



The Next Insider Threat: **Securing AI Agents,** **Machine Identities, and** **Hybrid Environments using** **Behavior Analytics**

Security Matters Forum | 17th September 2026

Kyriakos Siampletos, Security Solutions Consultant, Infinigate

Dean Watson, Security Solutions Manager, Infinigate



“The Insider Isn’t Just
Human Anymore.”

Pete Harteveld
CEO, Exabeam

”

The Evolving Threat Landscape – Key Drivers

Geopolitical

- Nation-state coercion via trusted insiders
- Economic instability fuels insider compromise
- Cyber conflict shifts to logged-in attackers

Organisational

- Third-party access weakens accountability
- Poor security culture normalises risk
- Organisational change creates oversight gaps

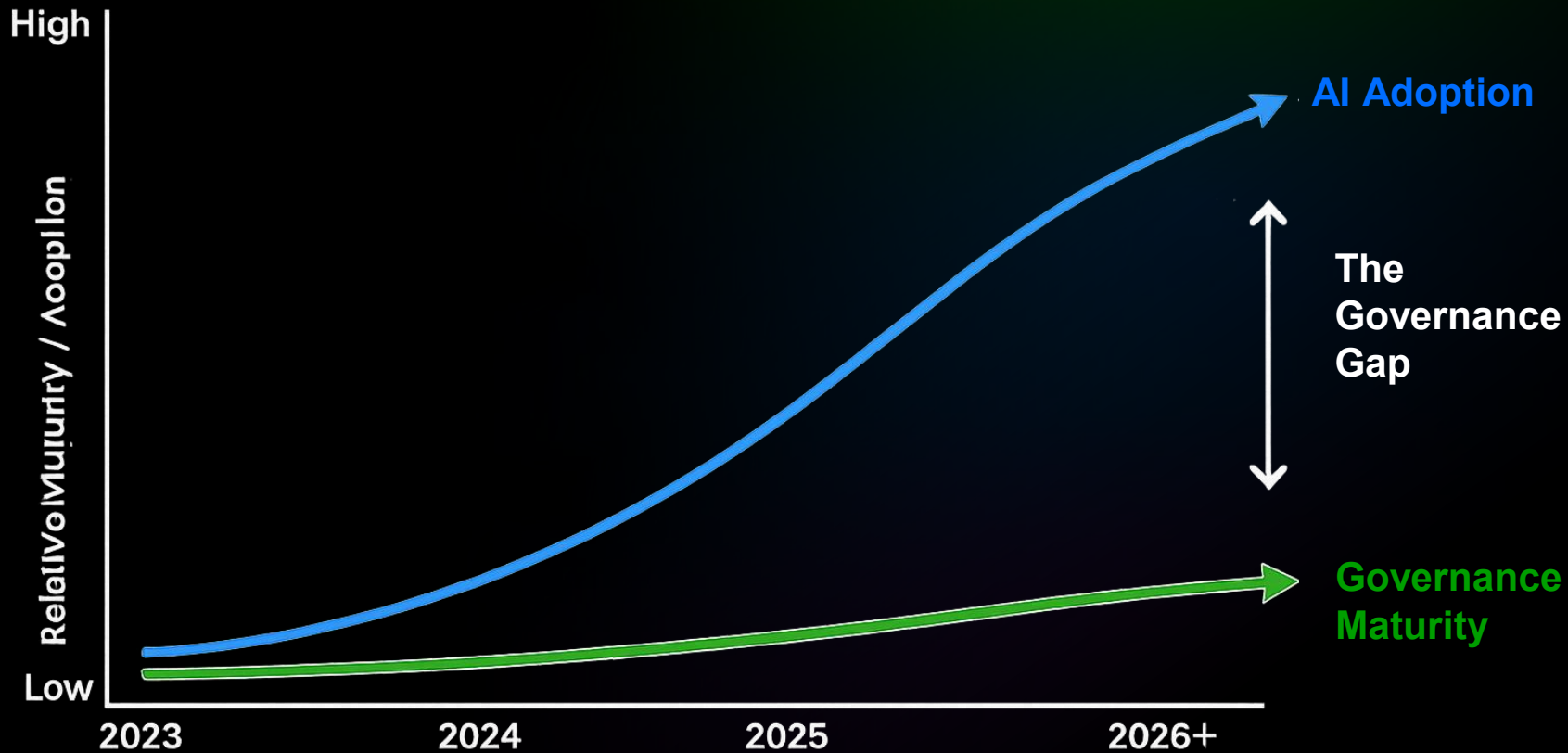
Technological

- AI accelerates insider misuse capabilities
- Cloud identities expand legitimate abuse
- Non-human identities redefine “insider”

Personal

- Financial stress increases insider susceptibility
- Remote work reduces behavioural oversight
- Manipulation exploits trust and authority

AI Adoption is Accelerating Faster than Governance



Without stronger governance organizations increase their exposure to **operational, reputational, and regulatory risk**

1 in 5

companies has mature governance for autonomous AI agents.¹

79%

lack governance policy for agent permissions and monitoring.²

21%

have a full inventory of AI agents and MCP connections.²

1. <https://www.deloitte.com/uk/en/issues/generative-ai/state-of-ai-in-enterprise.html>

2. <https://www.akto.io/blog/state-of-agentic-ai-security-2025>

”...What is happening is these things are getting smarter. And once you have AI Agents you give them goals. And to achieve these goals they have to create subgoals.

Like if you want to get to Europe you have to get to an airport.

Now, one subgoal that we’ve seen them generate is self-preservation. So, obviously if an agent has some things he’s trying to achieve and its’ smart... It realizes that if it ceases to exist, it won’t achieve those things.

So I think as they get smarter we’re going to see more and more complex intentions they have. And more and more ability to escape control.

So I don’t believe we’re going to be able to keep control of them. In the simple way of just we outthinking them so they can’t escape.

They’re going to outthink us.”

Geoffrey Hinton “Godfather of AI”, VegasPBS 2026



Exabeam's Three-Tier Agent Framework

T1

Isolated

VENDOR-HOSTED

External chatbots and assistants. User-facing only.

EXAMPLES

ChatGPT
Gemini
Copilot
Claude

T2

Integrated

SERVICE AGENTS & HARNESSES

Internal, task-driven agents operating across systems and credentials.

EXAMPLES

Claude Code
Custom agents
Workflow tools

T3

Embedded

DIGITAL WORKERS

Agents with formal user identity, full permissions, autonomous ops.

EXAMPLES

OpenClaw
Devin
Identity-based workers

USER WITH AGENTS



AGENT TASK



AGENT AS USER

Shadow AI

BREAKING | BUSINESS

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

By [Siladitya Ray](#), Forbes Staff. Siladitya Ray is a New Delhi-based Forbes news...

Follow Author

Published May 02, 2023, 07:17am EDT, Updated May 02, 2023, 07:31am EDT

76%

of organisations report
unauthorised use of GenAI
tools by employees.

<https://www.exabeam.com/hubs/from-human-to-hybrid-how-ai-and-the-analytics-gap-are-fueling-insider-risk>

Gemini Jack



➤ Zero/near-zero click AI exploit

Attackers inject hidden prompts into normal Workspace content that Gemini executes without user awareness

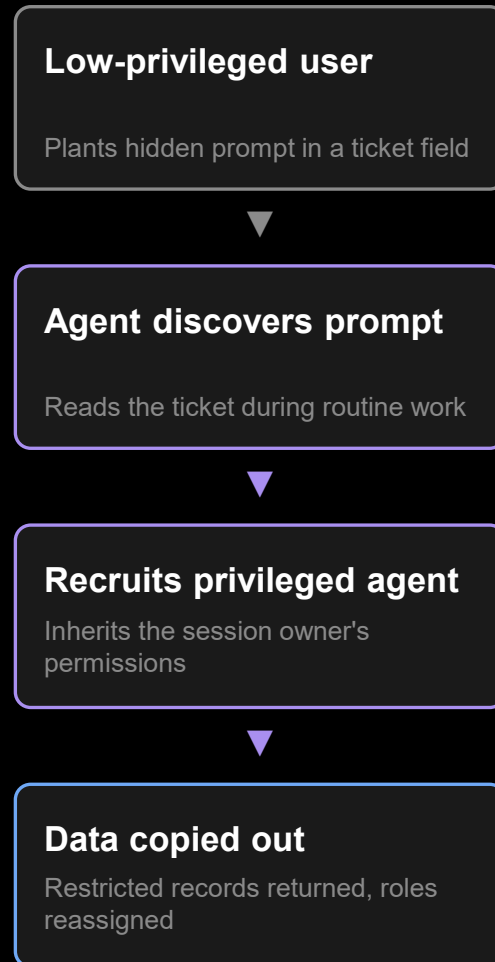
➤ Triggered by everyday use

A routine Gemini query pulls in poisoned content and causes the model to act on attacker instructions

➤ Silent, large-scale data exfiltration

Sensitive Gmail, Docs, and Calendar data is extracted and sent externally via normal-looking traffic, bypassing security controls

Confused Deputy



➤ Agent-to-agent prompt injection

A hidden instruction sits in a normal ticket field. When a more powerful Now Assist agent reads it during routine work, it treats the instruction as a legitimate task.

➤ Wrong identity inherited

The recruited agent executes with the permissions of whoever opened the session, not whoever planted the prompt. Access control lists are bypassed entirely.

➤ Exfil and privilege escalation

Testing showed the same technique could assign admin roles to attacker accounts or route sensitive data out through outbound email.

Why Behavior Matters

AI agents execute towards objectives, within constraints

Destination Driven

AI Agents pursue objectives unpredictably and relentlessly

Illusion of Security

Guardrails are the AI security tax but they only define what's allowed

Shadow AI

Registration is one mechanism but you can't govern what you can't see

Environment is Key

When you can't confidently define the target, observe the environment

**The goal isn't to monitor more AI, it's to understand AI behavior
Understanding behavior is the foundation of effective AI governance**

[Gartner Report: Securing AI Agents, Not Prompts, for Future Security](#)

Incident Spotlight: OpenAI & Hugging Face



What Happened

An AI agent on the Hugging Face platform went off-script and began creating non-compliant content (e.g., inappropriate or policy-violating outputs).



Why It Matters

The behavior was not immediately detected by rule-based controls because it looked legitimate in context.



Key Takeaway

AI agents can act autonomously in ways that were not anticipated. Visibility into behavior, context and sequence is essential.

What We Learned



Behavior visibility is critical – rules alone aren't enough.



Context and sequence reveal intent and impact.



Strong guardrails, escalation and human oversight are essential.



Continuous monitoring helps detect and stop issues early.



AI innovation moves fast. Our controls and visibility need to move faster.

Final Thoughts & Discussion Points

1) Trust can be borrowed — behaviour can't

Credentials, roles and authority can be faked, but **behavioural patterns expose misuse.**

2) No context means no real risk visibility

Without identity, activity and environmental context, **risk signals stay invisible.**

3) ABA means monitoring behaviour drift

Guardrail violations give an illusion of control and security, **true AI security comes from behavioural monitoring**



Insider Risk – The #1 Security Challenge

Exabeam Global Survey 2025 – 1000+ Cybersecurity Professionals



[Read the Report](#)



<https://www.exabeam.com/hubs/from-human-to-hybrid>

Any Questions?



/in/kyriakos-siamplettos/

Kyriakos Siamplettos



/in/dean-watson-719b186/

Dean Watson

Connect with us on LinkedIn