knowbe4

Securing The Hybrid Workforce: Protecting Humans and AI Agents in a New Era

WWW.KNOWBE4.COM

KnowBe4, Inc. | 33 N Garden Ave, Suite 1200, Clearwater, FL 33755
Tel: 855-KNOWBE4 (566-9234) | Email: Sales@KnowBe4.com

CR-0663 0326

Introduction

The workforce has changed.

Employees are no longer working alone. They are increasingly collaborating with AI copilots, assistants and autonomous agents that help write code, summarize incidents, draft communications, analyze data and answer customer questions. What began as simple productivity tools has quickly evolved into a new class of digital colleagues embedded across enterprise workflows.

This shift is accelerating rapidly. Goldman Sachs estimates that agentic AI could represent roughly 60% of software market value by 2030, signaling a fundamental change in how work gets done. The modern workforce is no longer purely human, it is a hybrid team of people and AI agents working together to drive productivity and decision-making.

But this transformation introduces a new security challenge that many organizations are not prepared to address.

For decades, cybersecurity teams have focused on protecting systems, networks and endpoints while also reducing human risk through security awareness, culture-building and training. These efforts were designed to defend against phishing, social engineering and other forms of cognitive manipulation.

Today, that risk surface has expanded. Humans are no longer the only actors interacting with sensitive data, generating content or making decisions inside the enterprise. AI agents now perform many of these same tasks — often with speed and confidence, but without a full understanding of organizational policy, context or risk tolerance.

Unlike employees, AI agents cannot be improved through traditional training or awareness programs. Organizations cannot fully inspect or control the prompts shaping responses, yet these agents are already influencing business outcomes.

This creates a familiar challenge in a new form. Security teams must manage how humans interact with the technology they use. In the AI era, risk increasingly resides in the interaction layer between humans and intelligent systems.

A useful way to think about an AI agent is as a highly capable but highly impressionable intern. It processes vast amounts of information and executes instructions quickly and confidently. It is eager to help, but it lacks judgment, context and an inherent understanding of organizational risk.

Without the right guardrails, even the most capable assistant can do things you wouldn't expect.

The Emerging Human-AI Attack Surface

AI agents are incredibly powerful and inherently vulnerable. Attackers are already learning how to exploit the dynamic between humans and AI. In many cases, the attack does not target the human or the AI alone. Instead, adversaries exploit the trust relationship between them.

The introduction of AI into everyday workflows has created a new category of security risk. Threat actors are using AI to scale familiar attacks, generating more convincing phishing emails, producing deepfake content for social engineering and automating reconnaissance. At the same time, they are developing techniques specifically designed to manipulate AI systems:

- **Prompt Injection Attacks**
  Malicious inputs designed to influence an AI agent's behavior, potentially causing it to bypass controls, reveal sensitive data or perform unintended actions.
- **AI Agent Impersonation**
  Rogue agents that mimic legitimate enterprise tools to collect credentials, sensitive information or workflow data from unsuspecting employees.
- **Human–AI Social Engineering**
  Attacks that exploit trust in AI-generated outputs, potentially turning compromised agents into insider threats.

In many cases, no malware is deployed and no credentials are stolen. The employee simply issues an instruction, and the AI agent executes it as designed. Both the human and the AI behave as expected, which is precisely where the vulnerability exists.

The Trust Problem

Humans naturally trust tools that consistently help them succeed. When AI systems produce clear summaries, useful recommendations or polished content, users quickly develop confidence in their outputs. Over time, this can create automation bias — the tendency to trust automated decisions even when they may be incorrect or manipulated.

Attackers can exploit this trust. A malicious prompt hidden within a document, dataset or webpage can influence an AI-generated response, presenting manipulated information in a confident and authoritative tone. To the employee, the output appears credible, even though it has been shaped by an attacker.

Traditional security architectures were built to protect systems, networks and endpoints. While these controls remain essential, they do not fully address the risks created by human–AI collaboration. The most significant gap now exists at the interaction layer, where employees rely on AI to interpret information, generate content and support decisions.

Securing this environment requires protecting both sides of the modern workforce. Organizations must empower employees to recognize manipulation attempts and safely use AI tools, while also ensuring AI agents are resilient against prompt injection, data exposure and unauthorized actions.

The Future of Security Is Dual Defense

The boundary between human and AI in cybersecurity is disappearing. As AI agents become embedded in daily workflows, the organizations that adapt their security strategies to this reality will be the ones that remain resilient. The message is clear: cybersecurity is no longer just about protecting systems from humans or humans from systems. It is about securing the interaction between them, because within that interaction lies both the greatest vulnerability and the strongest defense.

To meet this challenge, organizations must adopt a dual defense strategy that strengthens both the human element and the AI systems themselves.

1. Strengthen Human Resilience

In an AI-augmented workplace, the role of the employee is shifting from task execution to oversight and judgment. Security awareness training must evolve accordingly. It is no longer enough to teach users how to spot a phishing email. Employees must develop digital mindfulness — a combination of healthy skepticism, contextual awareness and the ability to critically evaluate AI-generated outputs.

This starts with education. Employees need to understand both the capabilities and the limitations of AI agents, including how they can be manipulated. They should be trained to recognize anomalous or unexpected AI behavior, such as outputs that conflict with policy or context. Just as importantly, organizations must establish simple but effective verification protocols for high-risk actions, especially those initiated or assisted by AI. When something feels urgent, unusual or out of scope, employees should know how to pause, validate and escalate.

2. Harden the AI Agent

At the same time, AI systems must be treated as a new class of enterprise asset that requires governance, monitoring and control. This begins with implementing strong input validation to prevent malicious prompts or poisoned data from influencing agent behavior. AI outputs should also be continuously analyzed to detect anomalies, policy violations or signs of manipulation.

Equally important is defining clear role boundaries for AI agents. Systems should be designed to operate within strict scopes, refusing requests that fall outside authorized tasks. This must be reinforced by well-defined AI usage policies that govern how agents interact with data, systems and users.

Critical Capabilities

Managing risk in this hybrid environment requires visibility into how AI agents operate, how they interact with users and how they access data and systems. To provide complete coverage from discovery to defense, effective AI agent security products should be built around four core pillars:

Discover
Automatically identify and catalog every AI agent operating within your environment — including agents introduced through sanctioned platforms and those deployed without formal approval. Continuous discovery helps eliminate blind spots and ensures security teams understand the full scope of AI usage.

Monitor
Capture every agent invocation, prompt interaction and tool call in a comprehensive, searchable audit trail. Rich telemetry, including metadata from AI digital assistants such as Microsoft Copilot, enables teams to understand how agents are being used and where potential risk may emerge.

Detect
Apply advanced detection capabilities to identify risky behavior in real time. This includes identifying prompt injection attempts, privilege escalation, sensitive data exposure and anomalous agent activity across languages and use cases. Behavioral analysis helps surface both known and emerging threats that traditional controls may miss.

Protect
Enable organizations to define and enforce their desired security posture. Teams can choose to passively monitor activity, generate alerts or actively intercept and prevent high-risk operations in real time. Flexible enforcement allows security teams to balance productivity and protection as AI adoption evolves.

Lastly, for IT and security professionals, scalability is essential. Managing risk across a growing population of human users and AI agents requires centralized visibility and operational efficiency. Key capabilities should include:

Visibility and Auditing
Centralized dashboards provide a clear view of human-to-agent and agent-to-system interactions. Detailed audit logs enable security teams to detect anomalous behavior, investigate incidents, and identify activity that falls outside of defined policies or intended use cases.

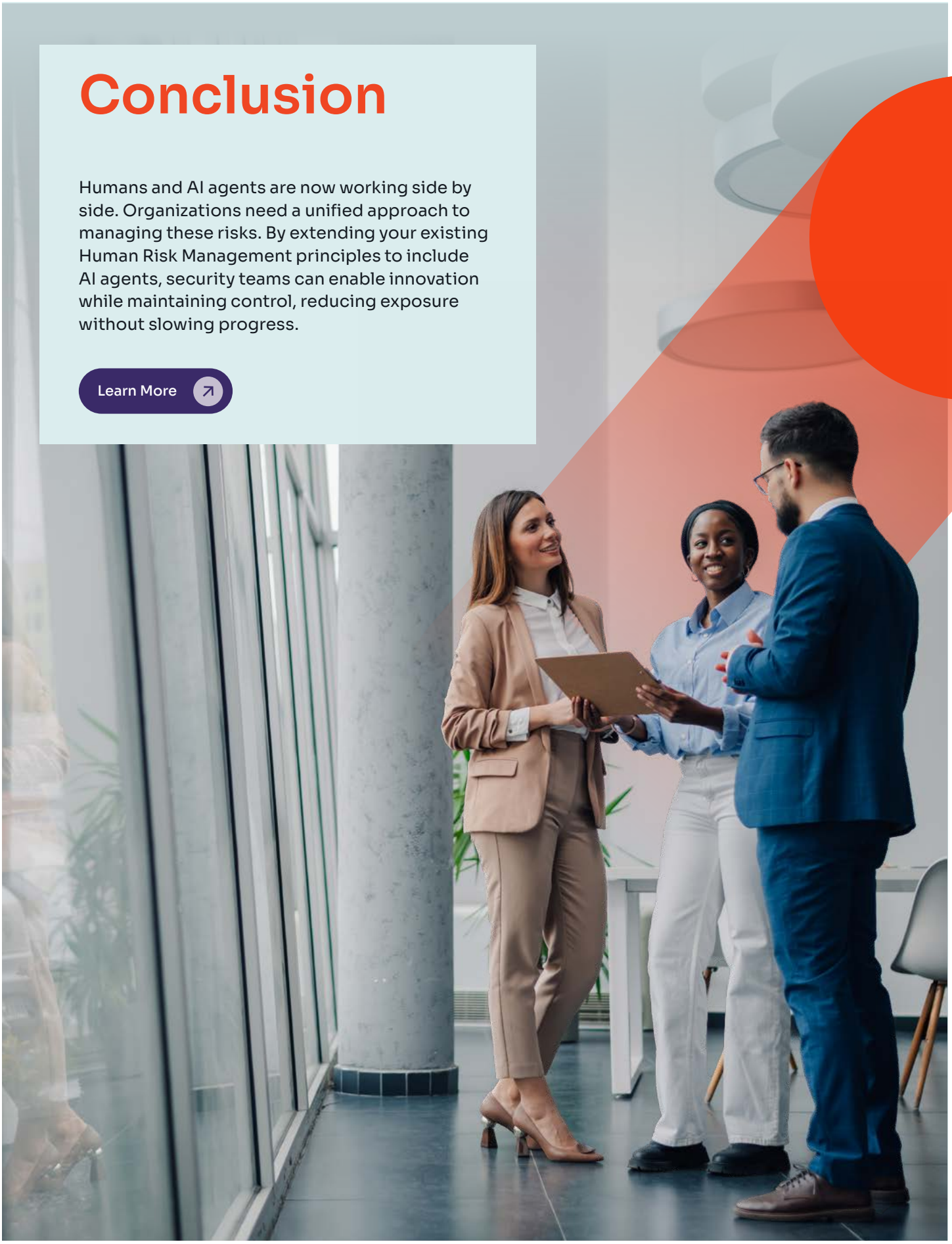
Dynamic Policy Enforcement
Security teams can define risk-based policies through API-driven controls, enabling consistent governance of AI usage across environments. Policies can be adjusted as new use cases emerge, ensuring AI adoption remains aligned with organizational security and compliance requirements.

Integrated Risk Scoring
Interaction data feeds into the broader human risk management framework, enabling organizations to assess risk across both human and AI behaviors. By incorporating human-agent interactions into risk scoring models, security leaders gain a more complete understanding of exposure and can prioritize mitigation efforts more effectively.

Conclusion

Humans and AI agents are now working side by side. Organizations need a unified approach to managing these risks. By extending your existing Human Risk Management principles to include AI agents, security teams can enable innovation while maintaining control, reducing exposure without slowing progress.

Learn More



Free Phishing Security Test
Find out what percentage of your employees are Phish-prone with your free Phishing Security Test

Free Email Exposure Check
Find out which of your users emails are exposed before bad actors do

Free Automated Security Awareness Program
Create a customized Security Awareness Program for your organization

Free Domain Spoof Test
Find out if hackers can spoof an email address of your own domain

Free Phish Alert Button
Your employees now have a safe way to report phishing attacks with one click

About KnowBe4

KnowBe4 empowers the human and AI workforce to make safer security decisions every day. Trusted by over 70,000 organizations worldwide, we help strengthen security culture and manage risk. Our comprehensive AI-driven HRM+ platform includes awareness and compliance training, cloud email security, real-time coaching, crowdsourced anti-phishing, AI Defense Agents, and more. As the only global security platform of its kind, KnowBe4 provides personalized content, tools, and techniques to keep the modern workforce safe from phishing, vishing, deepfakes, and emerging threats.

For more information, please visit www.KnowBe4.com.

knowbe4

KnowBe4, Inc. | 33 N Garden Ave, Suite 1200, Clearwater, FL 33755
855-KNOWBE4 (566-9234) | www.KnowBe4.com | Sales@KnowBe4.com