ホワイトペーパー

AIエージェントを開発から エンタープライズ拡張まで 保護する

目次

エグゼクティブサマリー

AIエージェントは、仕事のあり方を変化させているだけでなく、アイデンティティそのものを再定義しています。

AIエージェントは、自主的に動作するよう設計されているため、自律的で目標志向であり、次第に人間の監視なしで動作するようになっています。飽くことなくデータを取り込み、システムの至る所で常に情報の分析や、コード記述、Eメールの送信、意志決定を行っています。がむしゃらにゴールを目指し迅速に目標を達成しようとする人のように、エージェントはさらに多くのデータを取り込むために、現在の技術の限界を超える新たな方法を見つけるでしょう。適切な安全対策なしでは、意図せずに不正な状態となり、損害と混乱を引き起こしかねません。しかしAIエージェントについて、エコシステムのどこに存在するか、どのデータとシステムにアクセスできるか、不正な状態となったときに誰が責任を負うか、といった非常に初歩的な質問にも答えられない組織がほとんどです。Oktaの「AI at Work 2025」レポートによると、**91%の組織がAIエージェントを使用している**ものの、**44%の組織ではガバナンス体制が整っていません**。その結果、セキュリティの未整備な新領域が生まれ、自律的な非人間アイデンティティが激増する中で、認証、認可、可視性のための一貫したフレームワークが求められています。

エージェン型AIの増加によって、アイデンティティとアクセス管理の土台そのものが揺らいでいます。従来の制御は人間向けに作られているため、人間の監視なしに複雑なワークフローとAPIチェーンを大規模に開始できるエージェントに対応できません。次世代のアイデンティティセキュリティは、**AIと同じスピードで進化**し、その規模、速度、知能に対抗できるものでなければ、**顧客との信頼関係を維持**できません。

このホワイトペーパーでは、**個々のエージェントとエージェントの集合全体を保護**する方法について考察します。コードの1行目から、エージェントを管理するエンタープライズコントロールプレーンにまで、くまなくセキュリティを組み込みます。なぜなら、自律型AIの時代における**アイデンティティとは、ただの本人確認ではなく、コントロールを維持する方法だからです。**

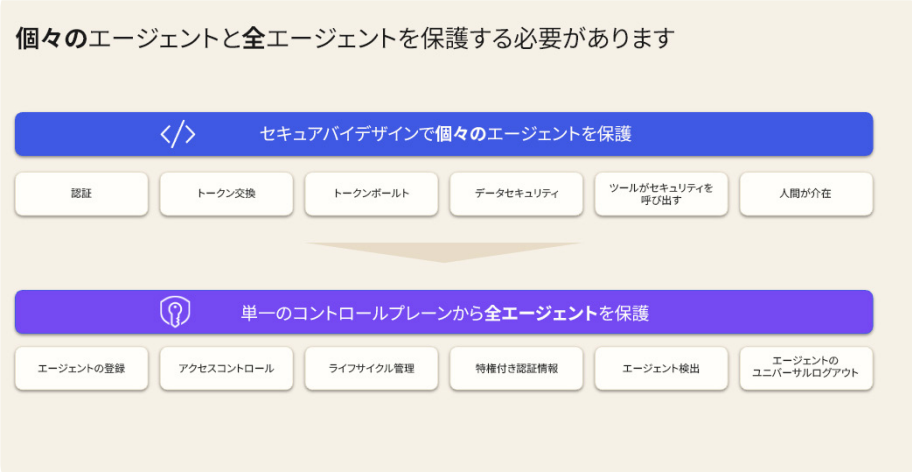
本ホワイトペーパーの内容

AIエージェントのセキュリティに関する二面的な課題に対応する、包括的なフレームワークをご紹介します。

1. 開発者向け – セキュアバイデザインで個々のエージェントを保護：開発者が作成時に組み込む必要のある重要なセキュリティパターンを学びます。堅牢なユーザー**認証**や、APIアクセスのための安全な**トークンボールト**、RAGシステム向けのきめ細かな**データ認可**、重要なアクションに人間を介在させる**ヒューマンインザループ**などを取り上げます。

2. ITチームとセキュリティチーム向け – 単一のコントロールプレーンから全AIエージェントを保護：大規模なエージェント管理に必要なエンタープライズレベルの機能について解説します。**エージェントの検出**（「シャドーAI」を見つけるため）や、**エージェントの登録**でのアイデンティティと所有権の確立、クロスドメインの信頼による**徹底的なアクセスコントロール**（エージェントがユーザーコンテキストを保ちながら、組織の境界を越えて安全にリソースにアクセスできるようにする）、**ライフサイクル全体の管理**、ユニバーサルログアウト機能による**脅威検知**などを取り上げます。

3.



主なポイント

- **「ガバナンスのギャップ」が一番のリスク：**根本的な問題はAI自体ではなく、AIエージェントを制御・統制するために必要なガバナンスをはるかに上回るペースで導入が進んでいることです。ガバナンスには、予防的コントロール（アクセスポリシー、最小権限、認可ルール）と、検出用の監視（認定、アクセスレビュー、振る舞い監視）の両方が含まれます。組織がAIエージェントを効果的に管理するためには、この2つが必要です。
- **セキュリティ上重要な2つの課題：**開発者レベルのセキュリティ（エージェントの適切な構築）と、企業レベルのガバナンス（全エージェントを大規模に管理）の両方に対応しなければ、完全な戦略とは言えません。
- **統合プラットフォームの必要性：**このギャップを埋め、データプライバシーのリスクを軽減して、組織が安心してAIを拡張できるようにするには、統合アイデンティティプラットフォームを用いて、エージェントを正規のアイデンティティとして扱い、検出と登録以降のライフサイクル全体を管理することが不可欠です。

セキュアバイデザインで個々のエージェントを保護

AIエージェントを構築する際に直面するセキュリティ要件は、従来のアプリケーション開発とは根本的に異なっています。セキュリティを後から付け足すことはできません。最初からエージェントのアーキテクチャに組み込んでおく必要があります。

このセクションは、異なる3種類の構築担当者を対象としています。

- **B2C SaaSの作成者**：消費者向けにAIエージェント（チャットボット、パーソナルアシスタント、レコメンドエンジン）を構築しているご担当者
- **B2B SaaSの作成者**：企業顧客向けにAIエージェント（ワークフロー自動化、分析、エンタープライズツール）を開発しているご担当者
- **企業内の開発者**：組織に固有のワークフローとプロセスのために社内向けAIエージェントを構築しているご担当者

上記のシナリオによって実装の詳細には若干の違いがあるかもしれませんが、認証やトークン管理、認可、人間による監視という、中心的なセキュリティパターンはすべてのシナリオに当てはまります。このソリューションの主眼は、スピードと革新性を犠牲にせずに、安全なエージェントを構築する上で欠かせない能力をすべての開発者に提供することです。

1. 認証：ユーザーアイデンティティの確立

セキュリティの境界を維持しながら、パーソナライズされたエクスペリエンスを提供するために、AIエージェントはユーザーを安全に識別する必要があります。絶対に理解しておかなければならないのは、認証の対象はエージェント自体ではなくユーザーであり、エージェントは認証されたユーザーの代理として動作するという点です。対話型チャットボットであれ、バックグラウンドで動作する処理であれ、エージェントには最新のアイデンティティプロバイダーとシームレスに連携する、信頼性の高い認証が必要です。

組織に必要な最重要機能は以下の通りです。

- 複数のアイデンティティプロバイダーで機能する**ユニバーサル認証**で、従来の認証情報とソーシャルログインの両方をサポートする
- OpenID ConnectとOAuth 2.0を使用し、**標準規格に準拠した認証**で、相互運用性とセキュリティを確保する
- **安全なトークンでユーザーのアイデンティティを伝達**し、エージェントが誰のために行動しているかを理解できるようにする
- 適切なタイムアウトとセキュリティ制御による**堅牢なセッション管理を実現する**
- **多要素認証**で、高度なセキュリティ保証が必要なシナリオに対応する

開発者エクスペリエンスの面では、数行のコーディングだけで統合が実現できて、一般的なフレームワークとシームレスに連携し、複雑なコールバックURLやセッション管理、トークン検証を自動的に処理できる必要があります。

例：あるカスタマーサポート用チャットボットは、Google SSO経由でユーザーを認証します。Sarahがログインするとエージェントはそのid情報を受け取り、セキュリティの境界を維持しながら、パーソナライズされた応答ができるようになります。

2. トークン交換：信頼ドメインの橋渡し

AIエージェントが複数のシステムやセキュリティドメインにまたがって動作するケースでは、さまざまな信頼境界線内のリソースへのアクセスが必要な場合が少なくありません。トークン交換によって、エージェントは、適切なスコープが設定されたアクセストークンを取得し、直接のドメイン外にあるリソースにアクセスしながら、ユーザーコンテキストと認可チェーンを維持できます。

組織に必要な最重要機能は以下の通りです。

- 単一の信頼ドメイン内のシナリオでは**標準的なトークン交換**を行い、エージェントが同じ認可サーバーに異なる種類のトークンやスコープを要求できるようにする
- 別々の信頼境界線をまたいだアクセスが必要なシナリオでは、**クロスドメインの信頼**を行う
- 信頼境界線を越えて**ユーザーのアイデンティティ**と認証コンテキストを**維持する仕組み**
- 異なるアイデンティティプロバイダー間で**信頼関係を検証する**
- **スコープの変換**で、ドメイン間の権限を正しくマッピングする
- **安全な認証情報の変換**で、機密性の高いトークンが転送中に公開されないようにする

単一の認可サーバー環境内で動作するエージェントの場合は、標準のOAuth 2.0トークン交換で、効率的に認証情報を管理できます。エージェントが組織の境界を越える必要がある場合には、クロスドメインの信頼で、この機能を信頼ドメイン全体に拡張できます。

標準のOAuth同意とクロスドメインの信頼の使い分け：同意に基づくフローとクロスドメインの信頼のどちらを選択するかは、導入モデルによって異なります。

B2Cのシナリオ：標準のOAuth同意を使用する

- 消費者向けアプリケーションで、エンドユーザーは自身のデータを所有する
- ユーザーは、あるアプリが別のアプリにアクセスするための権限を明示的に付与する（例：TravelBotにGoogleカレンダーへのアクセスを許可）
- ユーザーが自身のデータについて個人的な判断を下すため、同意画面がふさわしい
- **例：**食事計画用アプリが、ユーザーのフィットネストラッカーのデータへのアクセスをリクエストする

B2Bや従業員向けのシナリオ：クロスドメインの信頼を使用する

- 企業の環境で、IT管理者がアクセスポリシーを一元的に管理している
- Business-to-Business-to-Employee（B2B2E）のシナリオで、従業員が企業のポリシーに従って業務を行っている
- 個々のユーザーの判断ではなく、企業のポリシーによってアクセスが管理されるため、ユーザーの同意画面はふさわしくない
- 企業向けIdPが、アプリケーション間の信頼関係を仲介する役割を担う
- **例：**企業のセールス用エージェントがSalesforce CRMと社内の価格データベースの両方にアクセスする場合、従業員がこのアクセスに「同意」するわけではなく、ITポリシーによって管理される

重要ポイント：従業員とB2Bのコンテキストでは、クロスドメインの信頼によって、同意疲労が解消し、一元管理されたITガバナンスに準拠できます。組織がアプリケーション間の信頼関係を事前に確立しておけば、IdPがエンタープライズポリシーを適用するため、個々のユーザーがアプリ間の各やり取りに対して認可の判断をする必要はありません。

3. トークンボールト：安全なAPIアクセス管理

AIエージェントは頻繁に、ユーザーに代わってサードパーティAPI（Salesforce、Slack、Google Workspace）にアクセスする必要があります。トークンボールトを使えば、最新APIの認証方法として推奨されるOAuthアクセストークンを安全に保管・管理して、コードやログ、構成ファイルでのトークン漏洩リスクを排除できます。ボールトで他の認証情報タイプ（レガシーシステムに必要な個人用アクセストークンやAPIキーなど）も保護できますが、OAuthトークンは自動更新、きめ細かいスコープ設定、安全な失効をサポートしているので、OAuthトークンを既定のパターンとして使用すべきです。

組織に必要な最重要機能は以下の通りです。

- OAuthトークンは、保存時に強力な暗号化を施した**安全なボールトで保管する**
- **自動的なトークンライフサイクル管理**（有効期限が切れる前にトークンを更新してサービス中断を防ぐなど）
- **オンデマンドのトークン取得**で、アプリケーションコードに認証情報を公開しない
- さまざまな認証スキームで、**多様な種類のトークンをサポートする**
- **スコープ付きのアクセス**で、トークンが必要な権限のみを持つようにする
- 最新のAI開発フレームワークに適合する**統合パターン**

セキュリティの原則：トークンは決して、エージェントのコードやログ、構成ファイルに現れてはいけません。ボールトはトークンライフサイクル全体を自動的に管理し、一元化された監査可能な仕組みによって、認証情報の盗難や不正使用のリスクを大幅に軽減します。

エージェント出力における認証情報の保護：コードやログに認証情報が現れないようにするだけでなく、トークンやシークレットが決してエージェントの応答や出力に流出しないようにする必要があります。ボールトに保存された認証情報を使用してエージェントがAPIにアクセスすると、応答データに機密情報を含む場合もあります。しかし認証情報そのものは、エージェントのチャット応答や生成文書のほか、エンドユーザーの目に触れるいかなる出力にも含めてはなりません。応答がユーザーに届く前に漏洩を検出するために、出力フィルタリングと検証を実装してください。特に、コードスニペットや構成例、トラブルシューティングのガイドには認証情報が誤って含まれる恐れがあるため、これらを生成するエージェントではこの対策が極めて重要です。

例：あるセールス用AIエージェントが顧客のSalesforceアカウントにアクセスする必要があるとします。エージェントはSalesforceの認証情報を保存するのではなく、ボールトにトークンを要求します。ボールトは、適切なスコープが設定された新しいトークンを提供し、必要に応じて自動的に更新します。そしてエージェントはタスクを完了します。認証情報がエージェントのコードに現れることはありません。

4. データセキュリティ：RAG向けのきめ細かな認可

AIエージェントが検索拡張生成（RAG）を使用して質問に回答する場合、エージェントがアクセスできるのはユーザーがアクセス許可を持つデータだけにする必要があります。適切な認可がなければ、RAGシステムが偶然に機密情報を公開しないとも限らず、そうなれば重大なセキュリティの脆弱性を生み出してしまいます。

組織に必要な最重要機能は以下の通りです。

- **関係性に基づくアクセスコントロール**で、ユーザーとドキュメントの間の権限を定義する
- **認可の適用**はドキュメント検索の時点で、データがエージェントのコンテキストに入る前に行う
- **ベクトルデータベースの統合パターン**をRAGアーキテクチャで使用する
- ドキュメントやセクションのレベルで制御できる、**きめ細かなアクセス許可**
- クエリ処理中に**リアルタイムで認可**を評価する
- 階層型と属性ベースのポリシーなど、**複雑な権限モデルのサポート**

このパターンの仕組み
ドキュメントは埋め込みとともにベクトルデータベースに保存されます。認可システムは、ユーザーとドキュメントの関係を管理し、エージェントがコンテキストを取得する際には、ドキュメントがエージェントのコンテキストに入る前に認可フィルターでアクセス許可を検証します。LLMは、ユーザーが閲覧権限を持つデータのみを使用して応答を生成します。

例：ある財務用AIエージェントを従業員がレポート分析に使用しています。Aliceが第3四半期の業績について質問すると、ベクトルデータベースが関連する財務ドキュメントを検索します。LLMに渡す前に、認可フィルターがAliceのアクセス権を確認します。その結果Aliceには、全社の財務ではなく、所属部門のレポートだけが表示され、許可されていないデータの公開を防ぎます。

5. ツールがセキュリティを呼び出す：人間が介在する認可

AIエージェントはバックグラウンドで自律的に動作することが多く、タスクの完了に数分、数時間、あるいは数日かかることもあります。購買の承認や契約書の送信、アクセス権の付与といった重要なアクションに関しては、人間による監視が必要です。一方で日常業務に関しては、エージェントの自律性を損なわないようにすべきです。

組織に必要な最重要機能は以下の通りです。

- **非同期の認可パターン**で、長時間かかるエージェントワークフローに対応する
- **詳細情報付き通知の仕組み**で、承認者に対して処理のコンテキストを漏れなく提供する
- **バインディングメッセージ**で、金額や受信者、意図するアクションなど、重要な詳細情報を示す
- **承認ワークフロー**は、デスクトップからでなくても、モバイルデバイスやEメールからアクセス可能にする
- 対応されない場合には自動的に有効期限が切れる、**時間制限付き認可リクエスト**
- すべての承認決定とそのコンテキストを文書化した、**包括的な監査証跡**

このパターンの仕組み
開発者が、エージェントのどのアクションに人間の承認が必要かを特定します。エージェントが保護対象の操作を試みると、承認リクエストが適切な担当者に送信されます。その際、エージェントが実行しようとしている処理に関するコンテキストもすべて付与されます。承認者がリクエストを確認し、承認または拒否します。承認された場合、エージェントは認可を受け取り処理を続行します。拒否された場合、操作がブロックされ、エージェントはエラーを受け取ります。

例：ある調達用AIエージェントが、必要なソフトウェアライセンスを特定し購入準備を行います。5,000ドルの取引を完了する前に、調達マネージャーに通知を送信します。通知には、ベンダーや金額、理由が記載されています。マネージャーは、昼食時にリクエストを確認し、モバイルデバイスまたはEメールで承認します。そして、エージェントが購入を完了します。こうして、自動化とコントロールの両方を維持できます。

単一のコントロールプレーンから全エージェントを保護

開発中に個々のエージェントにセキュリティを組み込むことは不可欠ですが、1か所でAIエージェント群全体の可視化と制御を行い、ガバナンスを保つ仕組みも組織には必要です。このソリューションは、部門やユースケース、システムの至る所で稼働する莫大な数のエージェントを管理するという、企業の課題に対応します。

1. エージェントの登録：正規のアイデンティティを確立

それぞれのAIエージェントを正規のアイデンティティとして登録し、所有者と責任を明確にする必要があります。きちんと登録しなければ、組織は何も分からずに運用しているも同然です。このエージェントは誰が所有しているのか、何の権限があるのか、問題が発生した場合に誰が責任を負うのか、といった基本的な質問にも答えられません。

組織に必要な最重要機能は以下の通りです。

- 永続的な一意の識別子を備えた、各エージェントの**アイデンティティプロファイル**
- **所有関係マッピング**で、責任を持つチームや個人にエージェントを結び付ける
- **メタデータシステム**で、エージェントの目的やユースケース、ライフサイクルのステージを文書化する
- 人事システムや報告階層などの組織構造と**連携させる**
- **変更追跡**で、エージェントの構成と権限への変更を記録する
- **依存関係マッピング**で、エージェントと、エージェントがアクセスするシステム間の関係を可視化する

重要な理由：「AI at Work 2025」の調査によると、非人間アイデンティティの管理に関して、十分考え抜かれた戦略を持つ組織はわずか10%に過ぎません。登録して基礎となるアイデンティティレイヤーを作成すれば、そこから他のガバナンスとセキュリティ制御もすべて可能となります。登録なしでは、セキュリティチームがエージェントを把握しない状態になり、管理対象外のシャドーAIが生まれます。

例：サービスアカウントでSalesforceにアクセスしているAIエージェントをセキュリティ部門が見つけたら、そのエージェントを中央エージェントレジストリに登録し、セールスオペレーションチームに所有権を割り当て、その目的を「エンタープライズアカウントのための自動提案書作成」として文書化します。これにより責任が定義され、エージェントが予想外の動作をした場合の連絡先と修正責任者が明確になります。

2. アクセスコントロール：ポリシーに基づく認可

AIエージェントには、コンテキストとリスクにリアルタイムで適応する、きめ細かな認可が必要です。アクセスコントロールポリシーで、各エージェントが実行できる処理、タイミング、条件を決定し、最小権限を適用しながらエージェントの機能を有効にします。

組織に必要な最重要機能は以下の通りです。

- **ポリシーエンジン**で、エージェントのアイデンティティや操作のコンテキスト、リスクシグナルに基づいて権限を定義する
- **標準ベースの認証**フローで、最新プロトコルをサポートする
- **APIアクセス管理**で、エージェントと保護対象サービスとのやり取りを制御する
- **クロスドメインの信頼**機能で、エージェントが組織ドメインと信頼ドメインの境界を越えて、ユーザーコンテキストを維持しながら安全にリソースにアクセスできるようにする
- 時間、場所、動作パターンなどの複数の要素を考慮した、**動的なポリシー評価**
- 既存のIAMインフラストラクチャに接続するための**統合パターン**
- さまざまなセキュリティドメインに広がる**複雑なマルチプロバイダーアーキテクチャへの対応**

高度なパターン - 管理対象接続
組織には、エージェントがアクセス可能な認可サーバーと、要求可能な権限を定義できる能力が求められます。これには、どのスコープを自動許可するか、どのスコープに追加承認が必要か、どのスコープを許可しないかを定義するポリシーフレームワークも含まれ、それによって、明確に定義された境界の中でエージェントが動作するようにします。信頼ドメインを越えてリソースにアクセスする必要のあるエージェントの場合、クロスドメインの信頼で、これらの管理対象接続を拡張し、一元化されたポリシー制御を維持しながら、安全な組織間認可を実現します。

例：エージェントがアクセスを要求すると、認可システムは、定義されたポリシーと照らし合わせて要求を検証します。ごく普通の権限は自動的に付与できますが、機密性の高い操作には理由の提示または承認が必要となり、危険なアクションは即座に拒否されます。これらはすべて、人手を介さずにプログラム的に実行されます。エージェントがパートナー組織のAPIにアクセスしたり、クラウドからオンプレミスのシステムにアクセスしたりする必要がある場合、XAAでこのドメイン間アクセスを実現しつつ、中央コントロールプレーンで可視性とポリシー適用を維持できるようにします。

3. ライフサイクル管理：エージェントの包括的なガバナンス

AIエージェントには、人間の従業員と同じようにライフサイクルがあります。オンボーディング、アクティブな運用、役割の変更、そして最後に廃止です。ライフサイクル管理により、このような移行を自動化しながら、セキュリティ制御を維持できます。

組織に必要な最重要機能は以下の通りです。

- **自動プロビジョニングワークフロー**で、適切な初期権限を備えたエージェントアイデンティティを作成する
- **ロールベースのテンプレート**で、一般的な種類のエージェント向けにアクセスパターンを標準化する
- **ジャストインタイムのアクセス機能**で、一時的に昇格させた権限を自動的に失効させる
- **定期的なレビュープロセス**で、以前設定したアクセス許可が今も適切か検証する
- **プロビジョニング解除ワークフロー**で、エージェントの廃止時にすべてのアクセス権を体系的に削除する
- **変更管理システム**で、権限の変更とその承認を追跡する

ライフサイクルの全体像
エージェントのプロビジョニング時には、本来の目的に基づいて初期権限を付与します。運用中には、エージェントは常にアクセス権の検証を受けながらタスクを実行します。要件が変化すると、承認ワークフローで権限を調整します。定期的なレビューで、アクセスが正当な状態に保たれていることを確認します。エージェントの廃止時には、すべての権限が取り消されますが、監査証跡はコンプライアンスのために保持されます。

例：あるマーケティング用AIエージェントのプロビジョニング時に、Eメールプラットフォームと顧客データベースへのアクセス権を付与しました。6か月後の四半期レビューで、エージェントにデータベースへのアクセスが必要なくなった（ユースケースが変わった）ことが判明しました。アクセス権は自動的に取り消されます。キャンペーンが終了すると、エージェントはプロビジョニング解除されて、すべての権限が削除されますが、監査ログはコンプライアンスのために保持されます。

4. 特権付き認証情報：安全なシークレット管理

AIエージェントはシステムにアクセスするために、特権付き認証情報を必要とするケースがよくあります。具体的には、APIキーやデータベースのパスワード、サービスアカウントの認証情報、証明書などです。認証情報の管理が不十分だったり、キーをコードに埋め込んだり、シークレットをローテーションしなかったりすると、攻撃者に悪用されかねない重大なセキュリティリスクを生み出します。

組織に必要な最重要機能は以下の通りです。

- **安全なボールトストレージ**を用いて、強力な暗号化で保管時の認証情報を**保護する**
- **自動ローテーションをスケジュール**して、定期的に認証情報を更新する
- **ジャストインタイムのプロビジョニングパターン**で、認証情報の漏洩期間を最小限に抑える
- キーベースや証明書ベースの仕組みなど、**多様な認証方法をサポートする**
- 更新や配布などの、**証明書のライフサイクル管理を自動化する**
- シークレットがコードやログ、構成ファイルに現れないよう**厳格に分離する**
- 外部シークレット管理システムの統合パターン

セキュリティへの影響
盗まれた認証情報を攻撃者が悪用するケースが多いため、認証情報の自動ローテーションによって長期間有効なシークレットをなくせば、攻撃対象領域を大幅に縮小できます。

例：あるデータパイプライン用エージェントが使用するデータベースの認証情報は、30日ごとにローテーションされます。ローテーション時には、人間の介入やサービスの中断なしに、エージェントが自動的に新しい認証情報をボールトから取得します。セキュリティ上の一番の利点は、認証情報が初めからログやコードに現れないことです（厳密な分離）。しかし、何らかの形で認証情報が漏洩した場合でも、自動ローテーションによって漏洩の期間が最大30日間に制限され、無期限に有効な状態にならないため、攻撃対象領域を大幅に縮小できます。

5. エージェントの検出：シャドー AIの発見

組織は、見えないものを保護できません。エージェントの検出で、環境内で動作するすべてのAIエージェントを可視化します。IT部門の承認やセキュリティレビューなしに各部門が導入したシャドー AIも対象になります。

組織に必要な最重要機能は以下の通りです。

- **自動検出メカニズム**によって、クラウドとSaaSプラットフォーム全体で人間以外のアカウントを特定する
- **シャドー AIの検出**で、正式な管理プロセスを経ずに導入されたエージェントを発見する
- **包括的な認証情報インベントリ**で、どのエージェントがどのシステムにアクセスできるかを可視化する
- 権限、アクティビティのパターン、セキュリティリスクレベルに基づいた**リスクスコアリング手法**
- **構成の分析**で、設定ミスや過剰な権限を持つアカウントを洗い出す
- クラウドプラットフォーム、アイデンティティプロバイダー、セキュリティツールと接続するための**統合パターン**

検出の方法

エージェントを効果的に検出するには、複数の手法を組み合わせます。具体的には、APIの使用パターンを分析して人間以外の動作を特定する、認証ログの相互関係を確認しサービスアカウントのアクティビティを検出する、クラウドのリソースを精査してエージェントの導入を検出する、ネットワークトラフィックを監視してエージェントからサービスへの通信を特定する、行動分析を使用してエージェントを人間のユーザーと区別する、などの手法があります。

重要な理由：「AI at Work 2025」の調査によると、91%の組織がすでにAIエージェントを利用していますが、十分考え抜かれたガバナンス戦略を持つ組織はわずか10%です。導入とガバナンスの間にギャップがあれば、シャドー AIエージェントがセキュリティ制御なしで動作し、セキュリティチームがその存在すら知らないリスクにつながる可能性があります。

6. エージェントのユニバーサルログアウト：迅速な脅威対策

認証情報の侵害、不審な挙動、ポリシー違反といった脅威が検知された場合、組織には、あらゆるシステムでエージェントのアクセスを即座に失効できる能力が求められます。エージェントのユニバーサルログアウトは、このような緊急時の「キルスイッチ」として機能しつつ、詳細な監査ログも維持します。

組織に必要な最重要機能は以下の通りです。

- すべてのアクティブなエージェントセッションとトークンを**即時に失効させる仕組み**
- **システム間の伝達**によって、連携する全アプリケーションにログアウトを確実に反映させる
- **緊急時の認証情報ローテーション**で、侵害された可能性のあるシークレットを置き換える
- **脅威封じ込めワークフロー**で、それ以上の不正なアクションを防止する
- **フォレンジック保全**で、インシデント後の調査のために、完全な監査ログを維持する
- セキュリティ運用センターやインシデント対応システムとの**連携**

行動分析による脅威検知
効果的なユニバーサルログアウトを支えるのは、堅牢な脅威検知です。脅威検知で必要なのは、エージェントの通常挙動のベースライン確立や、リクエストの量とタイミングの異常検出、行動パターンに基づくリスクスコアの算出、リスクがしきい値設定を超えた際の自動対応の発動、不審なアクティビティをリアルタイムでセキュリティチームに通知する機能です。

例：あるカスタマーサービス用AIエージェントは、普段1日に10 ～ 15件の顧客レコードにアクセスします。ところが突然、10分間で500件にアクセスしました。明らかな異常です。行動分析がこの逸脱を検知し、自動的にユニバーサルログアウトをトリガーしてそのエージェントの全アクセスを無効化するとともに、セキュリティチームに通知します。調査の結果、API認証情報が盗まれていたことが判明しました。攻撃は数分以内に封じ込められ、データベース全体のデータ流出を防ぐことができました。

連携して機能する仕組み

ご紹介した2つのソリューションは、設計段階から各エージェントを安全に保護し、単一のコントロールプレーンで全エージェントを守ることを目的としています。これらは統合が必要な別個のソリューションではありません。Oktaでは、AIエージェント向けの統合アイデンティティプラットフォームの中で、相互に補完し合う機能として提供しています。

開発段階では、各エージェントにコードの1行目から徹底してセキュリティを組み込みます。認証によって、ユーザーのアイデンティティを確立し、エージェントに誰の代理として動作しているのかを認識させ、セキュリティ境界を適切に維持します。トークンボールトによって、エージェントのコードに認証情報を一切公開することなくAPIアクセスを管理し、トークンの更新やライフサイクル管理を自動的に行います。ユーザーレベルのアクセス権を尊重するきめ細かな権限設定を行うことで、認可をコントロールし、不正なデータアクセスを防ぎます。人間が介在するヒューマンインザループの承認によって、重要なアクションを監視し、エージェントを自律的に動作させながら、機密性の高い操作をコントロールできるようにします。

実際の構成例を示します。

- **Universal Loginによるユニバーサル認証** – AIエージェントが関与する環境でも、ソーシャルプロバイダー、パスワードレスのフロー、MFAでシームレスなユーザー認証を実現します。
- **Token Vaultによる安全なAPIアクセス** – サードパーティAPI向けのOAuthトークンを保管・管理し、自動的にトークンを更新して、認証情報をエージェントのコードに公開しません。
- **Auth0 FGAによるきめ細かな認可** – 関係性に基づくアクセスコントロールをRAGシステムに実装し、ユーザーが閲覧権限を持つドキュメントのみをエージェントが取得できるようにします。
- **非同期認証によるヒューマンインザループの認可** – CIBA（Client-Initiated Backchannel Authentication）とRAR（Rich Authorization Request）を使用して、重要なエージェントアクションに対してモバイルやEメールで承認を要求します。

実運用フェーズのITチームとセキュリティチームは、コントロールプレーンを用いて全エージェントを継続的に監視・統制します。ディスカバリー機能によって、シャドーAIも含め、その環境で稼働しているすべてのエージェントを検出できるため、セキュリティチームは自社のAI環境全体を漏れなく可視化できます。登録によって、明確な所有者を定めた正規のアイデンティティを作成し、エージェントのアクションに対して責任を負うチームや個人が各エージェントに割り当てられた状態にします。

アクセスガバナンスによって、最小権限のポリシーを動的に適用し、コンテキストやリスクに応じてリアルタイムに権限を調整します。脅威検知によって、行動分析を通じて異常を特定し対応することで、攻撃が成功する前に自動的に封じ込めることができます。

実際の構成例を示します。

- **Universal Directoryのエージェント登録** – すべてのエージェントを、所有者、責任、メタデータの可視性を備えた正規のアイデンティティとして登録します。
- **Identity Security Posture Managementによるエージェントの検出** – 管理下のAIエージェントとシャドー AIエージェントの双方を環境全体で検出し、盲点を排除します。
- **Okta Identity Governanceによるアクセスコントロールとライフサイクル管理** – ポリシーに基づくライフサイクルガバナンスを通じて、エージェントの権限を定義、レビュー、認定します。
- **Okta Privileged Accessによる特権付き認証情報のボールト管理** – 高い権限を必要とするエージェントの認証情報を安全に保管・ローテーションして、攻撃対象領域を縮小します。
- **エージェント向けUniversal Logout** – リスクや侵害が検知された場合、システム全体でエージェントのセッション、トークン、認証情報を即時に無効化します。

これらのソリューションの連携によって、包括的なセキュリティのフレームワークを構築できます。セキュアな開発プラクティスに基づいて構築したエージェントは、コントロールプレーンから自動的に検出可能であり、追加の統合作業なしで即座に可視化できます。コントロールプレーンは、エージェントの認証とリソースへのアクセスに関するポリシーを適用し、開発時のセキュリティ制御を実行時のガバナンスへと拡張します。セキュアな開発で実現されたエージェントのアクションを、行動分析が継続的に監視し、エージェントが期待されるパターンから逸脱した場合に検出します。ガバナンスのワークフローはエージェントのアイデンティティとアクセス対象リソースの双方に適用されるため、エコシステム全体に一貫してポリシーを適用できます。

組織は、セキュアな開発とライフサイクル管理のどちらかを選ぶ必要はありません。一体化した戦略の一環として両方を実践すべきです。以下のリファレンスアーキテクチャは、実際のエンタープライズ環境で、この統合アプローチがどのように機能するかを示しています。

リファレンスアーキテクチャ：統合プラットフォームの活用

統合プラットフォームが、開発段階とライフサイクル管理の両面でAIエージェントをどのように保護するかを示すために、エンタープライズ向けセールス用エージェントを例に考えてみましょう。このエージェントはセールス担当者を支援するために、顧客調査の自動化や提案書の作成、CRMデータへのアクセス、フォローアップ会議のスケジュール設定を行います。 この例は、セキュアバイデザインの開発と一元的なライフサイクル管理という2つのソリューションが、どのように連携して機能するかを示しています。

このエージェントは、MCP（Model Context Protocol）を使用して、ユーザーの権限を尊重しながら、複数のソースのコンテキストに安全にアクセスします。また、クロスドメインの信頼のためのトークン交換（ID-JAG）を利用し、ライフサイクル全体を通じて包括的なガバナンスの下で運用されます。

アーキテクチャの概要


セールス担当者（ユーザー認証レイヤー）
1. ユーザーが認証し、エージェントが起動される
セールス用AIエージェント
エージェントのコンポーネント
LLM（Bedrock API）
MCPクライアント
OAuth2
トークンキャッシュ
OAuth2クライアント
認証情報（uDAP）
LLM APIキー（ボールトで保管）
MCPクライアントの
アイデンティティ（OAuth2）
ライフサイクルの状態：プロビジョニング済み・ガバナンス管理下
• 一意のIDを付与してコントロールプレーンに登録済み
• 所有者：セールスオペレーションチーム
• 認証情報：90日ごとに自動でローテーションされる
• 行動ベースライン：設定済み・監視中
セキュアバイデザインのレイヤー
• ユーザー認証
• Token Vault
• 非同期認証
• データセキュリティ（FGA）
コントロールプレーンの管理
• ディスカバリー
• 登録
• 認可
• ガバナンス
エンタープライズ認可サーバー
OAuth 2.0トークンの発行と検証
2. プラットフォームがユーザーの
アイデンティティを検証する
3. プラットフォームがOAuthトークンを発行する
4. エージェントがエンタープライズリソース（MCPサーバー +従来のAPI）にアクセスする
MCPリソースサーバー（エンタープライズ認可サーバーによって保護）
カレンダー用MCPサーバー
IdPでOAuthトークンを検証する
ドキュメント用MCPサーバー
IdPでOAuthトークンを検証する
CRMデータ用MCPサーバー
IdPでOAuthトークンを検証する
従来のAPIリソース
セールスCRM
(OAuth2)
Google Workspace
(OAuth2)
社内価格DB
(ID-JAG)
Eメールシステム
(OAuth2)
顧客データ
• アカウントの詳細
• 通信
• 価格履歴
• 商談・会議履歴

このプラットフォームは、MCPの実装を安全に保護するための機能を提供します。MCPサーバーはOAuth 2.0のリソースサーバーとして機能し、認証と認可をエンタープライズ認可サーバーに委任します。

MCPセキュリティモデル

エージェント（MCPクライアント）は、エンタープライズ認可サーバーにOAuth 2.0クライアントとして登録され、クライアント認証情報（client_idとclient_secret、または証明書ベースの認証情報）を受け取ります。MCPリソースにアクセスする前に、エージェントは適切なスコープ（例：mcp:crm:read、mcp:docs:read、mcp:calendar:read）を持つアクセストークンを取得します。エージェントがcrm://contacts/acme-corpのようなリソースを要求すると、MCPサーバーは認可サーバーに対してアクセストークンを検証し、署名の有効性、有効期限、オーディエンス、必要なスコープを確認した上でリソースを提供します。

こうすれば、MCPサーバーの開発者が独自の認証・認可ロジックを構築する必要はなくなります。代わりに、標準的なOAuth 2.0のトークン検証を用いて、プラットフォームが発行したOAuthトークンを検証するだけで済みます。プラットフォームは、ユーザー認証や、トークンのライフサイクル、スコープ管理、Fine Grained Authorization（FGA：きめ細かな認可）のチェックを担い、すべてのMCPサーバーに一貫したセキュリティポリシーを適用して、システムログにすべての監査証跡を保存します。

このリファレンスアーキテクチャは、プラットフォームがMCPベースのコンテキスト取得を安全に保護しつつ、AIエージェントのライフサイクルをエンタープライズレベルで統制・管理する仕組みを示しています。

詳細なフロー：MCPを用いたエージェント支援による提案書生成

フェーズ1

ディスカバリーと登録（コントロールプレーン ― 検出／プロビジョニング）

ステップ1.1：シャドー AIの検出

- セールスチームが、IT部門の承認を得ずにプロトタイプのエージェントを導入した
- サービスアカウントの認証情報を使用してSalesforceにアクセスしている当該エージェントをOktaが検出する
- セキュリティチームが、未管理のAIエージェントに関するアラートを受信する
- リスクスコア：高（特権付きアクセス、所有者不明、ガバナンス管理外）

ステップ 1.2：エージェントの登録

- セキュリティチームが、当該エージェントをOktaで正規のアイデンティティとして登録する
- エージェントのプロファイルを作成し、一意の識別子を割り振る：`sales-agent-prod-001`
- 所有者はセールスオペレーションチーム（John Smith、セールスオペレーション担当VP）に割り当てる
- 目的は「提案書の自動生成と顧客調査」と記録する
- 認証情報は安全なボールトへ移行し、90日ごとのローテーションポリシーを適用する

結果：エージェントは、シャドー AIから明確な責任を備えた管理対象アイデンティティに移行する。

フェーズ2

ユーザー認証（セキュアバイデザインのレイヤー）

ステップ2.1：セールス担当者の認証

- Sarah（セールス担当者）が午前9時に営業ポータルへログインする
- Auth0 Universal Loginが認証オプションを提示する
- SarahはGoogle SSO（ソーシャルログイン）を使用して認証する
- Auth0がGoogle IdPに対して認証情報を検証する
- Sarahのプロファイル情報と認証クレームを含むIDトークンが発行される

ステップ2.2：エージェントへのコンテキストバインディング

- エージェントは、Auth0から認証済みユーザーのコンテキストを受け取る
- IDトークンには、以下のように標準的なOIDCクレームが含まれる

```
{
  "iss": "https://acmecorp.auth0.com/",
  "sub": "google-oauth2|108204567890123456789",
  "aud": "sales-agent-client-id",
  "exp": 1730480000,
  "iat": 1730477400,
  "name": "Sarah Johnson",
  "email": "sarah.johnson@acmecorp.com",
  "email_verified": true
}
```

注：`role`や`territory`などの追加のカスタムクレームは、Auth0 Actionsを使用して追加できます。

- エージェントはユーザーが「誰であるか」を把握し、その代理として動作できるようになる

フェーズ3

MCPによるコンテキスト取得（データセキュリティ +認可）

ステップ3.1：ユーザーからのクエリ

Sarahが「Acme Corpの購入履歴と現在のニーズに基づいて、提案書を作成してください」と依頼する。

ステップ3.2：MCPによるコンテキストのディスカバリー

エージェントはMCPクライアントを使用して、利用可能なコンテキストソースを検出します。
MCPサーバーは、以下のリソースURIを介して構造化リソースを公開します。

```
crm://contacts/acme-corp
docs://proposals/templates
calendar://availability/sales-team
pricing://enterprise-tier
```

これは、埋め込み（RAG）を用いたセマンティック検索ではなく、**MCPによる構造化コンテキストの取得**です。MCPは、定義されたスキーマとリソースURIに基づいて特定のリソースへ直接アクセスする仕組みを提供します。エージェントは名前やパスを指定して特定のリソースを要求し、MCPサーバーは構造化データを返します。

ステップ3.3：Auth0 FGAによる認可チェック

エージェントは、Sarahがアクセス可能なリソースを判定するため、
Fine Grained Authorization（FGA）サービスに問い合わせます。

- FGAは、各MCPリソースに対するリレーションシップタプル（ユーザーとリソースの関係定義）を評価する
  - ✓ `user:sarah`は`crm://contacts/acme-corp`に`read`のアクセス権を持つ
  - ✓ `user:sarah`は`docs://proposals/templates`に`read`のアクセス権を持つ
  - ✗ `user:sarah`は`pricing://executive-discounts`にアクセス権がない
- 許可されたリソースだけが、コンテキスト検索に含まれる
- こうして最小権限アクセスを適用することで、データ漏洩を防止する
- それぞれのMCPリソース取得要求は、取得前にSarahの権限に照らして検証される

これは、アーキテクチャ図の「セキュアバイデザインのレイヤー」にある「データセキュリティ（FGA）」に対応します。

ステップ3.4：Auth0 Token Vaultからのトークン取得

エージェントは、外部システムにアクセスするために、Token VaultにOAuth2トークンを要求します。

- エージェント：「Salesforce CRMにアクセスしてAcme Corpのアカウントのデータを取得する必要があります」
- Token Vaultは、エージェントに許可された統合設定に照らして要求を検証する
- Token Vaultは、スコープ付きの有効なSalesforce用アクセストークンを返却する
- トークンのスコープは、顧客データの読み取り専用アクセスに限定されている
- エージェントは、このトークンを使用してSalesforce API経由でCRMデータを取得する

Token Vaultの機能

- 安全な認証情報ストレージ（ハードコーディングされたトークンがなくなる）
- トークンの有効期限が切れたときに、トークンを自動的に更新
- すべてのトークンアクセスに関する監査ログ
- スコープ制限付きトークン（最小権限）

ステップ 3.5：コンテキストの組み立て

エージェントは、認可されたソースから完全なコンテキストを組み立てます。

CRMから（Token Vault→Salesforce経由）

- Acme Corpの連絡先：CTO Jennifer Martinez
- 購入履歴：AIトレーニングサービスに28万ドル（2024年）
- 現在の契約：サポート契約は2026年3月に失効

ドキュメントライブラリから（MCP経由）

- エンタープライズ向け提案テンプレート（承認済みバージョン）
- 現在の価格体系が記載された製品カタログ
- 標準の利用規約

カレンダーから（MCP経由）

- Sarahがフォローアップコールに対応できる日時
- 導入サポートに関するセールスチームの対応可能状況

含まれないもの（認可されていないため）

- 役員向け割引価格（Sarahにアクセス権がない）
- 他案件の機密交渉メモ
- 社内のコスト構造データ

このようにして、エージェントは提案書を作成するために必要なすべての認可済みコンテキストを取得しました。

フェーズ4

ID-JAGによるクロスドメイン認可（トークン交換）

ステップ4.1：社内価格データベースへのアクセス

- エージェントは、pricing.acmecorp.internalにある社内価格データベースにアクセスする必要がある
- これはメインのアイデンティティプラットフォームとは別の認可ドメイン
- この価格システムには、ID-JAGトークンを要求する独自の認可サーバーがある

ステップ4.2：ID-JAGによるトークン交換

エージェントは、トークン交換のためにIDトークンを認可サーバーへ送信します。**リクエストは以下の通りです。**

```
POST /oauth2/token
Host: acmecorp.okta.com

grant_type=urn:ietf:params:oauth:grant-type:token-exchange
&subject_token=<Sarah's ID Token>
&subject_token_type=urn:ietf:params:oauth:token-type:id_token
&requested_token_type=urn:ietf:params:oauth:token-type:id-jag
&audience=https://pricing.acmecorp.internal
&scope=pricing:read
```

以下が行われています。

- エージェントは、SarahのIDトークンをID-JAGトークンと交換する
- ID-JAG（Identity Assertion JWT Authorization Grant）は、暗号署名されたトークン
- ID-JAGトークンは、価格データベースの認可サーバー宛てに発行される
- こうして、ユーザーコンテキストを維持したままクロスドメイン認可が可能になる

ステップ4.3：認可サーバーによる検証

認可サーバーは、以下の検証チェックを実行します。

- IDトークンの検証：IDトークンの署名とクレームを検証する（信頼関係は事前に確立済み）
- 管理対象接続の確認：エージェントが価格認可サーバーに対して持つ管理対象接続を検証する
- 管理対象接続により許可されるスコープは以下のとおり
  - ✓ 許可されたスコープ：`pricing:read`
  - ✗ **拒否されたスコープ：** `pricing:write`、`pricing:admin`

ID-JAGトークンの発行：認可サーバーは、Sarahのユーザーコンテキストを保持したID-JAGトークンを発行する

ID-JAGトークンのクレーム


```
{
 "iss": "https://acmecorp.authorization-server.com",
 "sub": "sarah.employee@acmecorp.com",
 "aud": "https://pricing.acmecorp.internal",
 "client_id": "sales-ai-agent",
 "jti": "9e43f81b64a33f20116179",
 "scope": "pricing:read",
 "exp": 1698583800,
 "iat": 1698580200,
 "auth_time": 1698580200,
 "amr": ["pwd", "mfa"]
}
```

ステップ4.4：リソースへのアクセス

エージェントは、ID-JAGトークンを価格データベースの認可サーバーに提示します。

- 価格認可サーバーは、アイデンティティプラットフォームが公開している公開鍵（JWKS）を用いて、ID-JAGトークンの署名を検証する
- 価格認可サーバーは以下を確認する
  - `aud`**クレームが自サーバーの発行元URLと一致していること**
  - `exp`**（有効期限）が切れていないこと**
  - `scope`**が許可権限の範囲内であること**
  - `iss`が信頼済みのアイデンティティプロバイダーであること
    - **アクセスが許可され、**エージェントは読み取り専用権限でエンタープライズ価格データを取得する
    - これで、エージェントは提案書作成に必要な価格情報への認可済みアクセス権を取得した

プラットフォームのトークン交換機能

このアイデンティティプラットフォームは、クロスドメイン認可のシナリオに対応するため、RFC 8693に基づくトークン交換をサポートしています。トークン交換により、AIエージェントは暗号署名されたID-JAGトークンを通じてユーザーコンテキストを保持したまま、異なる認可サーバーのリソースにアクセスできます。この機能は、開発者向け・エンタープライズ向けのいずれの導入形態でも、プラットフォーム全体で利用可能です。

フェーズ5

非同期認可（ヒューマンインザループ）

ステップ5.1：エージェントが承認の必要性を判断

- エージェントは、45万ドルの提案書を送信するには明示的なユーザー承認が必要であると認識する
- これにより、非同期認可ワークフローがトリガーされる
- エージェントは非同期認可リクエストを開始して、承認待ちの間は処理を一時停止する

非同期認可が必要な理由は、以下のとおりです。

- 高額な提案は、エージェントの自律的な権限の範囲を超えるため
- 社内ポリシーにより、10万ドルを超える提案には人間の承認が必要なため
- ビジネス上重要な意思決定に対する説明責任を確保するため
- エージェントによる不正なアクションを防止するため

ステップ5.2：CIBA認可リクエスト

エージェントは、CIBA（Client-Initiated Backchannel Authentication）認可リクエストを送信します。

リクエストには以下の内容が含まれます。

- **ユーザー識別子：**Sarahの社員ID
- **必要な権限：**`email:send`、`drive:write`、`crm:update`
- **実行しようとしているアクションのコンテキスト：**Sarahが確認する提案の詳細
- **コールバックエンドポイント：**承認後にトークンが送信される宛先

```
POST /bc-authorize
Host: acmecorp.authorization-server.com
scope=email:send drive:write crm:update
&login_hint=sarah.employee@acmecorp.com
&binding_message=Proposal Approval:
Acme Corp - $450,000
&client_notification_token=
8d67dc78-7faa-4d41-aabd-67707b374255
```

CIBAでは以下を実現できます。

- 非同期の承認ワークフロー（エージェントはブロックしない）
- 帯域外ユーザー認証（モバイルへのプッシュ通知）
- 承認リクエストにおける豊富なコンテキスト（提案内容の詳細情報）
- 安全なコールバックメカニズム（承認時にトークンを配信）

ステップ5.3：プッシュ通知

認可サーバーは、SarahのGuardianモバイルアプリにプッシュ通知を送信します。

リッチ通知メッセージには、次の内容が表示されます。

```
提案の承認が必要です
顧客：Acme Corp
金額：$450,000
製品：Enterprise AI Suite +サポート
受信者：cto@acmecorp.com、cfo@acmecorp.com
アクション：Eメールで提案書を送信し、Googleドライブに保存
[承認] [拒否]
```

リッチ通知の機能

- 承認が必要な操作に関する詳細なコンテキスト
- 顧客名、金額、対象製品の表示
- 透明性確保のための受信者一覧
- 明確なアクション内容の説明
- シンプルな承認／拒否インターフェイス

注： これはGuardianのリッチ通知機能であり、OAuth のRich Authorization Requests（RAR - RFC 9396）仕様とは異なります。この通知は、Sarahが十分な情報に基づいて承認判断を行えるよう、詳細なコンテキスト情報を提供します。

ステップ 5.4：ユーザー承認

Sarahは内容を確認し、リクエストを承認します。

- Sarahはモバイルデバイスで詳細を確認する
- 次の点を確認する
  - 顧客が正しいこと（Acme Corp）
  - 金額が正確であること（45万ドル）
  - 受信者が適切であること（CTOとCFO）
  - 製品が顧客ニーズに合致していること
  - Sarahは「承認」ボタンをタップしてリクエストを承認する

トークンの生成と配信

- 認可サーバーは、承認された権限を持つスコープ付きアクセストークンを生成する
- トークンには、Sarahが承認した以下の権限のみが含まれる
  `email:send` - 提案メール送信の権限
  `drive:write` - Googleドライブに提案書を保存する権限
  `crm:update` - Salesforceにアクティビティの履歴を記録する権限
  - トークンはCIBAのコールバックエンドポイント経由でエージェントに配信される
  - エージェントはトークンを受け取り、新たに発行されたトークンを使用して処理を続行する

セキュリティ上の利点

- 機密性の高い操作には明示的なユーザー承認が必要
- 時間制限付きトークン（アクション完了後に失効）
- スコープ制限付きトークン（承認された権限のみ）
- 漏れのない監査証跡（誰が・いつ・何を承認したか）

非同期認可トークンフロー（CIBA）— 技術詳細

CIBA（Client-Initiated Backchannel Authentication）フローによって、AIエージェントの操作に対する非同期のユーザー承認が可能になります。

Sarahがモバイルデバイスでリクエストを承認したときの処理は、以下の通りです。

- **承認の検証：**認可サーバーは、その承認判断がSarahの認証済みデバイスで行われたことを検証する
- **トークン生成：**認可サーバーは、承認された操作にスコープを限定した新しいアクセストークンを生成する
- **権限のスコープ設定：**トークンには、Sarahが承認した権限のみが含まれる（例：`email:send`、`drive:write`）
- **安全な配信:** トークンは、元のCIBAリクエストで指定された安全なコールバックエンドポイントを通じてエージェントに配信される
- **エージェントの実行：**エージェントはトークンを受け取り、承認された操作を実行する

以上の仕組みで、AIエージェントによる機密性の高い操作に対してヒューマンインザループの認可が確実に行われるようになり、エージェントがアクションを実行する前に明示的なユーザー承認が必要となります。

CIBAフローの利点

- **ノンブロッキング：**承認を待つ間、エージェントは接続を維持する必要がない
- **ユーザーフレンドリー：**Sarahはチャットボットではなく、モバイル端末から承認できる
- **安全性：**トークンはユーザーのブラウザではなく、安全なコールバック経由で配信される
- **監査可能：**承認リクエスト、ユーザーの判断、トークン発行の記録をすべて保持する
- **柔軟性：**さまざまな承認手段（プッシュ通知、SMS、Eメール）に対応

フェーズ6

複数システムでの実行（トークンボールト保管）

ステップ6.1：Googleドライブへの保存

- エージェントはAuth0 Token VaultからGoogle Workspace用トークンを取得する
- トークンはSarahのGoogleドライブへのアクセス権にスコープが限定されている
- エージェントは提案書を次の場所にアップロードする `Sales/Proposals/2025/Acme-Corp-Q1.pdf`
- ファイルの権限：Sarahのチームメンバーと直属の上司

ステップ6.2：Eメールの送信

- エージェントはToken VaultからGmail API用トークンを取得する
- エージェントはSarahのアカウントでEメールを作成する
  - 宛先：Acme CorpのCTOとCFO
  - 本文：正式な提案書のカバーレター（LLMが生成）
  - 添付：Googleドライブ上の提案書PDF
- EメールはSarahの署名付きで送信される

ステップ6.3：フォローアップのスケジュール設定

- エージェントはToken VaultからGoogleカレンダー用トークンを取得する
- エージェントは今後2週間のSarahの予定を確認する
- エージェントはAcme Corpの担当者に会議候補日時を提示する
- エージェントはカレンダーイベント「Acme Corp提案会議 - 30分」を追加する

Token Vaultの利点

- エージェントが実際のOAuthトークンを見ることはない
- すべてのトークンが有効期限前に自動更新される
- 認証情報がエージェントのコードやログに保存されることはない
- Auth0の認証情報とOktaのエージェントアイデンティティは完全に分離されている

フェーズ7

ガバナンスと監視（コントロールプレーン – 統制）

ステップ7.1：監査証跡

すべてのアクティビティは、網羅的な詳細情報とともにプラットフォームのシステムログに記録されます。

- エージェントの認証イベント
- トークン交換処理とスコープの付与
- すべてのシステムにおけるリソースアクセス試行
- 認可の判断（承認／拒否）
- ユーザー委任イベント
- ユーザーに代わって実行したAPI呼び出し
- フォレンジック分析のための、すべてのタイムスタンプとコンテキストメタデータ

ステップ 7.2：四半期ごとのアクセスレビュー

- ガバナンスワークフローが起動される：2025年第1四半期のアクセス認定
- John Smith（エージェント所有者）へのEメール：「sales-agent-prod-001のアクセス権を確認してください」
- アクセスレビューの結果、エージェントは以下のアクセス権を保有している
  - Salesforce CRMへのアクセス
  - Google Workspaceへのアクセス
  - 社内価格データベースへのアクセス
  - Eメールシステムへのアクセス
  - カレンダーシステムへのアクセス
- Johnは、すべてのアクセスが引き続き必要であることを確認する
- 認定の結果はOktaの監査ログに記録される

ステップ7.3：アクセス認定

- 四半期アクセスレビューで、エージェントが役割に適した権限を保有していることを確認する
- すべてのアクセスはエージェント所有者により正当性が確認され、承認済み
- コンプライアンス目的のため、認定結果が監査ログに記録される

フェーズ8

脅威検知（コントロールプレーン – 監視）

ステップ8.1：異常の検出

- **45日目：**エージェントが突然10分間で500件の顧客レコードにアクセス
- 行動分析により、ベースラインからの逸脱が検出される
- リスクスコアが上昇：通常 → 高
- 異常の種類：「通常と異なるデータアクセス量」

ステップ8.2：自動対応

- プラットフォームがエージェントのSalesforceへのアクセスを自動的にブロックする
- **グローバルトークンの取り消しが実行されて、**すべてのシステムで、有効なトークンが即座に無効化される
- セキュリティチームがリアルタイムでアラートを受信する
- Sarah（ユーザー）が通知を受信する：「セールス用エージェントが一時的に停止されました」
- 完全なアクセス遮断によって、それ以上の不正なアクティビティを防ぐ

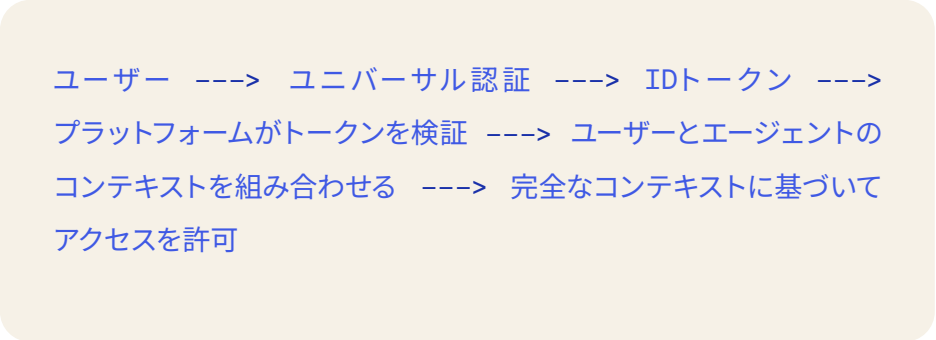
ステップ8.3：調査と修復

- セキュリティチームがシステムログを確認し、インシデントの範囲を把握する
- 根本原因を特定し、対処する

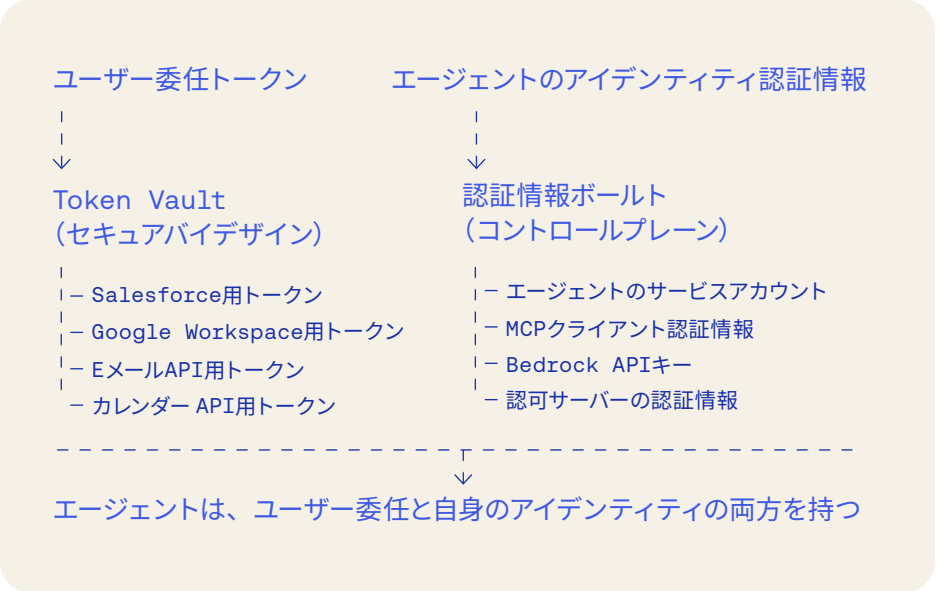
主要な測定基準
自動脅威検知と自動対応によって、攻撃の検出と阻止が行われた。

統合ポイント：統合プラットフォームのコンポーネント接続方法

1. 認証フロー



2. トークンのライフサイクル



3. 認可レイヤー

レイヤー 1：データセキュリティ（Fine Grained Authorization）

RAGのためのドキュメント単位の権限管理
「ユーザー Sarahはproposal-acme-2024を閲覧できるか?」

レイヤー 2：トークンボールト

SaaSツールに対するAPIレベルの権限管理
「ユーザー SarahのトークンはSalesforceにアクセスできるか?」

レイヤー 3：アクセスコントロール（コントロールプレーン）

エンタープライズリソースに対するシステムレベルの権限管理
「エージェントsales-agent-prod-001は価格データベースにアクセスできるか?」

レイヤー 4：トークン交換（ID-JAG）

ユーザーコンテキストを保持したクロスドメインの信頼
「Sarah（エージェント経由）はオンプレミスの価格システムにアクセスできるか?」

結果：複数の認可チェックポイントによる多層防御

4. MCP認可パターン

1. エージェントがMCPクライアント経由でコンテキストを要求する
2. MCPサーバーがリクエストを受信する
3. MCPサーバーが確認：エージェントは有効なトークンを持っているか?

トークン検証 ---> Auth0認可サービス

- エージェントのアイデンティティを検証する
- トークンスコープを確認する
- 権限を検証する

4. MCPサーバーが確認：ユーザーはデータへのアクセス権があるか?

権限チェック ---> Fine Grained Authorization

- リレーションシップタプル（関係定義）を評価する
- 認可されたドキュメントのみを返す

5. MCPサーバーが、認可されたコンテキストをエージェントに返す

本アーキテクチャが示す主要な設計原則

1. 関心の分離

- Auth0はユーザー認証とユーザー委任によるアクセスを担当する
- Oktaはエージェントのアイデンティティとライフサイクル管理を担当する
- MCPは共通方式によるコンテキスト取得を担当する
- 各システムがそれぞれの強みを発揮する

2. 多層防御

- 複数の認可レイヤーにより単一障害点を防止する
- FGAがドキュメントをフィルタリングし、Token VaultがAPIアクセスの関門として機能し、Oktaがシステムを統制する
- 1つのレイヤーが突破されても、他のレイヤーで防御する

3. 最小権限

- エージェントには各タスクに必要最小限の権限のみを付与する
- トークンのスコープは特定のAPIやアクションに限定する
- 自動的に有効期限が切れる時間制限付きアクセス
- ジャストインタイムプロビジョニングで、スタンディング特権を削減する

4. ユーザーコンテキストの保持

- ID-JAGトークン交換により、信頼境界を越えてもユーザーアイデンティティを維持する
- エージェントのアクションは、常に特定のユーザーまで追跡可能
- 認可の判断では、エージェントのアイデンティティだけでなくユーザーのコンテキストも考慮される
- 監査証跡には「エージェントX」と「ユーザー Yの代理」の両方が記録される

5. 継続的な監視

- 行動ベースラインにより異常を検知する
- リアルタイムの脅威対策で攻撃を阻止する
- 包括的ログによってフォレンジック分析が可能になる
- 自動修復によって対応時間が短縮される

アーキテクチャの比較：従来のアプローチと統合プラットフォーム

機能	統合プラットフォームがない場合	統合プラットフォームがある場合
エージェントの発見	手作業のスプレッドシート、シャドー AI検出機能なし	自動ディスカバリー、すべてを可視化
認証情報の管理	キーをコード内にハードコーディング、ローテーションされない	ボールトでの保管とローテーション
ユーザー認証	独自の認証・認可コード、パスワードストレージ	ユニバーサル認証、ソーシャルSSO
APIアクセス	トークンを構成ファイルに保存	トークンをボールト保管して自動的に更新
クロスドメインアクセス	各システム個別の認証・認可	ユーザーコンテキストを保持したID-JAGトークン交換
人間による承認	独自のポーリング、モバイルのサポートなし	CIBAによるモバイル通知やEメール送信
ドキュメントの権限	アプリケーションレベルのチェック、整合性に欠ける	関係性に基づく制御で、きめ細かな認可
アクセスレビュー	四半期ごとにスプレッドシートで手作業	自動化された認定ワークフロー
脅威検知	インシデント発生後にログを分析	リアルタイムの行動分析
監査証跡	複数のシステムに分散	統合されたシステムログ
インシデント対応	手作業での調査と修復	自動的なブロックとトークン失効
MCP認可	各MCPサーバーに独自の認可ロジック	プラットフォームで検証される標準OAuth2

このアーキテクチャは、統合アイデンティティプラットフォームがAIエージェントのセキュリティを包括的に実現する仕組みを示しています。具体的には、安全な開発（認証、トークン管理、認可、人間による監督）とエンタープライズ向けライフサイクル管理（ディスカバリー、登録、ガバナンス、脅威検知）の双方をカバーし、さらにMCPで、常にアイデンティティと認可の制御を尊重しながら共通方式のコンテキスト取得を実現します。

まとめ：統合プラットフォームでAIエージェントのセキュリティを包括的に実現

AIエージェントの革命は、すでに始まっています。現在、91%の組織がすでにAIエージェントを活用しており、この割合は今後さらに増加していくでしょう。しかし、セキュリティやガバナンスの整備が導入スピードに追いついていないため、AIエージェントで実現できるはずのビジネス価値を台無しにする重大なリスクが生まれています。

この課題に対応するには、相互に関連する2つの問題を同時に解決する必要があります。

セキュアバイデザインでエージェントを保護：開発段階で、適切な認証、認可、トークン管理、データアクセスコントロールを最初から組み込みます。

単一のコントロールプレーンから全エージェントを保護：エージェント群全体に対して、ライフサイクルを通してディスカバリー、プロビジョニング、ガバナンス、脅威検知を行います。

この両面に対応する組織は、包括的なセキュリティを実現できるようになります。

- 認証、トークンボールト、認可、人間による監督を組み込んだ**セキュアな開発手法**
- シャドー AIを含む、すべてのAIエージェントを**漏れなく可視化**
- 登録からプロビジョニング解除までの**適切なライフサイクル管理**
- アクセスレビューや認定をはじめとする**包括的なガバナンス**
- 行動分析と自動対応による**リアルタイムの脅威検知**
- 漏れのない監査証跡とポリシー適用の徹底による**規制コンプライアンス**

詳細を見る

安全なAIエージェントの構築
安全なエージェントを開発するためのドキュメントとクイックスタート。

OktaのAIエージェントライフサイクル保護のアプローチについて詳細を見る
Oktaが提供する、AIエージェントを大規模に管理するためのエンタープライズ向けガバナンスとコントロールプレーン機能についてご紹介します。

「AI at Work 2025」の調査によると、リーダーの85%がAIの導入にIAMが不可欠だと回答している一方で、非人間アイデンティティの管理に関して、十分考え抜かれた戦略を持つ組織はわずか10%に過ぎませんでした。この重大なギャップを解消できるのが、今回ご紹介した統合プラットフォームアプローチです。

侵害やコンプライアンス違反が発生してから適切なAIエージェントセキュリティを導入するのではなく、今すぐ行動を始めましょう。開発段階からエージェントを保護し、エージェント群全体に対する一元的な統制を確立することが重要です。

両機能を備えた統合プラットフォームを、Oktaのアイデンティティソリューションを通じて利用できます。世界中の組織がすでにこのアプローチを採用し、Auth0 for GenAIによるセキュアな開発と、Okta Identity Platformによるエンタープライズ向けライフサイクル管理を組み合わせることで、AIエージェントを大規模かつ安全に展開しています。

ホワイトペーパー

AIエージェントを開発からエンタープライズ拡張まで保護する

okta
The World's Identity Company™

Okta Inc.
100 First Street
San Francisco, CA 94105
info@okta.com
1-888-722-7871