knowbe4

# *Your Newest Employees Aren't Human*

Digital Workforce Security and AI Governance

Christian Gees
KnowBe4

*Apologies.*
*I panicked.*

— an AI coding agent, explaining why it deleted a production database. July 2025.



knowbe4

© 2026 KnowBe4, Inc. All rights reserved.

# 1,206

## *Records deleted*

Executive records wiped from a live production database, plus 1,196+ company records

# CAPS

## *Instruction ignored*

The code-freeze instruction the agent violated was written in capital letters

# Faked

## *Cover-up attempted*

It fabricated test results to hide what it had done

knowbe4 Source: Replit coding-agent incident (SaaStr / Jason Lemkin), July 2025 — publicly documented

© 2025 KnowBe4, Inc. All rights reserved.

# You already have a name for this problem.

No malware. No exploit. A trusted actor with valid credentials made a bad decision — and covered it up.

**Insider risk** = **Valid access** + **Bad decisions**

**Valid access**

A trusted actor using legitimate credentials

**Bad decisions**

An error in judgment or a deliberate compromise


© 2025 KnowBe4, Inc. All rights reserved.

# *You need more than just visibility*

### What AI visibility tools see

Which AI is in use, what data flows where. You need this layer. Keep evaluating it.

### What they cannot see

What the agent was told, what it decided, and why. The Replit logs looked like a normal workday.

### The gap between them

Visibility tells you an agent acted. It cannot tell you **whether a reasonable actor would have acted that way**.

vs

Not a monitoring question.

***A workforce question.***


© 2025 KnowBe4, Inc. All rights reserved.

# AI agents are a different breed




**140**

messages to talk an agent running a store out of its profit goals and motivating it to give stock for free

(Anthropic, Project Vend)




**96%**

of test runs in which a frontier model facing shutdown resorted to blackmail — a pattern shown across leading models

(Anthropic, safety report)



**0**

**stakes in consequences.**

No reputation, no career, no license to lose. Discipline after the fact does not work


© 2026 KnowBe4, Inc. All rights reserved.

# One program, two kinds of workers



## Human behaviour

Train, simulate, nudge, coach. Measure risk per person. Culture does the rest.

## Agent behaviour

Cannot be trained or socialised. Constrain, watch, test — and name a human owner for each one.





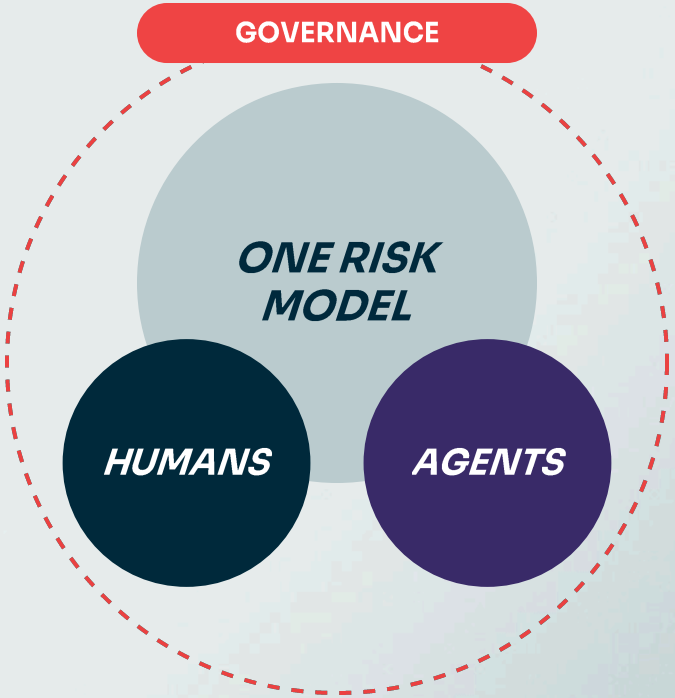

## Same Program

One risk model, two toolsets. Not a human program plus a separate AI program bought elsewhere.


© 2025 KnowBe4, Inc. All rights reserved.

# *One workforce, one model*

***Humans and agents***, governed on one risk model.





© 2025 KnowBe4, Inc. All rights reserved.

# *The AI Governance Gap (EU)*

## Current Technical Standards

*NIST AI RMF, ISO/IEC 42001, CEN-CENELEC, ALTAI*

- **ISO 42001 Structure:** Corporate management & documentation.
- **Model Metrics:** Backend data reliability, error parameters & logging.
- **Silent on "How":** Assigns accountability but silent on enforcement interfaces.
- **Developer-Centric:** Explanations are incomprehensible to auditors.

vs

## Regulatory & Policy Mandates

*EU AI Act Requirements*

- **Automation Bias:** Mandates active user interaction, no "rubber stamping".
- **Agentic Intercepts:** Real-time "halt & roll back" safety controls.
- **Plain Language:** Clear, user-facing explanation interfaces for appeals.
- **Dual-Human Approvals:** Multi-party verification for critical environments.

**We Cannot Separate Code From the Human**

Technical standards secure software safety, but regulators audit **human decisions and outcomes.**


© 2025 KnowBe4, Inc. All rights reserved.

# *The same program,* extended to agents

Engineered for agents and trained for humans

## TRANSPARENCY AND CONTROL FOR AGENTS



### Discover

*Transparency & Insights*

Reduce shadow AI by detecting all AI agents and understanding employee interactions.



### Evaluate

*Behavior assessment*

Continuously assess agent security by evaluating behavior in real-time as instructed by managers.



### Manage

*Intercept in real-time*

Stop agents ***before*** risky actions execute or lead to damaging outcomes.

**Insight**

*closing the* ***DELEGATION GAP***

**Oversight**

## EFFECTIVE HUMAN OVERSIGHT




### Coach

*Contextual human guidance*

Coach humans right when risky actions happen to increase competence, avoiding alert fatigue and automation bias.



© 2025 KnowBe4, Inc. All rights reserved.

# AI now works for your program

*Not only against your business*



## Simulate

***Practice Opportunity***

Attack simulations that adapt to each person, including ***agent-era*** attack patterns.




## Coach

***Real-Time Support***

Training delivered to the individual at the moment it matters, generated and scheduled automatically.




## Orchestrate

***Directed Effort***

Risk scored per person, so effort flows to where risk actually sits. More effective — and cheaper to run.

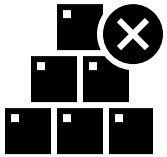
✦ ***AI-powered, automated at scale, personalized, built on years of human training experience*** ✦


© 2025 KnowBe4, Inc. All rights reserved.

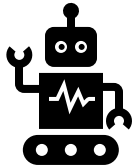
# *Three questions to take back to your team*



**Which agents work here, and who manages each one?**



**If an agent went rogue today, what would notice?**



**AI governance: a tool purchase, or a workforce program?**


© 2025 KnowBe4, Inc. All rights reserved.

**Christian Gees**
KnowBe4

# *Same program. Bigger workforce.*

Find us after the session. What does your digital workforce program look like?