Actualmente en vista previa técnica

Administrador de riesgos de los agentes

Proteja a su fuerza laboral de agentes de IA

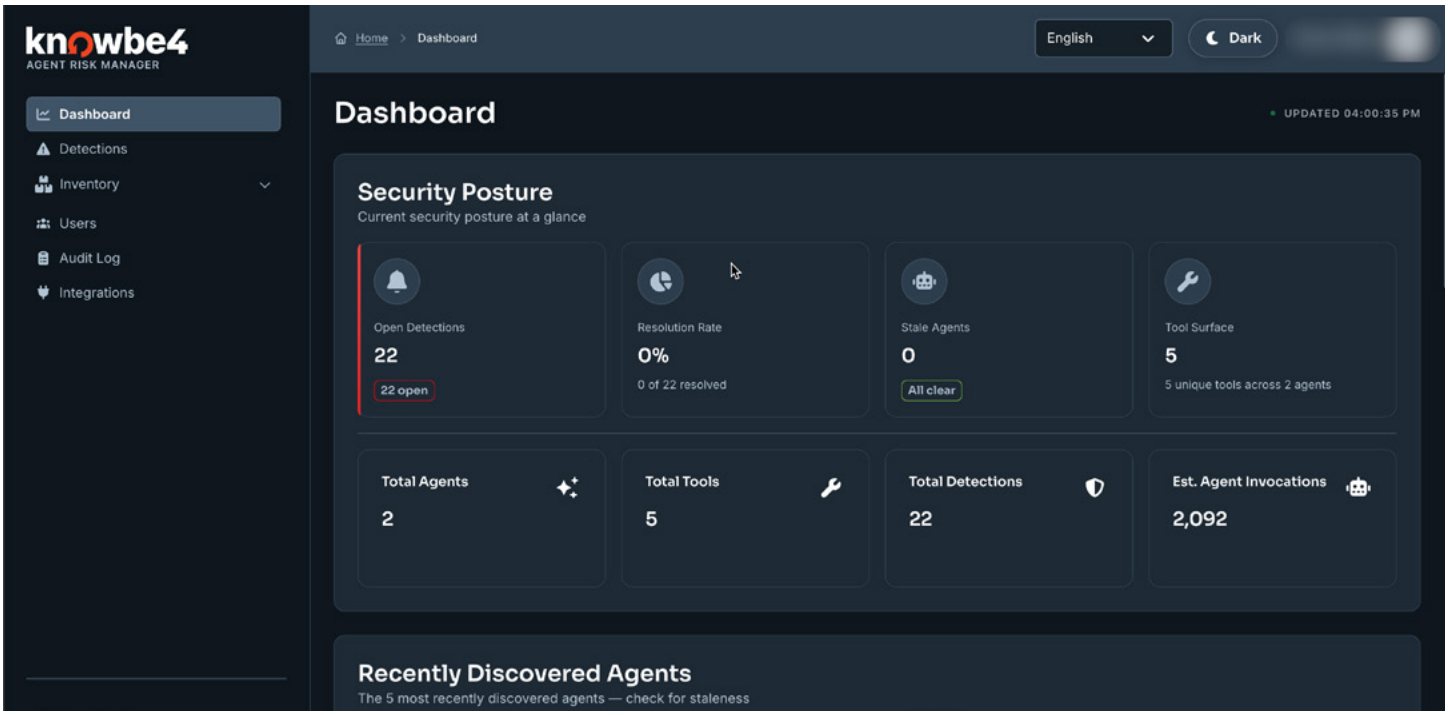
Aspectos destacados

- Desarrollado para entornos de Copilot (Microsoft), Claude (Anthropic), Gemini (Google) y ChatGPT (OpenAI)
- Detección de amenazas en tiempo real
- Pista de auditoría completa desde el primer día
- No se requiere ninguna infraestructura

Vea a todos los agentes de IA. Detecte todas las amenazas. Mantenga el control.

El auge de los agentes de IA está presentando riesgos nuevos, dado que se hacen cargo de los flujos de trabajo críticos con gran desenvoltura, sin comprender de manera integral sus políticas corporativas. Las herramientas de seguridad tradicionales, como SIEM (Security Information and Event Management) y DLP (Data Loss Prevention), no están equipadas para monitorear las complejas entradas y salidas de estos agentes. Esto genera un punto ciego en la seguridad de la IA, lo cual deja a su organización en un estado vulnerable ante explotaciones sutiles como la inyección indirecta de *prompts* o el aumento paulatino de permisos, en el cual los agentes exponen datos delicados de manera accidental.

El Administrador de riesgos de los agentes de KnowBe4 aborda este desafío. Se basa en más de 15 años de experiencia y datos sobre el riesgo humano, y está diseñado específicamente para las organizaciones que utilizan agentes de IA. El Administrador de riesgos de los agentes elimina las brechas de seguridad gracias a que le brinda a su equipo de seguridad visibilidad en tiempo real, detección automatizada de amenazas y control activo sobre los agentes de IA de su organización y la manera en que interactúan los usuarios con ellos.



Beneficios clave

Control automatizado

Descubra cada agente de su inquilino al instante y sin configurar nada. Usted puede verlo todo, desde herramientas oficiales hasta la IA no autorizada, sin ningún esfuerzo.

Costos predecibles

Proteja su presupuesto e infraestructura del abuso de recursos o de costos descontrolados de API (Application Programming Interface) provocados por llamadas de IA maliciosas o ineficientes.

Asesoramiento contextualizado en tiempo real

Cuando bloqueamos una acción riesgosa, explicamos por qué lo hicimos. Intercepte las amenazas y ofrezca asesoramiento inmediato a los usuarios.

Reducción de riesgos permanente

Hay datos que demuestran que el 70 % de los usuarios que reciben nuestro asesoramiento en tiempo real no repite jamás el mismo comportamiento riesgoso. Mejore el uso de los agentes de IA y la destreza con los *prompts* para reducir el riesgo de la organización a largo plazo.

Alineación real del comportamiento

Moldee de manera segura el comportamiento de la IA desde afuera. Garantice interacciones seguras y consistentes sin la necesidad de modificar los modelos subyacentes ni confiar en capas de seguridad opacas de terceros.

Puntaje de riesgo holístico

Elimine el punto ciego de la IA: unifique los datos del comportamiento humano y de la IA en un puntaje único, para obtener así un panorama claro del verdadero perfil de riesgo de su organización.

Cobertura completa desde el descubrimiento hasta la defensa

Funciones clave

Detección de amenazas

Detecte las amenazas antes de que provoquen daños

El centro de detección del Administrador de riesgos de los agentes les ofrece a sus analistas información en tiempo real sobre cada evento riesgoso, categorizado por tipo de amenaza y gravedad. Los indicadores visuales de riesgos le muestran de un vistazo qué categorías de detección son las más activas, para que siempre sepa a dónde mirar primero.

Área de impacto

Comprenda el área de impacto de cada herramienta

En la vista Red de herramientas se muestra un gráfico interactivo de fuerzas dirigidas donde figura qué agentes comparten qué herramientas. El tamaño de los nodos escala en función de la cantidad de agentes, para que pueda ver de inmediato qué herramientas tienen la mayor área de impacto potencial si se vulneran.

Pista de auditoría completa

La pista que abarca hasta el identificador de la conversación

En el Registro de auditoría se capturan todos los eventos, como las invocaciones benignas de herramientas, los desencadenantes de detecciones y los descubrimientos de esquemas, junto con metadatos que permiten rastrear las acciones de un usuario hasta llegar al flujo de detección.

Puntaje de riesgo de los usuarios

Entérese de qué usuarios representan el mayor riesgo de IA

El Administrador de riesgos de los agentes calcula de manera automática un puntaje de riesgo para cada usuario cuyos agentes hayan desencadenado una detección. Revele al instante quiénes son los usuarios con mayor riesgo e investigue en detalle los eventos específicos que justifican su puntaje.

Cómo funciona

El Administrador de riesgos de los agentes establece una interfaz centralizada para monitorear y proteger a la creciente fuerza laboral de "entidades no humanas" y a los humanos que interactúan con ellas. Se integra perfectamente con su proveedor de agentes de IA a fin de brindarle una capa de seguridad externa que opera de afuera hacia adentro y no requiere la modificación de sus modelos subyacentes de IA.

1. Conectar
Vincule sus proveedores de agentes (Microsoft Copilot, ChatGPT, Gemini, Claude) a través de un proceso de incorporación guiado que solo lleva unos minutos.

2. Interceptar
El Administrador de riesgos de los agentes monitorea automáticamente todas las interacciones con usuarios y ejecuciones de herramientas por parte de agentes.

3. Analizar
Las interacciones se ejecutan mediante motores paralelos de detección para identificar inyecciones de *prompts*, filtraciones de PII (información de identificación personal), abuso de recursos y mucho más.

4. Actuar
Si se detecta una amenaza, el Administrador de riesgos de los agentes genera una alerta o bloquea de manera activa la operación, desencadena el asesoramiento en tiempo real para los usuarios, y registra el evento completo para su investigación.

5. Investigar
A través del tablero, su equipo de seguridad puede evaluar las detecciones, revisar la pista de auditoría, actualizar estados y ajustar las políticas.

Seis motores de detección. Ningún punto ciego.

El Administrador de riesgos de los agentes incluye una lógica de detección dedicada para cada categoría principal de ataque mediante agentes de IA.

1. **Inyección de *prompts***
   Bloquea los intentos de evasión y las inyecciones indirectas que convierten a las herramientas de productividad en "agentes del caos".

2. **Información delicada**
   Busca números de identificación nacional, contraseñas y PII, y censura los datos automáticamente para impedir filtraciones de DLP.

3. **Consumo ilimitado**
   Protege su presupuesto e infraestructura del abuso de recursos y de las llamadas de API excesivas.

4. **Seguridad del contenido**
   Señala el contenido inapropiado, dañino o que infringe las políticas, presente en entradas y salidas, antes de que llegue a los usuarios finales.

5. **Escalada de privilegios**
   Impide que los agentes accedan a recursos o realicen acciones que estén por fuera de sus permisos otorgados, lo cual brinda un control crucial de los agentes con privilegios elevados.

6. **Extralimitación de agentes**
   Identifica a los agentes que actúan por fuera de su alcance operativo previsto, y detecta las desviaciones antes de que se conviertan en incidentes de seguridad o cumplimiento.

¿Todo listo para proteger a sus agentes de IA? **Más información**

KnowBe4 Brazil | R. Gomes de Carvalho, 911 | Sala 208 - Vila Olímpia | CEP: 04547-003 | São Paulo-SP
Tel.: +55 (0800) 761 2668 | www.KnowBe4.com/es | Sales@KnowBe4.com