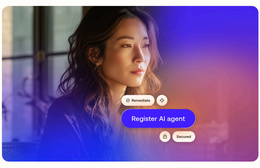
セキュアなエージェンティック企業のための設計指針

すべてのセキュリティおよびITリーダーが、エージェントが制御不能な規模に拡大する前に答えるべき3つの問い



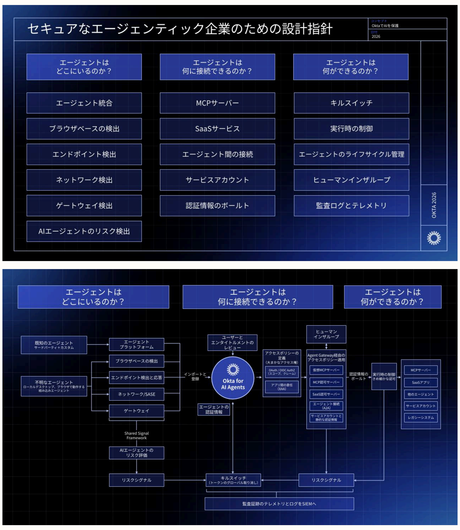
AIセキュリティの中心にあるアイデンティティギャップ

かつてソフトウェアは、指示されたことを実行していました。しかし今では、自ら判断して行動し、独自に接続を行う存在となっています。AIエージェントはすでに、従業員の業務を支援し、顧客対応を担い、サプライチェーン全体で稼働しており、それを取り巻くセキュリティよりも速いペースで拡大しています。過去10年にわたり、組織はボールト、最小権限、継続的認証によって、人間のアイデンティティセキュリティを強化してきました。しかし、AIエージェントの急増により、新たなアイデンティティギャップが生じています。誰でもエージェントを立ち上げることができ、エージェントはさらに別のエージェントを生成でき、それぞれがアプリ、API、SaaSツール、データシステムをまたいで接続を行います。その結果、特権アクセスを持つ数千もの新たなエンティティが、機械的なスピードで動作することになります。しかもその多くは、既存のセキュリティ制御の範囲外で動いています。

だからこそ、エージェントを正規のアイデンティティとして取り扱う必要があります。セキュアなエージェント型エンタープライズの運用は、エージェントが制御不能な規模に拡大する前に、明確な説明責任と可視性を確立することから始まります。そのために、組織は次の3つの問いに答えることができなければなりません。

1. エージェントはどこにいるのか？
2. エージェントは何に接続できるのか？
3. エージェントは何ができるのか？

これらの問いは、AIエージェントを保護するための運用上の設計指針を定義するものです。これらの問いに答えることは、単に現状を把握するためではありません。セキュアなエージェント型エンタープライズを運用するには、適切なシステム、アイデンティティ統制、ガバナンスモデルが必要です。



Q1：エージェントはどこにいるのか？

お客様から最も多く寄せられる懸念事項は「可視性」です。

貴社の環境にエージェントがいくつ存在するか、チームに尋ねてみてください。ほとんどの場合、正確な数を答えられません。従業員がブラウザで立ち上げたエージェントや、デスクトップでひそかに稼働しているエージェントは、多くの場合、認識も制御もされていません。

エージェントがどこで構築・デプロイされたかに関わらず、それを検出できる必要があります。

- **エージェント型プラットフォームの統合：**主要なサードパーティープラットフォーム上のエージェントや、自社で構築したカスタムビルドのエージェントを、アイデンティティプロバイダーに登録します。作成の時点で可視化できなければ、制御できません。
- **ブラウザベースの検出：**アイデンティティプロバイダーの範囲外のブラウザや拡張機能を介して動作するシャドーエージェントを検出します。これらは、従業員が許可なく立ち上げるエージェントです。
- **エンドポイント検出：**管理対象デバイス上で稼働しているエージェントを特定します。これは、モバイルデバイス管理やエンドポイントセキュリティと連携する必要があります。
- **ネットワーク検出：**ネットワークレイヤーで、認可されていないエージェント間やエージェントとツール間のトラフィックを検出します。エージェント同士、またサービスとの間で通信が発生するため、これらの接続を把握する必要があります。
- **ゲートウェイ検出：**API、MCP、エージェントのゲートウェイとやり取りする未登録のAIエージェントおよびOAuthクライアントを特定して管理します。エージェントがAPIを呼び出す場合、認証とログ記録を行う必要があります。
- **AIエージェントのリスク評価：**継続的に監視を行い、AIエージェントが悪用される脆弱性につながる構成ミスを特定します。あらゆるエージェントアイデンティティのセキュリティ態勢を分析することで、リスクを事前に検出します。

どのようなスタックを使用している場合でも、これらのすべてのソースからのシグナルを取り込む必要があります。検出レイヤーは、プラットフォーム、ツール、チームを横断して機能する必要があります。断片的な可視性は、可視性がないのと同じです。

Q2：エージェントは何に接続できるのか？

エージェントを可視化できたら、エージェントがアクセスできるすべてのリソースをマッピングし、アクセスポリシーを適用する必要があります。接続パスが一元管理されていない場合、1つのエージェントが優書されただけで、機械的なスピードで環境全体に連鎖的にアクセスが拡大してしまいます。

侵害されたエージェントによる影響が及ぶ範囲は、その接続によって定義されます。

- **MCPサーバーとリソース：**Model Context Protocolサーバーは、エージェントにツールとデータソースへのアクセスを提供します。これには、内部（アプリ、API、データベース、知的財産）と外部（Slack、GitHub、Notionなどのサードパーティ MCP）の両方が含まれます。これにより、セキュリティ境界は、エージェントがアクセスできるすべてのリソースにまで広がります。
- **SaaSアプリケーション：**エージェントは、従業員が日常的に利用しているのと同じSaaSツールに接続します。違いは、エージェントの方がはるかに速く動作し、より多くのデータにアクセスできる点です。侵害されたエージェントは、人間では到底不可能な速度で、接続されているすべてのSaaSアプリからデータを持ち出したり、変更を加えたりすることができます。
- **エージェント間の接続：**自律型エンティティ間のハンドシェイクと認可を保護します。エージェントが別のエージェントを呼び出すことで、ラテラルムーブメントが始まります。データの交換やタスクの委任の前に、双方がアイデンティティを検証する必要があります。
- **サービスアカウント：**長期間有効な静的認証情報の無秩序な増加を排除します。これらは、エージェントが従来のマシン間通信パターンから継承する「マスターキー」のようなものです。静的認証情報はすべて、悪用される可能性のある永続的なバックドアになります。
- **認証情報のボールト：**シークレットを自動的に保護し、ローテーションします。ローテーションされていないトークンは、攻撃者に対して扉を開いているようなものです。認証情報は、ボールトで管理し、動的に発行し、頻繁にローテーションする必要があります。どのエージェントは、タスク終了後も有効な認証情報を使用して動作すべきではありません。

こうした接続はすべてSIEMにログ記録する必要があります。見えないものは保護できません。すべてのエージェントの接続について、アクセス対象、アクセス日時、使用された認証情報を含め、すべての情報をセキュリティオペレーションセンターに送信し、監視と調査に活用する必要があります。

Q3：エージェントは何ができるのか？

エージェントがどこにいて、何に接続できるかを知るだけでは不十分です。実際の動作を制御し、中断できなければ意味がありません。エージェントがデータの持ち出しや認可されていないプロセスの立ち上げを開始した場合、迅速に対応する必要があります。

- **キルスイッチ：**エージェントが本来の任務から逸脱した場合、機密データに予期せずアクセスした場合、または脅威が検出された場合は、リスクを封じ込めるために、すべてのシステムでアクセス権の取り消しを即座に実行できる必要があります。
- **実行時の制御：**エージェントが達成しようとしている内容に基づいて、リアルタイムで認可を行います。コンテキスト、一連の処理の流れ、データの量を評価します。顧客レコード10件に対するクエリと、10,000件に対するクエリでは、明らかに挙動が異なります。プロンプトインジェクション攻撃を検知し、アクションが実行される前にツールレベルでポリシーを適用します。
- **エージェントライフサイクル管理：**初日には適切だったエージェントの権限が、90日後も適切であることはほとんどありません。アクセス権を継続的に見直して最小権限を維持し、認定を自動化し、エージェントの廃止時や従業員の離職時には直ちにアクセス権の取り消しを行います。
- **ヒューマンインザループ承認：**機密性の高い、またはリスクを伴う可能性のあるエージェントの動作には、人間の承認を必須とします。破壊的な操作、大量のデータへのアクセス、エージェントの特権のエスカレーションを防止します。
- **監査ログとテレメトリ：**すべてのエージェントの動作をログに記録し、SIEMに送信する必要があります。すべてのツール呼び出し、すべての認可の判断、すべてのアクセス試行です。実行時の制御やキルスイッチは、完全な可視性があって初めて機能します。

設計指針はあくまでベースライン

「エージェントはどこにいるのか」「エージェントは何に接続できるのか」「エージェントは何ができるのか」という3つの問いは、達成すべき目標ではありません。本番環境でAIエージェントを運用するために必要な最低限の基準です。

これらの問いに答えられない組織は、状況を把握できない状態で運用を進めているようなものです。取締役会や監査役から問われたとき、または侵害が発生して規制当局から問われたときに、「わかりません」では通用しません。

先手を打つ組織は、すでにこれを認識しています。業界をリードする企業は、最初のインシデントが起きてから対応に追られるのを待ってはいません。すでにAIエージェントを正規のアイデンティティとして扱い、エージェント導入の初期段階から、可視化・制御・ガバナンスを組み込んでいます。エージェントが制御不能な規模に拡大する前に、3つの問いに答えています。

この設計指針は、一度限りの監査ではありません。エージェントが数十から数千に増えるに伴い、この3つの問いに答え続けることは、継続的な運用規律となります。人間のアイデンティティセキュリティの構築に尽力してきたこの10年間の成果が、エージェントによって損なわれる事態は避ける必要があります。

Oktaは、大規模なAIエージェント運用におけるセキュリティ対策に取り組む大手企業との継続的な提携作業に基づいて、この設計指針を作成しました。Okta Platformによる実装方法については、okta.com/ai-agentsをご覧ください。