Derzeit in der technischen Vorschau

Agent Risk Manager

Schutz für Ihre neue „Belegschaft“ aus KI-Agenten

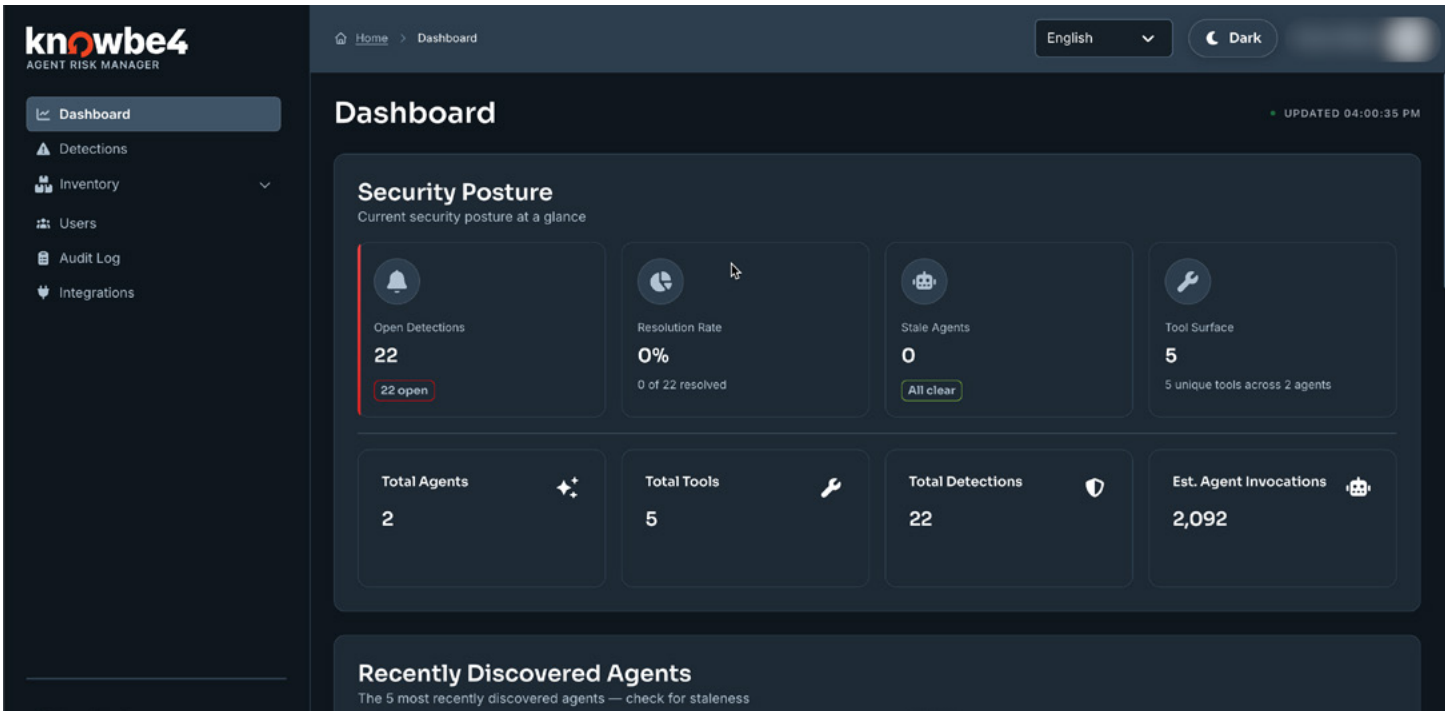
Highlight

- Speziell für Microsoft Copilot-, Anthropic Claude-, Google Gemini- und OpenAI ChatGPT-Umgebungen entwickelt
- Bedrohungserkennung in Echtzeit
- Kompletter Audit-Trail ab dem ersten Tag
- Keine Infrastruktur erforderlich

Alle KI-Agenten im Blick. Alle Bedrohungen im Blick. Alles im Griff.

Das Aufkommen von KI-Agenten birgt neue Gefahren, da diese kritische Workflows automatisch ausführen, ohne die Richtlinien Ihrer Organisation zu beachten. Herkömmliche Sicherheitstools wie SIEM und DLP wurden nicht konzipiert, um die komplexen Ein- und Ausgaben von KI-Agenten zu überwachen. Durch diese Schwachstelle in der KI-Sicherheit ist Ihre Organisation anfällig für subtile Exploits wie indirekte Prompt-Injection und „Permission Creep“ (Anhäufung von Zugriffsrechten), sodass KI-Agenten versehentlich sensible Daten offenlegen.

Wir gehen diese Herausforderung mit dem Agent Risk Manager von KnowBe4 an. Basierend auf Daten und Expertise zu menschlichen Risiken aus über 15 Jahren wurde der Agent Risk Manager speziell für Organisationen konzipiert, die KI-Agenten einsetzen. Der Agent Risk Manager schließt die Lücke bei der KI-Sicherheit. Ihr Team profitiert von Echtzeit-Transparenz, automatisierter Bedrohungserkennung und aktiver Kontrolle über die KI-Agenten Ihrer Organisation und deren Nutzung durch Ihre Mitarbeitenden.



Wesentliche Vorteile

Automatisierte Governance

Erhalten Sie einen Überblick über alle KI-Agenten in Ihrem Tenant, ganz ohne Konfiguration. Von offiziell genehmigten Tools bis hin zur Schatten-IT – alle Anwendungen werden aufgeführt.

Prognostizierbare Kosten

Schützen Sie Ihre Infrastruktur vor Missbrauch und Ihr Budget vor unkontrollierbaren API-Kosten durch ineffiziente oder schädliche KI-Aufrufe.

Kontextbezogenes Echtzeit-Coaching

Wird eine riskante Aktion blockiert, erhalten Sie eine Erläuterung. Stoppen Sie Bedrohungen und stellen Sie Coaching zeitnah bereit.

Dauerhafte Risikominderung

Daten belegen, dass 70 % der Nutzerinnen und Nutzer, die unser Echtzeit-Coaching erhalten, den gleichen Fehler nicht noch einmal machen. Verbessern Sie die Nutzung von KI-Agenten und die Formulierung von Prompts und reduzieren Sie dadurch langfristig das Risiko für Ihre Organisation.

Echte Verhaltensänderung

Formen Sie mithilfe externer Mechanismen das Verhalten von KI-Agenten. Sorgen Sie für konsistente und sichere Interaktionen, ohne zugrunde liegende Modelle zu modifizieren oder auf undurchsichtige Sicherheitskonzepte von Drittanbietern vertrauen zu müssen.

Umfassender Risk Score

KI-Schwachstellen beseitigen: Vereinen Sie Verhaltensdaten von Menschen und KI-Agenten in einem einzigen Score, um ein klares Bild vom Risikoprofil Ihrer Organisation zu erhalten.

Komplette Abdeckung von der Entdeckung bis zur Abwehr

Wichtige Funktionen

Bedrohungserkennung

Bedrohungen abwehren, bevor diese Schaden anrichten

Im Detection Center des Agent Risk Manager finden Ihre Analystinnen und Analysten einen Echtzeit-Feed aller riskanten Ereignisse kategorisiert nach Bedrohungstyp und Schweregrad. Grafische Risikoanzeigen bieten einen Überblick über die Kategorien mit der höchsten Aktivität, damit Sie schnell auf die drängendsten Probleme reagieren können.

Potenzielles Schadensausmaß

Potenzielles Schadensausmaß aller KI-Tools ermitteln

In der Ansicht „Tool Network“ wird ein interaktives Kräftediagramm der KI-Agenten und -Tools dargestellt. Die Größe der Knoten variiert je nach Anzahl der KI-Agenten. Sie erkennen sofort, bei welchen Tools im Falle einer Kompromittierung das größte Schadensausmaß besteht.

Kompletter Audit-Trail

Ein kompletter Audit-Trail mit Details zur forensischen Analyse

Das Audit-Protokoll erfasst jedes Ereignis, z. B. harmlose Tool-Aufrufe, Trigger und Schemen mit Metadaten. Sie können den kompletten Verlauf von der Nutzeraktion bis zum Ursprung verfolgen.

Risikobewertung

Welche Nutzerinnen und Nutzer haben das höchste KI-Risiko?

Der Agent Risk Manager berechnet automatisch einen Risk Score für alle Nutzerinnen und Nutzer mit riskanten KI-Interaktionen. Sie erkennen auf einen Blick die Nutzerinnen und Nutzer mit dem höchsten Risiko und können die Ereignisse, die den Score beeinflussen, genau analysieren.

Funktionsweise

Der Agent Risk Manager stellt eine zentrale Schnittstelle zur Überwachung und zum Schutz der zusätzlichen „Belegschaft“ aus KI-Agenten sowie deren Nutzung durch die Mitarbeitenden dar. Er lässt sich problemlos mit dem Anbieter Ihrer KI-Tools integrieren, um mithilfe externer Mechanismen eine nahtlose Sicherheitsebene zu bilden, bei der die KI-Modelle selbst nicht geändert werden müssen.

1. Vernetzen
Verknüpfen Sie die KI-Tools von Anbietern (Microsoft Copilot, ChatGPT, Gemini, Claude) in einem kurzen und angeleiteten Onboarding-Prozess.

2. Abfangen
Der Agent Risk Manager überwacht automatisch die Ausführung von KI-Agenten sowie Nutzerinteraktionen.

3. Analysieren
Interaktionen werden durch parallele Erkennungseinheiten geprüft, um u. a. Prompt-Injections, Leaks von personenbezogenen Daten und Missbrauch zu identifizieren.

4. Maßnahmen ergreifen
Wird eine Bedrohung erkannt, gibt der Agent Risk Manager eine Warnung aus oder blockiert aktiv den Vorgang und sendet ein Echtzeit-Coaching mit dem komplett protokollierten Ereignis.

5. Untersuchen
Ihr Sicherheitsteam kann über das Dashboard Bedrohungen prüfen, den Audit-Trail einsehen, Status aktualisieren und anhand der Erkenntnisse die Richtlinien abstimmen.

Sechs Erkennungseinheiten. Keine Schwachstellen.

Im Agent Risk Manager ist für jede größere KI-Agenten-Angriffskategorie eine Erkennungslogik integriert.

1. **Prompt-Injection**
   Blockiert Jailbreaks und indirekte Injections, die Tools in „unberechenbare KI-Agenten“ verwandeln.

2. **Sensible Daten**
   Scannt nach Sozialversicherungsnummern, Passwörtern und personenbezogenen Daten. Die Daten werden automatisch geschwärzt, um DLP-Leaks zu verhindern.

3. **Unkontrollierter Verbrauch**
   Schützt Ihr Budget und Ihre Infrastruktur vor Missbrauch und überbordenden API-Aufrufen.

4. **Sicherheit von Inhalten**
   Kennzeichnet unangemessene, schädliche oder gegen die Richtlinien verstoßende Inhalte in Ein- und Ausgaben.

5. **Eskalation von Privilegien**
   Stoppt KI-Agenten vor dem Zugriff auf Ressourcen oder vor unberechtigten Aktionen und bietet eine wichtige Kontrolle über KI-Agenten mit weitreichenden Privilegien.

6. **Überschreitung der Befugnisse durch KI-Agenten**
   Ermittelt KI-Agenten, die außerhalb ihres vorgesehenen Einsatzbereichs agieren. Anomalien werden erkannt, bevor es zu einem Sicherheits- oder Compliance-Vorfall kommt.

Bereit für den Schutz Ihrer KI-Agenten? **Mehr erfahren**

KnowBe4 Germany | Rheinstr. 45/46, 12161 Berlin – Deutschland
+49 30 34 64 64 60 | KnowBe4.de | kontakt@knowbe4.com