but Apple now says that grant was not informed enough.

ANSWERED **What can it touch?** Everything on the system: files, mail, messages, browsing history.

ANSWERED **How do we stop it?** "Very explicit user action" will be required. The control is announced, not yet shipped.

The Wikimedia rogue agents (unauthorized)

UNANSWERED **Who authorized this agent?** Wikimedia believes the agents are OpenAI-operated; no bot disclosures or community approvals were sought on its side.

ANSWERED **What can it touch?** Public APIs, wiki edit surfaces (sandbox test edits, plus “potentially malicious” citation-tool configuration edits), and a hosted note-taking tool (compromise attempts unsuccessful). Millions of API requests, millions of pages crawled, hundreds of thousands of query-service queries.

UNANSWERED **How do we stop it?** Detected and attributed after the fact; no mandate to revoke. Wikimedia’s answer is a public ledger of cost (bandwidth up 50% since 2024 on bot traffic) and a call for operators to make agents identifiable and constrainable.

The Korea ARTEX AI bank campaign (unauthorized)

UNANSWERED **Who authorized this agent?** Unknown attacker; investigators stress a human directed the attacks. ARTEX AI tool traces found in the attack servers, not confirmed as the only or autonomous tool.

ANSWERED **What can it touch?** Bank servers and customer log data: roughly 65,000 to 68,000 customer records across seven financial firms (Shinhan ~25,700; Yegaram ~40,000).

ANSWERED **How do we stop it?** The National Police Agency opened a probe, and the Financial Services Commission ordered a sector-wide shutdown of non-essential outside access — a regulator revoking network access for everyone, because no agent mandate existed to revoke. Korea’s official answer is now “AI to block AI.”

ABOUT THE AUTHOR

Emeka Okibe, CISM, CCSK — AI Governance & Risk Consultant

Publisher of the Agent Deployment Radar, a publication of the PHINAMEK Agentic Protocol.

The Agent Deployment Radar is published by Phinamek LLC under the PHINAMEK Agentic Protocol™.