Currently in technical preview

Agent Risk Manager

Secure your AI agent workforce

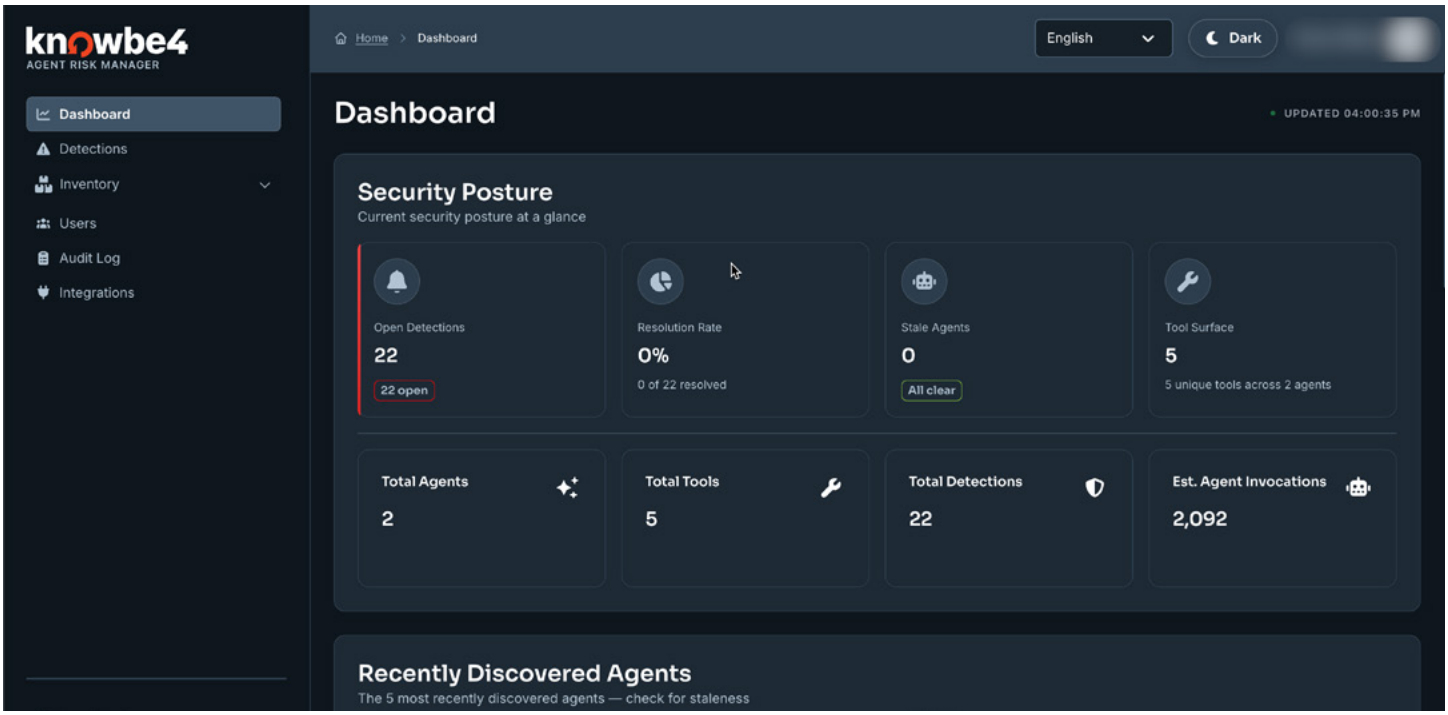
Highlight

- Purpose-built for Microsoft Copilot, Anthropic Claude, Google Gemini and OpenAI ChatGPT environments
- Real-time threat detection
- Full audit trail from day one
- No infrastructure required

See every AI agent. Detect every threat. Stay in control.

The rise of AI agents introduces new risks as they confidently handle critical workflows without a built-in understanding of your corporate policies. Traditional security tools such as SIEM and DLP are ill-equipped to monitor the complex inputs and outputs of these agents. This creates an AI security blind spot, leaving your organisation vulnerable to subtle exploits such as indirect prompt injections and 'permission creep', where agents accidentally expose sensitive data.

KnowBe4's Agent Risk Manager addresses this challenge. Leveraging over 15 years of expertise and human risk data, it is specifically designed for organisations using AI agents. Agent Risk Manager eliminates the security gap by providing your team with real-time visibility, automated threat detection, and active control over your organisation's AI agents and how your users interact with them.



Key benefits

Automated governance

Gain instant, zero-configuration discovery of every agent in your tenant. From official tools to 'shadow AI', you see it all without lifting a finger.

Predictable costs

Protect your budget and infrastructure from resource abuse and runaway API costs caused by inefficient or malicious AI calls.

Real-time contextual coaching

When a risky action is blocked, we explain it. Intercept threats and deliver immediate, in-the-moment coaching to your users.

Permanent risk reduction

Data shows that 70% of users who receive our real-time coaching never repeat the same risky behaviour. Improve AI agent use and prompt proficiency, reducing long-term organisational risk.

True behavioural alignment

Shape AI behaviour safely from the outside. Ensure consistent, secure interactions without needing to modify underlying models or trust opaque, third-party safety layers.

Holistic risk score

Close the AI blind spot: Unify human and AI behaviour data into a single score, giving you a clear picture of your organisation's true risk profile.

Complete coverage, from discovery to defence

Key features

Threat detection

Catch threats before they cause damage

Agent Risk Manager's detection centre gives your analysts a real-time feed of every risky event, categorised by threat type and severity. Visual risk gauges show at a glance which detection categories are most active, so you always know where to look first.

Blast radius

Understand the blast radius of every tool

The Tool Network view renders an interactive force-directed graph showing which agents share which tools. Node size scales with agent count, so you immediately see which tools have the highest potential blast radius if compromised.

Complete audit trail

A full audit trail down to the conversation ID

The Audit Log captures every event, such as benign tool invocations, detection triggers and schema discoveries, with metadata that takes you from a user's action all the way through the detection pipeline.

User risk scoring

Know which users are your highest AI risk

Agent Risk Manager automatically calculates a risk score for every user whose agents have triggered detections. Surface your riskiest users instantly, and drill down to the specific events driving their score.

How it works

Agent Risk Manager establishes a centralised interface to monitor and protect the growing workforce of 'non-human identities' and the humans interacting with them. It integrates with your AI agent provider to deliver a seamless, 'outside-in' security layer that doesn't require modifying your underlying AI models.

1. Connect
Link to your agent providers (Microsoft Copilot, ChatGPT, Gemini, Claude) via a guided, minutes-long onboarding process.

2. Intercept
Agent Risk Manager automatically monitors every agent tool execution and user interaction.

3. Analyse
Interactions run through parallel detection engines to identify prompt injections, PII leaks, resource abuse and more.

4. Action
If a threat is detected, Agent Risk Manager raises an alert or actively blocks the operation and triggers real-time user coaching with the full event logged for investigation.

5. Investigate
Via the dashboard, your security team can triage detections, review the audit trail, update statuses and fine-tune policies.

Six detection engines. Zero blind spots.

Agent Risk Manager includes purpose-built detection logic for every major AI agent attack category.

1. **Prompt injection**
Blocks jailbreaks and indirect injections that turn productivity tools into 'agents of chaos'.

2. **Sensitive information**
Scans for SSNs, passwords and PII, automatically redacting data to prevent DLP leaks.

3. **Unbounded consumption**
Protects your budget and infrastructure from resource abuse and excessive API calls.

4. **Content safety**
Flags inappropriate, harmful or policy-violating content in inputs and outputs before it reaches end users.

5. **Privilege escalation**
Stops agents from accessing resources or taking actions beyond their granted permissions, providing a critical control for high-privilege agents.

6. **Agent overstepping**
Identifies agents acting outside their intended operational scope, catching drift before it becomes a security or compliance incident.

Ready to secure your AI agents? **Learn more**

KnowBe4 UK, Ltd. | 1 Leeds City Office Park, Leeds, LS11 5BD
+44 (0) 1347 487512 | www.KnowBe4.com | Sales@KnowBe4.com