現在、テクニカルプレビュー版として提供中

Agent Risk Manager

AIエージェントを確実に守る

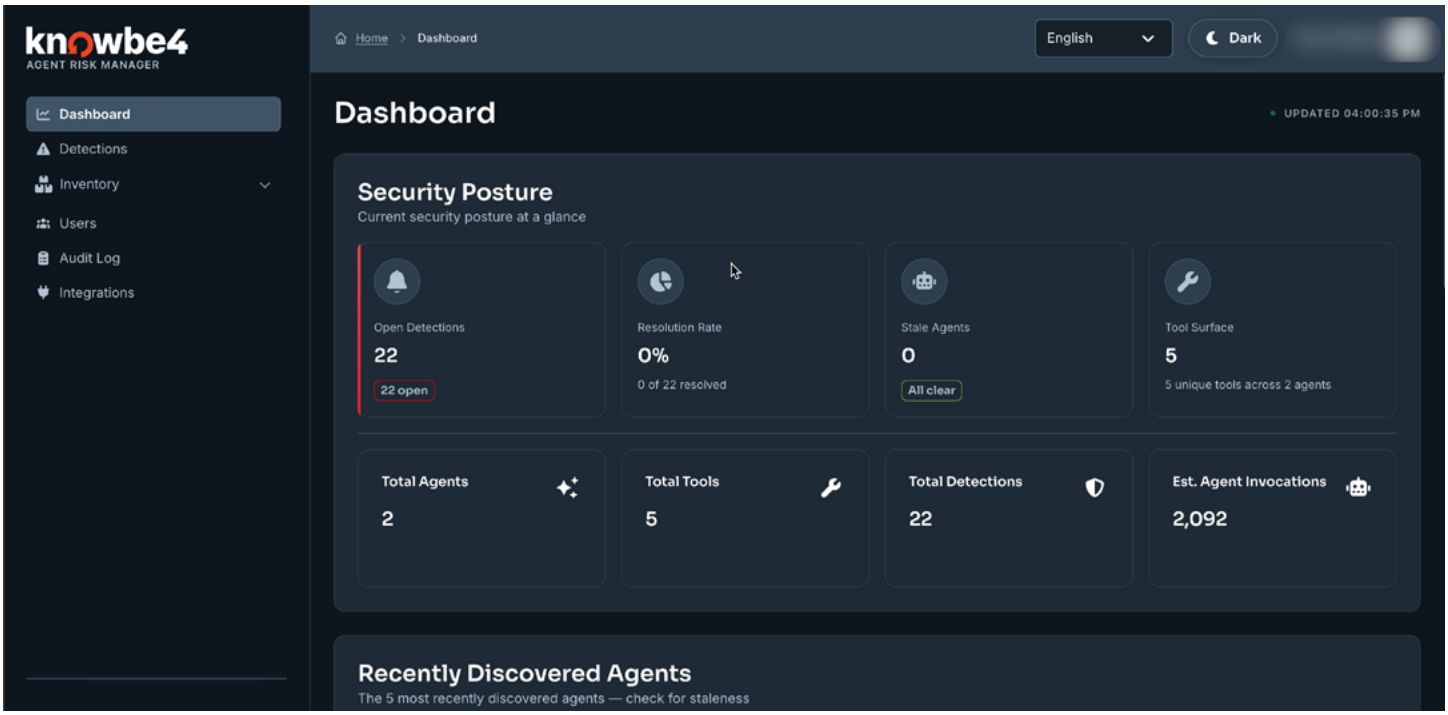
特長

- Microsoft Copilot、Anthropic Claude、Google Gemini、OpenAI ChatGPT環境向けに特化した設計
- リアルタイムで脅威を検知
- 導入初日から完全な監査ログを提供
- インフラストラクチャは一切不要

すべてのAIエージェントを可視化。あらゆる脅威を検知。常に制御を維持。

AIエージェントの普及に伴い、新たなリスクが生じています。というのも、AIエージェントは企業ポリシーを組み込んだ理解を持たないまま、重要なワークフローを自律的に実行してしまうためです。SIEM（セキュリティ情報およびイベント管理）やDLP（データ損失防止）といった従来のセキュリティツールでは、こうしたエージェントが行う複雑な入出力を監視することは困難です。その結果、AIセキュリティの死角が生まれ、組織は間接的なプロンプトインジェクションや、「特権クリープ」といった巧妙な攻撃にさらされます。これにより、エージェントが意図せず機密データを漏洩するリスクが発生します。

KnowBe4のAgent Risk Managerは、この課題に対応します。15年以上にわたる専門知識とヒューマンリスクに関するデータを活用し、AIエージェントを導入している組織向けに設計されたAgent Risk Managerが、リアルタイムでリスクを可視化し、自動で脅威を検知。さらに組織内のAIエージェントと、ユーザーとエージェントのやり取りを能動的に制御し、セキュリティ上のギャップを解消します。



主なメリット

ガバナンスの自動化

テナント内のすべてのエージェントをすぐに検出。面倒な設定は一切ありません。社内で承認されているツールから、無断利用の「シャドウAI」まで、あらゆるエージェントを簡単に確認できます。

予測可能なコスト構造

非効率的なAIの利用、またはAIの悪用によって引き起こされるリソースの不正利用や予期せぬAPI（アプリケーションプログラミング インターフェイス）コストの急増から、予算とインフラを保護します。

状況に応じたリアルタイムのコーチング

リスクを伴う行動がブロックされた場合は、その理由を説明。脅威を未然に防ぎ、ユーザーにその場ですぐにコーチングを行います。

継続的なリスク低減

リアルタイムコーチングを受けたユーザーの70%は同じ危険な行動を繰り返さないことが、データで示されています。AIエージェントの利用方法を改善し、プロンプト作成のスキル向上を図ることで、組織の長期的なリスクを低減します。

正しい行動アライメント

外部からの管理でAIの動きを安全に制御。基盤となるモデルを変更したり、サードパーティ製の不透明な保護レイヤーに頼ることなく、一貫性のある安全なやり取りを実現します。

包括的なリスクスコア

AIの死角を解消：
人間とAIの行動データを単一のスコアに統合することで、組織の真のリスクプロファイルを明確に把握できるようにします。

検知から防御まで、あらゆる局面を網羅
主な機能

脅威検知

被害が生じる前に脅威を防ぐ

Agent Risk Managerの検知センターは、脅威の種類や重大度別に分類されたあらゆるリスクイベントを、分析担当者へリアルタイムで提供します。リスク状況を視覚的に表示することで、最も活発な検知カテゴリを一目で把握でき、優先的に確認すべき領域を明確化します。

影響範囲

すべてのツールの影響範囲を把握する

ツールネットワーク・ビューでは、どのエージェントがどのツールを共有しているかを示す、インタラクティブな力学モデルによるグラフが表示されます。
ノードのサイズはエージェント数に応じて変化するため、どのツールが侵害された場合に最も広範囲に影響が及ぶ可能性があるかを一目で把握できます。

完全な監査証跡

会話IDまで遡れる完全な監査証跡を実施する

監査ログには、通常のツール実行、検知トリガー、スキーマの検出など、あらゆるイベントがメタデータとともに記録されるので、ユーザーの操作から検知パイプラインに至るまでの一連の過程を追跡することができます。

ユーザーのリスクスコアリング

AIリスクが最も高いユーザーを特定する

Agent Risk Managerは、エージェントが検知を発動したすべてのユーザーのリスクスコアを自動的に算出します。リスクの高いユーザーを即座に特定し、そのスコアに影響を与えている具体的な内容を詳細に分析します。

仕組み

Agent Risk Managerは、増加し続ける「非人間アイデンティティ」と、これとやり取りを行う人間を監視・保護するための一元的なインターフェースを提供します。AIエージェントプロバイダーとシームレスに連携し、基盤となるAIモデルを変更することなく、シームレスな「アウトサイド・イン」型のセキュリティレイヤーを実現します。

1. 連携
ガイドがあるのでオンボーディングはわずか数分で完了。これを通じて、エージェントプロバイダー（Microsoft Copilot、ChatGPT、Gemini、Claude）と連携します。

2. 仲介
Agent Risk Managerは、すべてのエージェントツールの実行とユーザーとのやりとりを自動的に監視します。

3. 分析
やりとりは並列検出（パラレルディテクション）エンジンによって処理され、不正なプロンプトインジェクション、PIIの漏洩、リソースの不正利用などが検知されます。

4. 初動対応
脅威が検出された場合、Agent Risk Managerはアラートを発するか、その操作を積極的にブロック。精査のためすべてのイベントを記録した上で、ユーザーへのリアルタイムのコーチングを開始します。

5. 調査
セキュリティチームは、検知結果の優先順位付け、監査ログの確認、ステータスの更新、ポリシーの微調整などをダッシュボードから行うことができます。

6つの検出エンジンで、あらゆる死角をなくす。

Agent Risk Managerには、主要なAIエージェント攻撃の各カテゴリに対応した専用検出ロジックが組み込まれています。

1 プロンプトインジェクション
生産性向上ツールを「混乱の要因」に変えてしまう、脱獄や間接インジェクションをブロックします。

2 機密情報
社会保障番号、パスワード、PIIをスキャンし、DLPによる情報漏洩を防ぐためにデータを自動的にマスキングします。

3 無制限消費
リソースの不正使用や過度なAPI呼び出しから、予算とインフラを保護します。

4 コンテンツの安全性
入力および出力に含まれる不適切、有害、またはポリシー違反のコンテンツを、エンドユーザーに届く前に検知します。

5 権限昇格
エージェントが許可された権限を超えてリソースにアクセスしたり、操作したりすることがないよう、高い権限を持つエージェントに対し重要な制御手段を提供します。

6 エージェントの暴走
想定された運用範囲外で活動しているエージェントを特定し、セキュリティやコンプライアンス上のインシデントに発展する前に、この逸脱を未然に防ぎます。

AIエージェントを守るために、ぜひ一度お問い合わせください。 **さらに詳しく**

KnowBe4 Japan合同会社 | 〒107-0052 東京都港区赤坂 9−7−1 ミッドタウン・タワー 18F | 03-4586-4540
www.knowbe4.com/ja | info@KnowBe4.jp