Disponível atualmente na versão prévia técnica

Agent Risk Manager

Proteja sua força de trabalho de agentes de IA

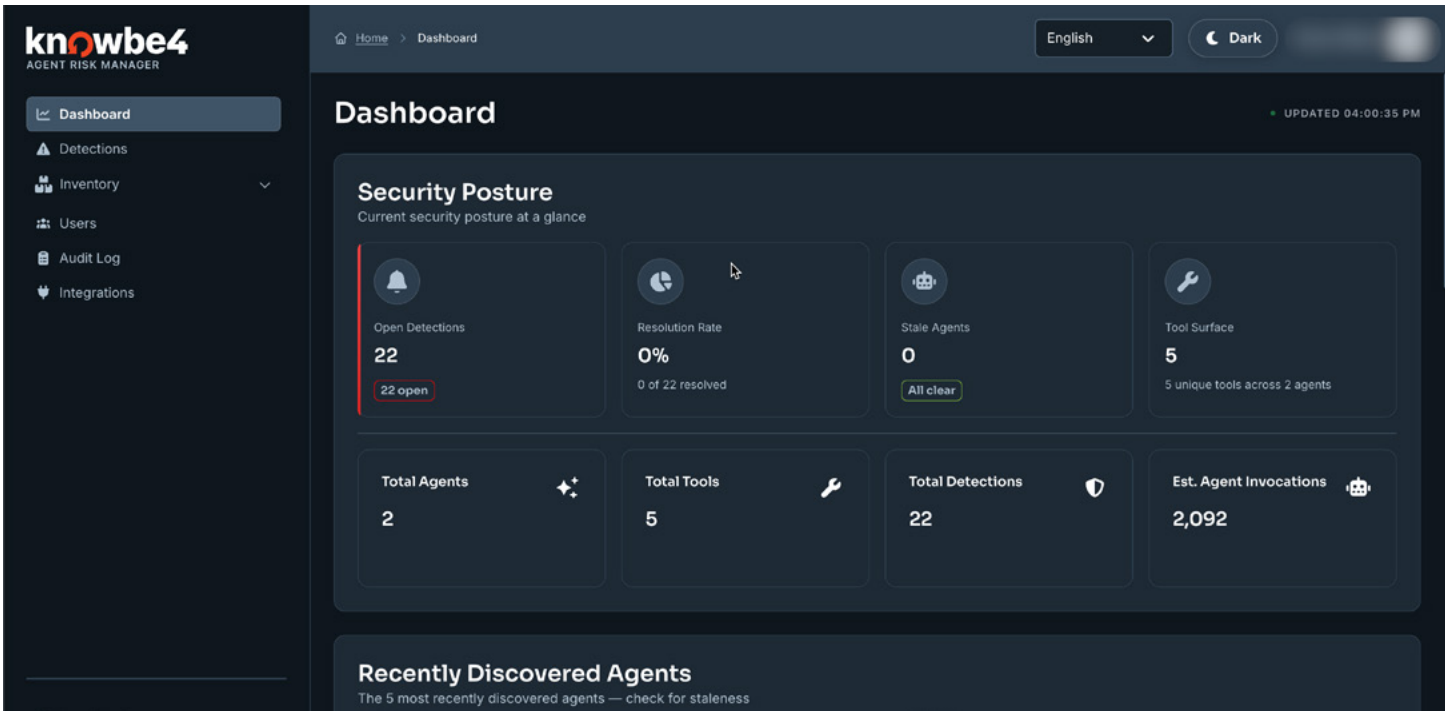
Destaques

- Desenvolvido especificamente para os ambientes do Microsoft Copilot, Anthropic Claude, Google Gemini e OpenAI ChatGPT
- Detecção de ameaças em tempo real
- Auditoria completa desde o primeiro dia
- Nenhuma infraestrutura necessária

Vigie todos os agentes de IA. Detecte todas as ameaças. Mantenha o controle.

O surgimento dos agentes de IA introduz novos riscos, uma vez que eles executam automaticamente fluxos de trabalho críticos de forma confiante, sem ter uma compreensão profunda das políticas da sua empresa. As ferramentas tradicionais de segurança, como de SIEM (Security Information and Event Management) e DLP (Data Loss Prevention), não estão preparadas para monitorar as entradas e saídas complexas desses agentes. Isso cria um ponto cego na segurança da IA, deixando sua organização vulnerável a explorações sutis, como injeções indiretas de prompts e o acúmulo gradual de permissões, que ocorre quando os agentes expõem inadvertidamente dados sigilosos.

O Agent Risk Manager da KnowBe4 resolve esse desafio. Com mais de 15 anos de dados e expertise em riscos gerados por falhas humanas, ele foi projetado especificamente para organizações que utilizam agentes de IA. O Agent Risk Manager elimina as lacunas de segurança, proporcionando à sua equipe visibilidade em tempo real, detecção automatizada de ameaças e controle ativo sobre os agentes de IA da sua organização e sobre a forma como os usuários interagem com eles.



Principais benefícios

Governança automatizada

Obtenha descoberta instantânea de todos os agentes no seu locatário, sem necessidade de configuração. De ferramentas oficiais à "IA oculta", você enxerga tudo sem esforço.

Custos previsíveis

Proteja seu orçamento e sua infraestrutura contra o uso indevido de recursos e custos descontrolados de API causados por chamadas de IA ineficientes ou maliciosas.

Coaching contextual em tempo real

Quando uma ação arriscada é bloqueada, nós explicamos o motivo. Intercepte ameaças e ofereça coaching imediato e em tempo real aos seus usuários.

Redução permanente de riscos

Os dados revelam que 70% dos usuários que recebem nosso coaching em tempo real nunca mais repetem o mesmo comportamento de risco. Otimize o uso de agentes de IA e aumente seu conhecimento sobre prompts, reduzindo o risco organizacional a longo prazo.

Alinhamento comportamental genuíno

Modele o comportamento da IA com segurança usando controles externos. Garanta interações consistentes e seguras sem precisar modificar os modelos subjacentes ou confiar em camadas de segurança obscuras de terceiros.

Nível de risco holístico

Elimine o ponto cego da IA: unifique os dados comportamentais humanos e da IA em um único nível, proporcionando uma visão clara do verdadeiro perfil de risco da sua organização.

Cobertura completa, da descoberta à defesa

Principais recursos

Detecção de ameaças

Detecte as ameaças antes que elas causem problemas

A central de detecção do Agent Risk Manager fornece aos seus analistas um fluxo em tempo real de todos os eventos de risco, categorizados por tipo e gravidade da ameaça. Os indicadores visuais de risco mostram imediatamente quais categorias de detecção estão mais ativas, para que você sempre saiba por onde começar a verificar.

Raio de alcance

Compreenda o raio de alcance de cada ferramenta

A visualização da rede de ferramentas gera um gráfico interativo baseado em relações, mostrando quais agentes compartilham quais ferramentas. O tamanho do nó varia de acordo com o número de agentes, permitindo que você identifique imediatamente quais ferramentas têm o maior raio de alcance potencial em caso de comprometimento.

Trilha de auditoria completa

Uma trilha de auditoria completa, que inclui até mesmo o ID da conversa

O log de auditoria registra todos os eventos, como chamadas de ferramentas não maliciosas, disparos de detecção e descobertas de esquemas, com metadados que permitem acompanhar desde a ação do usuário até o final do pipeline de detecção.

Nível de risco do usuário

Saiba quais usuários representam o maior risco de IA

O Agent Risk Manager calcula automaticamente um nível de risco para cada usuário cujos agentes tenham disparado alertas. Identifique instantaneamente seus usuários de maior risco e analise em detalhes os eventos específicos que estão contribuindo para o nível de risco deles.

Como funciona

O Agent Risk Manager oferece uma interface centralizada para monitorar e proteger a crescente força de trabalho composta por “identidades não humanas”, além das pessoas que interagem com elas. Ele se integra totalmente ao seu provedor de agentes de IA para oferecer uma camada de segurança contínua, “de fora para dentro”, que não exige modificações nos seus modelos de IA subjacentes.

1. Conecte-se
Conecte-se aos seus provedores de agentes (Microsoft Copilot, ChatGPT, Gemini, Claude) por meio de um processo de integração guiado que leva apenas alguns minutos.

2. Intercepte
O Agent Risk Manager monitora automaticamente todas as execuções de ferramentas de agentes e interações do usuário.

3. Analise
As interações são processadas por mecanismos de detecção paralelos para identificar injeções de prompts, vazamentos de informações de identificação pessoal (PII), uso indevido de recursos e muito mais.

4. Aja
Se for detectada uma ameaça, o Agent Risk Manager emite um alerta ou bloqueia ativamente a operação e aciona o coaching do usuário em tempo real, com o evento completo registrado para investigação.

5. Investigue
Por meio do painel, sua equipe de segurança pode classificar as detecções, analisar a trilha de auditoria, atualizar os status e ajustar as políticas.

Seis mecanismos de detecção. Nenhum ponto cego.

O Agent Risk Manager inclui uma lógica de detecção desenvolvida especificamente para cada uma das principais categorias de ataques de agentes de IA.

1. **Injeção de prompt**
Bloqueia invasões e injeções indiretas que transformam ferramentas de produtividade em “agentes do caos”.

2. **Informações sigilosas**
Verifica se há números de documentos de identidade, senhas e informações de identificação pessoal, ocultando automaticamente os dados para evitar vazamentos de DLP.

3. **Consumo sem limites**
Protege seu orçamento e sua infraestrutura contra o uso indevido de recursos e chamadas excessivas à API.

4. **Segurança do conteúdo**
Sinaliza conteúdo impróprio, prejudicial ou que viole as políticas nas entradas e saídas antes que ele chegue aos usuários finais.

5. **Elevação de privilégios**
Impede que os agentes acessem recursos ou executem ações que extrapolem as permissões concedidas, proporcionando um controle fundamental para agentes com privilégios elevados.

6. **Desvio de função de agentes**
Identifica agentes que atuam fora do escopo operacional previsto, detectando irregularidades antes que se transformem em um incidente de segurança ou de conformidade.

Que tal começar a proteger seus agentes de IA agora? **Saiba mais**

KnowBe4 Brazil | R. Gomes de Carvalho, 911 | Sala 208 - Vila Olímpia | CEP: 04547-003 | São Paulo-SP
Tel.: (0800) 761 2668 | www.KnowBe4.com/pt | Sales@KnowBe4.com