E-Book

AIエージェントの時代：アイデンティティ・信頼・コントロールの再考

目次

アイデンティティの新領域

91%

本番環境ですでにAIエージェントを利用している組織の割合

10%

非人間アイデンティティの管理に向けて十分な戦略を持っている組織の割合

50%以上

AIエージェントのガバナンスとコンプライアンスを最大の懸念事項の一つとして挙げている組織の割合

出典：

従来のIAMは、人間の速度に合わせて設計されています。しかし、AIは機械的なスピードで動きます。

組織において静的なチャットボットから自律型エージェントへの移行が進むにつれ、アイデンティティの本質が変わりつつあります。自律型エージェントは、情報を提供するだけでなく、複雑な環境をまたいで複数のステップを踏むタスクを実行します。それによって可視性に大規模なギャップが生じるため、セキュリティチームが気づかないうちに不正な操作が行われる可能性があります。

このeBookは、AIを活用する組織のセキュリティに関する包括的なソートリーダーシップシリーズ（7部構成）から重要な点を選んでまとめたものです。主要なトピックを要約しただけでなく、自律型エージェントへの対応に必要な新たなアーキテクチャ戦略についても深いインサイトをご紹介しています。

エージェントの速度でIAMを実現

機械的なスピードで動き、1分あたり最大6,000件の操作を実行できるAIエージェントが相手では、従来型の人間指向の認可モデルは機能しません。セキュリティを、同意に基づく手動の承認から自動化された実行時権限へと移行させることで、「パニック」に陥ったエージェントが原因で、数秒のうちに大量のデータが失われるといった事態を防ぐ必要があります。

推奨事項

- **ポリシー主導型のルール：**セキュリティが機械的なスピードに対応できるように、エージェントの速度に合わせて拡張できる自動ルールを導入します。
- **エフェメラル資格情報：**無期限に有効な認証情報ではなく、数分で無効になる認証情報を使えば、攻撃の機会が大幅に減少します。
- **関係性ベースのアクセス：**きめ細かい関係性ベースのアクセスコントロールを使用して、ミリ秒単位の認可チェックを可能にします。
- **継続的な評価：**「許可しておしまい」ではなく、エージェントが実行するあらゆる操作を逐一評価します。

現在のセキュリティスタックで、1秒間に100件のコマンドを実行する不正エージェントがデータベースを削除する前に、その動きを阻止できるでしょうか。エージェントの操作に依然として人間の承認を必要としているようでは、イノベーションが遅れるだけでなく、機械的なスピードでの侵害を無視することになります。

参考トピック：AIセキュリティ：エージェントの速度でIAMを実現

認可の存続が意図した長さを超えるケース

「認可のドリフト」は、ある特定のタスクのために発行されたデジタルキーが、その作業が完了した後も数か月にわたって有効なまま残っている場合に発生します。AIエージェントが普及した世界では、このような休眠状態のトークンは時限爆弾となります。なぜなら攻撃者が、パスワードを解読する必要すらなくSaaS間の正規な連携を乗っ取ることができるためです。

推奨事項

- **長期的に運用される委任アイデンティティ：**どのAIエージェントも、ユーザーとは別に、管理・監査・追跡が可能な独自のアイデンティティを持たなければなりません。
- **継続的に更新できる認可：**タスクやユーザー、環境の変化に応じてアクセス権を自動的に調整し、ゼロスタンディング特権の態勢を維持します。
- **システム間で即座にプロビジョニング解除：**共有シグナルによる失効を導入すれば、あるアプリケーションで失効した認証情報は直ちにあらゆる場所で無効になります。
- **リアルタイムのインテント検証：**一つひとつのアクションを、認証情報の発行時だけではなく、実行時にも最新のポリシーと照合して検証します。

タスクの完了時にエージェントの認証情報を失効させる、正式な自動プロセスは整備されているでしょうか。トークンが本来の用途を終えた後も有効なままになると、「サイレント」な侵害に対して無防備な状態を作り出すことになります。

参考トピック：AIエージェントのセキュリティ：認可の存続が意図した長さを超えるケース

AIセキュリティとコンプライアンスのための行動計画

AIエージェントのエコシステムを制御する準備はできていますか。AIアイデンティティセキュリティコンプライアンスチェックリストを確認し、次のステップに役立てましょう。

チェックリストで詳細を見る

複数のシステムにまたがるエージェントの信頼

AIエージェントが組織の境界を超えて独立したシステムにアクセスすると、制約に関する「記憶」が失われることが多くあります。多くのアイデンティティプロバイダーはトークンを個別に検証するため、1つの信頼ドメインで侵害が発生すると、その影響が他の何百ものドメインに波及するおそれがあります。

推奨事項

- **検証可能な委任：**暗号学的に検証可能な委任の仕組みを実装し、システム間を移動するアイデンティティが人間のものかエージェントのものかを明示的に区別します。
- **ポータブルな制約：**「読み取り専用」などのセキュリティ上の制約が、トークンと共に信頼ドメイン間を移動し、引き渡しの際に失われないように維持します。
- **失効管理の連携：**連携するリスクシグナル（IPSIEなど）を導入することで、リアルタイムのセキュリティアラートが異なるサービスプロバイダー間で共有されるようにします。
- **取得時間の制御：**リクエストの発信元ドメインに関係なく、エージェントがAPIを呼び出す瞬間にアクセス権を検証できるよう、きめ細かい認可を適用します。

エージェントがパートナーのシステムへと移動したとき、その時点での安全性を保証するのは誰でしょうか。静的で分散した信頼モデルでは、クロスドメインのトークンハイジャックに対抗する共通の防衛手段はありません。

参考トピック：AIセキュリティ：複数のシステムにまたがるエージェントの信頼

AIエージェントへの委任の保護

エージェントが専門的なタスクを処理するためにサブエージェントを生成する再帰的な委任は、セキュリティリスクが「マトリョーシカ」のように入れ子になった状態を作り出します。委任の経路を厳密に管理してスコープを縮小しない限り、たった1つの悪意のあるプロンプトが「エージェントセッションスマグリング」を引き起こし、サブエージェントが目に見えないところで不正操作を実行してしまう事態になりかねません。

推奨事項

- **帯域外検証：**重大な影響を伴う操作については、プッシュ通知や個別のUIを使用して、エージェントの主要なチャットチャネルの外で検証を行います。
- **コンテキストグラウンディング：**一つひとつのエージェントセッションを元のタスクに結びつけ、エージェントの動作が意図した目的から逸脱し始めた場合に、「セマンティックドリフト」を継続的に検知し、フラグを立てます。
- **アイデンティティと権限の検証：**暗号化された認証情報の提示を義務付けるゼロトラストアーキテクチャを導入することで、システムリソースのやり取りが始まる前にエージェントのアイデンティティや具体的な権限が検証されるようにします。
- **ユーザーへの可視性：**ユーザーに対し、徹底的な透明性、あらゆるツール呼び出しの表面化、バックグラウンド推論、実行ログをリアルタイムで提供することで、不正な指示が持ち込まれるリスクを抑えます。

ワークフロー内で3段階にわたって連鎖したサブエージェントによるあらゆる操作について、その経路を暗号的に証明することはできるでしょうか。検証可能な委任チェーンがないマルチエージェントのエコシステムは、ラテラルムーブメントの格好の標的となります。

参考トピック：エージェントセキュリティ：委任チェーン

企業におけるAIエージェント：リーダーが見過ごしてはならないセキュリティリスク

AIエージェントは、パスワードをリセットし、お金を動かし、コードをリリースします。このホワイトペーパーでは、5つの一般的なユースケースを取り上げ、アイデンティティリスクを明らかにします。

ホワイトペーパーで詳細を見る

認可のギャップを埋める

AIエージェントは、経営陣の高レベルな権限を使ってデータを取得し、その情報をSlackやTeamsなどの共有ワークスペースに出力することがあります。このような「認可のギャップ」があると、役員報酬や取締役員会の資料といった機密データが、権限のない受信者に誤って共有されてしまう可能性があります。

推奨事項

- **オーディエンス認識型の認可：**「権限の共通部分」をリアルタイムで計算することで、ワークスペースにいる全員に閲覧が認可されているデータのみをエージェントが取得するようになります。
- **スコープを設定した取得：**取得されたデータを後からフィルタリングするのではなく、取得の前にエージェントの認証情報にスコープを設定します。そうすれば、機密性の高いファイルへのアクセスを未然に防ぐことができます。
- **関係性ベースのアクセス：**静的なロールにとどまらず、誰がどのチャネルに属し、現在どのような権限を持っているかを踏まえた関係性ベースのモデルに移行します。
- **ポリシーの継続的な同期：**共有ワークスペースでのユーザーの参加や離脱に応じて「権限グラフ」の精度が維持されるよう、アイデンティティガバナンスと認証エンジンを統合します。

共有ワークスペースに展開されたすべてのAIエージェントについて、その出力が、参加者全員の権限の共通範囲（最小権限）に限定されていることを証明できるでしょうか。オーディエンスを考慮しないエージェントは、内部データ漏洩の最大のリスクとなります。

関連トピック：AIエージェントの認可ギャップ

サイバーフィジカルセキュリティ

ヘルスケアや製造業といった物理的な産業でAIエージェントが普及すると、認可エラーは、データ漏洩ばかりか安全上の危険をも引き起こすようになります。これからのセキュリティは、デジタルエージェントに起因する物理的な損害を阻止しなければなりません。

推奨事項

- **エンドツーエンドの追跡可能性：**委任トークを使ったCross App Access（XAA）を導入し、あらゆる自動アクションについて、実行したエージェントと元のユーザーの双方に紐づけて追跡できるようにします。
- **オンデマンドの認証情報発行：**Token Vaultを利用して、有効期間の長い静的な認証情報ではなく、スコープの設定された有効期間の短いトークンが実行時にのみ取得されるようにします。
- **人間参加型（HITL）の検証：**CIBA（Client-Initiated Backchannel Authentication）を使用すれば、影響の大きい操作や想定外の操作を行うエージェントに対し、人間による明示的な承認を求めることができます。
- **リアルタイムのポリシー適用：**静的なロールに基づくのではなく、意思決定の瞬間に安全性の制約や操作の制限を評価するきめ細かい認可に移行します。

重要なシステムにアクセスしようとするすべてのエージェントについて、認可の範囲やアクティブな認証情報を明確に定義できているでしょうか。エージェントに対して何が許可されているのか、またその理由を明確に説明できない場合、その可視性の欠如こそが、主な攻撃対象領域となります。

参考トピック：AIエージェント：サイバーフィジカルIAMの安全性

アイデンティティと認可を統合して自律的な信頼を実現

AIエージェントが自律的に動作するようになると、従来型のアイデンティティや認可モデルは機能しなくなります。人間のユーザーにアクセス権を付与するためのシステムでは、委任や間接的アクション、機械主導型の意思決定を適切に管理できません。自律的な信頼を確立するには、アイデンティティや認可を1つの統合コントロールレイヤーとして扱い、エージェントに何を許可するか、またその境界を定義する必要があります。

推奨事項

- **エージェントを独立したアイデンティティとして表現する：**AIエージェントを、ユーザーやアプリケーションの延長としてではなく、独自のアイデンティティとして扱います。
- **アイデンティティと認可の判断を統合する：**エージェントのあらゆるアクションに一貫したアクセス判断を適用することで、コントロールの断片化を防ぎます。
- **委任を明示的に考慮する：**ユーザーやシステム、他のエージェントの代理として行動しているエージェントを認識できるような認可モデルを設計します。
- **広すぎるアクセス権を減らす：**意図したスコープを超える操作をエージェントに許可するような永続的な権限を制限します。
- **アイデンティティを使って境界を定義する：**アイデンティティと認可を、エージェントの動作を監視するだけでなく制限するメカニズムとして適用します。

アイデンティティと認可が統合されたシステムとして機能すれば、エージェントの行動や委任に合わせて自律的な信頼を適用できるようになります。

参考トピック： 自律的な信頼のためのオペレーティングシステムとしてのアイデンティティと認可

Oktaを活用

Oktaは、人間の意図と機械的なスピードでの実行の間にあるギャップを埋めることで、統合されたコントロールプレーンとしてのアイデンティティセキュリティファブリックを実現します。一元的なレイヤーが、ポリシーに基づくリアルタイムの制御を可能にするため、あらゆる信頼ドメインでエージェントのアクションが管理・追跡され、セキュリティが維持されます。アイデンティティをインフラストラクチャの中核に据えることで、組織は、どの自動ワークフローにも安全性と認可が組み込まれているという確信を持ってAIエージェントを拡張することができます。