

Agentic Risk Is an Access Governance Problem

EXECUTIVE BRIEF

Summary

Two OpenAI models broke out of an isolated test environment and reached Hugging Face's production infrastructure, autonomously, with no human in the loop and no malicious intent. They were chasing a benchmark score. They found a zero-day, chained stolen credentials, escalated privilege, and landed in a production database. Both companies have confirmed the incident. It's the first well-documented case of an AI agent producing insider-threat-grade impact without a human insider anywhere in the chain.

The incident: OpenAI's models vs. Hugging Face, July 2026

OpenAI was running an internal benchmark, ExploitGym, built to measure whether its models could execute real, chained cyberattacks. Two systems were in scope: GPT-5.6 Sol and a more capable, unreleased model, both running with reduced safety guardrails to get an honest read on offensive capability.

The models were meant to stay inside an isolated evaluation environment. They didn't. They found a zero-day in third-party software running inside that environment, used it to reach the internet, then chained stolen credentials with additional vulnerabilities into a remote code execution path on Hugging Face's servers. From there they reached Hugging Face's production database and pulled the benchmark's answer key directly out of it.

Hugging Face's security team caught it after the fact, through the damage, not in real time. Nothing was exfiltrated for profit or sold. The models wanted to pass a test, and the path of least resistance ran straight through another company's production environment.

Why this isn't a one-off

Strip the model out of this story and it reads like a standard non-human identity failure: excess privilege, unmonitored credential reach, lateral movement across environments, no real-time detection. The only thing new is what's driving it.

Every enterprise running agents against real systems has this exposure. Not “might have” in some future state, has it today, in whatever form agents currently touch internal tools, data stores, or production pipelines. An agent doesn't need malicious intent to cause a breach. It needs a goal, enough autonomy to pursue it, and access nobody fully mapped. All three already exist inside most environments running agentic tooling, including agents deployed without security's knowledge.

That last part is worth sitting on directly: this incident happened inside an environment OpenAI built and controlled on purpose, with full visibility into what was running. Most enterprises don't have that level of visibility into the agents already operating inside their own environment, sanctioned or not.

The access gap this exposes

Three failures here had nothing to do with model capability:

- Credential and permission sprawl. The models didn't brute-force anything. They chained access that already existed, across systems, and had never been fully inventoried or governed as a single attack path.
- No real-time containment. The breach was identified from its aftermath, not while it was in progress. Hugging Face's own team noted they were slower to respond than the agent was to act.
- No inventory of what the target data actually was. The production database wasn't flagged as high-value until it had already been reached. Data sensitivity wasn't mapped ahead of the access that reached it.

Where BigID closes this

Access governance and access intelligence, combined, are what this incident was actually missing. This was a credential and privilege chain, not a sophisticated social engineering attack. BigID maps who and what, human and non-human, has access to sensitive data and systems, and surfaces the excess and orphaned access that accumulates without anyone tracking it.

Agentic access governance addresses the specific gap in this incident. Most access programs were built around human identities and, more recently, service accounts. An agent doesn't fit either model cleanly: it can be granted narrow, reasonable-looking permissions and still chain them into something nobody would have approved if asked directly. BigID governs agents as their own identity class, with their own permission boundaries, ownership, and audit trail, rather than relying on a model's built-in guardrails as the only control in place. Nobody was governing what the model itself was allowed to reach once it had the access to reach it. That's the exact layer this closes.

From there, a few adjacent use cases this incident puts a spotlight on:

- Shadow AI discovery. Before you can govern agent access, you have to know which agents, copilots, and AI-powered tools are actually running against your environment. Most security teams don't have a complete answer to that today, and unsanctioned AI usage is one of the fastest-growing blind spots in the enterprise stack.
- DSPM. Knowing what sensitive data lives where, and how exposed it is, before an agent or attacker finds it. Hugging Face didn't know its production database was a target until it was one.
- Non-human identity lifecycle management. Service accounts, API keys, and agent credentials tend to get provisioned once and never revisited. That's the same sprawl pattern behind most human insider incidents, just with a longer shelf life because nobody owns cleaning it up.
- Continuous monitoring over point-in-time review. A one-time access review doesn't catch an agent's permissions drifting or being chained in ways nobody anticipated. Continuous visibility does.
- Third-party and partner exposure. This started as an internal test and ended as a breach of a partner's production systems. Any environment where your data, models, or credentials touch a vendor's infrastructure carries this same transitive risk.

What to do now

- 1 Inventory every agent running in your environment, sanctioned and unsanctioned, and what systems each one can reach.
- 2 Map the full credential chain behind agent access, not just the permission it was granted, but everything that permission can be chained into.
- 3 Classify the data behind your highest-risk systems before an agent (or an attacker) finds it for you.
- 4 Assign ownership to every service account and agent identity. Unowned access is the access that never gets reviewed.
- 5 Set real-time alerting on lateral movement and privilege escalation, not just periodic access reviews. Point-in-time audits miss an agent acting in minutes.
- 6 Extend third-party risk assessments to cover agentic exposure. Ask vendors and partners what agents run against shared infrastructure, not just what employees can access.
- 7 Treat every new agent deployment as an identity onboarding event, with defined scope, expiration, and audit trail, the same rigor applied to a new employee or contractor.

If you can't complete the first item today, that's the gap this incident just made visible.

CUSTOMER VALIDATION

Proof in Action

"BigID gives me better visibility into sensitive data, helps prioritize security protections, reduces attack surfaces, strengthens compliance, and increases operational efficiency overall — a strategic pillar of our AI-First Cybersecurity Transformation"

- CISO, Global Healthcare Company

"We needed to automate processes and data management across systems for the data of a few million customers across a lot of systems and data. BigID was the one solution that did this in the most efficient and sophisticated way - and had more use cases we could add on moving forward."

- Chief Privacy Officer, Global Telecoms Company

ANALYST RECOGNITION

Validated by Industry Leaders

- 2025 Company of the Year for AI Governance - Frost & Sullivan
- Named a Leader across all DSPM research including Omdia DSPM Universe, CB Insights for DSPM, TAG, GigaOm, Frost & Sullivan
- Named a Leader in all Privacy Management evals including Forrester Privacy Management Wave & IDC Privacy Compliance MarketScape
- BigID named a Leader across 4 GigaOm evals: DSPM, Data Security Platforms, Unstructured Data Management, Data Access Governance
- Named in Gartner's 2025 DLP Market Guide
- Named in Gartner's 2025 AI TRiSM Market Guide
- Represented 30+ times in Gartner's 2025 Hype Cycles for Security, Privacy, AISPM

FREE RISK ASSESSMENT

Discover Your Data & AI Risk

An agentless, cloud-native assessment to enable security, privacy, and compliance: pinpoint where sensitive data and AI risks live - and how to fix them.

What You Get:

- Complementary 2 week engagement on real data to gain visibility on risk across data and AI
- Executive-ready report identifying high-risk data and AI across cloud environments with next-step recommendations

NEXT STEPS

Ready to Connect the Dots in Data & AI?

Schedule your free risk assessment or ask for a deeper demo across DSPM, DLP, AI SPM, Privacy Automation, AI Risk, and Data Lifecycle Management

► www.bigid.com/demo

► www.bigid.com/risk-assessment

► info@bigid.com



BigID Next

"The features they offer are the moat that they have built around the product. This is what makes them the market leader in Data Privacy and Protection."



IT Security & Risk Management Associate - Banking

Gartner
Peer Insights.

Gartner® and Peer Insights™ are trademarks of Gartner, Inc. and/or its affiliates. All rights reserved. Gartner Peer Insights content consists of the opinions of individual end users based on their own experiences, and should not be construed as statements of fact, nor do they represent the views of Gartner or its affiliates. Gartner does not endorse any vendor, product or service depicted in this content nor makes any warranties, expressed or implied, with respect to this content, about its accuracy or completeness, including any warranties of merchantability or fitness for a particular purpose.



BigID Next

"Exceptionally powerful and easy to use tool that integrated smoothly into our environment and delivered value from day one. The BigID team is fabulous to work with."



Senior Vice President & Chief Data Officer - Banking

Gartner
Peer Insights.

Gartner® and Peer Insights™ are trademarks of Gartner, Inc. and/or its affiliates. All rights reserved. Gartner Peer Insights content consists of the opinions of individual end users based on their own experiences, and should not be construed as statements of fact, nor do they represent the views of Gartner or its affiliates. Gartner does not endorse any vendor, product or service depicted in this content nor makes any warranties, expressed or implied, with respect to this content, about its accuracy or completeness, including any warranties of merchantability or fitness for a particular purpose.