Actuellement en version bêta

Agent Risk Manager

Sécurisez votre main-d'œuvre composée d'agents d'IA

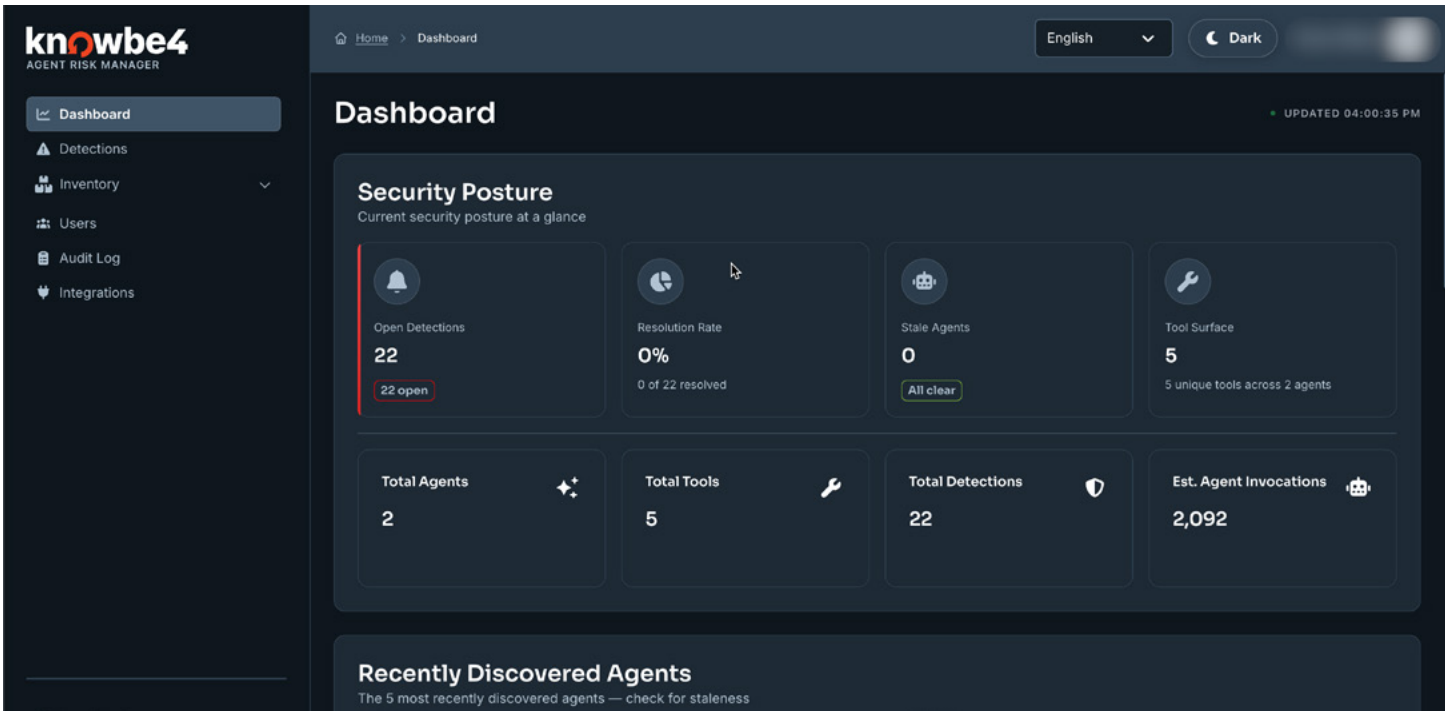
Points clés

- Conçu spécifiquement pour les environnements Microsoft Copilot, Anthropic Claude, Google Gemini et OpenAI ChatGPT
- Détection des menaces en temps réel
- Piste d'audit complète dès le premier jour
- Aucune infrastructure requise

Surveillez chaque agent d'IA. Détectez toutes les menaces. Gardez le contrôle.

L'essor des agents d'IA crée de nouveaux risques, car ils gèrent avec assurance des flux de travail critiques sans compréhension intrinsèque de vos politiques internes. Les outils de sécurité traditionnels, comme la SIEM (Security Information and Event Management, gestion des informations et des événements de sécurité) et la prévention des pertes de données, ne sont pas conçus pour surveiller les entrées et sorties complexes de ces agents. Cela crée des lacunes de sécurité liées à l'IA, et rend votre organisation vulnérable à des exploitations subtiles comme les injections d'invites indirectes et la « dérive des autorisations », à savoir quand un agent divulgue des informations sensibles par inadvertance.

Agent Risk Manager de KnowBe4 relève ce défi. S'appuyant sur plus de 15 ans de données et d'expertise sur les risques humains, il est pensé spécifiquement pour les organisations qui utilisent des agents d'IA. Agent Risk Manager comble les lacunes de sécurité en donnant à vos équipes une visibilité en temps réel, une détection automatisée des menaces, et un contrôle actif des agents d'IA et de la façon dont vos utilisateurs interagissent avec eux.



Principaux avantages

Gouvernance automatisée

Bénéficiez d'une découverte instantanée, sans configuration, de chaque agent de votre locataire. Des outils officiels à l'IA fantôme, vous disposez d'une visibilité complète, sans intervention manuelle.

Coûts prévisibles

Protégez votre budget et votre infrastructure contre toute utilisation abusive des ressources et des coûts d'API incontrôlés causés par des appels d'IA inefficaces ou malveillants.

Coaching contextuel en temps réel

Lorsqu'une action à risque est bloquée, une explication vous est fournie. Interceptez les menaces et fournissez à vos utilisateurs un coaching immédiat, au moment même où elles se produisent.

Réduction permanente des risques

Les données montrent que 70 % des utilisateurs qui bénéficient de notre coaching en temps réel ne reproduisent jamais le même comportement à risque. Améliorez l'utilisation des agents d'IA et la maîtrise des invites, tout en réduisant les risques organisationnels à long terme.

Véritable alignement comportemental

Façonnez le comportement de l'IA de manière sécurisée, par le biais de contrôles externes. Garantissez des interactions cohérentes et sûres, sans modifier les modèles sous-jacents ni dépendre de couches de sécurité tierces opaques.

Score de risque global

Comblez les lacunes de l'IA : unifiez les données comportementales du personnel humain et des agents d'IA en un score unique, afin d'obtenir une vision claire du véritable profil de risque de votre organisation.

Une couverture complète, de la découverte à la défense

Fonctionnalités clés

Détection des menaces

Détectez les menaces avant qu'elles ne causent des dommages

Le centre de détection d'Agent Risk Manager fournit à vos analystes un flux en temps réel de chaque événement à risque, classé par type de menace et niveau de gravité. Des indicateurs de risque visuels montrent en un coup d'œil les catégories de détection les plus actives, afin que vous sachiez toujours où concentrer votre attention.

Périmètre d'impact

Comprenez le périmètre d'impact de chaque outil

La vue du réseau d'outils affiche un graphe interactif à forces simulées, montrant quels agents partagent quels outils. La taille des nœuds varie en fonction du nombre d'agents, ce qui vous permet de voir immédiatement quels outils présentent le plus fort périmètre d'impact potentiel en cas de compromission.

Piste d'audit complète

Une piste d'audit complète jusqu'à l'ID de conversation

Le journal d'audit enregistre chaque événement, comme les appels d'outils anodins, les déclenchements de détection et les découvertes de schéma, avec des métadonnées permettant de suivre l'action d'un utilisateur tout au long du pipeline de détection.

Score de risque des utilisateurs

Identifiez les utilisateurs exposant le plus votre organisation aux risques liés à l'IA

Agent Risk Manager calcule automatiquement un score de risque pour chaque utilisateur dont les agents ont déclenché des détections. Identifiez instantanément vos utilisateurs les plus à risque et explorez en détail les événements spécifiques qui influencent leur score.

Fonctionnement

Agent Risk Manager établit une interface centralisée pour surveiller et protéger la main-d'œuvre en pleine croissance composée de systèmes automatisés et des humains qui interagissent avec ces derniers. Il s'intègre parfaitement à votre fournisseur d'agents d'IA pour fournir une couche de sécurité « de l'extérieur vers l'intérieur » fluide, sans nécessiter de modification de vos modèles d'IA sous-jacents.

1. Connexion
Connectez-vous à vos fournisseurs d'agents (Microsoft Copilot, ChatGPT, Gemini, Claude) par l'intermédiaire d'un processus d'intégration guidé de seulement quelques minutes.

2. Interception
Agent Risk Manager surveille automatiquement chaque exécution d'outil par les agents et chaque interaction utilisateur.

3. Analyse
Les interactions passent par des moteurs de détection parallèles afin d'identifier les injections d'invites, les fuites d'informations personnelles identifiables (IPI), l'utilisation abusive de ressources, et plus encore.

4. Action
Si une menace est détectée, Agent Risk Manager déclenche une alerte ou bloque activement l'opération, et lance un coaching en temps réel de l'utilisateur, l'ensemble de l'événement étant consigné pour investigation.

5. Investigation
Depuis le tableau de bord, votre équipe en charge de la sécurité peut trier les détections, consulter la piste d'audit, mettre à jour les statuts et affiner ses politiques.

Six moteurs de détection. Zéro angle mort.

Agent Risk Manager intègre une logique de détection spécialement conçue pour chaque grande catégorie d'attaques contre les agents d'IA.

1. **Injection d'invites**
Bloque les contournements des garde-fous de l'IA et les injections indirectes qui transforment les outils de productivité en « agents du chaos ».

2. **Informations sensibles**
Recherche les numéros de carte d'identité, mots de passe et informations personnelles identifiables (IPI), et masque automatiquement ces données afin d'empêcher toute fuite.

3. **Consommation non maîtrisée**
Protège votre budget et votre infrastructure contre l'utilisation abusive des ressources et les appels d'API excessifs.

4. **Sécurité des contenus**
Détecte les contenus inappropriés, nuisibles ou non conformes aux politiques dans les entrées et les sorties, avant qu'ils n'atteignent les utilisateurs finaux.

5. **Élévation de privilèges**
Empêche les agents d'accéder à des ressources ou d'effectuer des actions au-delà des autorisations qui leur sont accordées, pour un contrôle critique des agents aux privilèges élevés.

6. **Dépassement du périmètre de l'agent**
Identifie les agents qui agissent en dehors de leur périmètre opérationnel prévu, afin de détecter les dérives avant qu'elles ne deviennent des incidents de sécurité ou de conformité.

Vous souhaitez sécuriser votre main-d'œuvre composée d'agents d'IA ? **En savoir plus**

KnowBe4 France | 132 Rue Bossuet, Lyon, France, 69006
+31 (0)30 7996074 | www.KnowBe4.com/fr | Sales@KnowBe4.com