Agentic AIのための アイデンティティ管理

AIエージェントの世界における認可、認証、セキュリティの挑戦

編集責任者：Tobin South
2025年10月

エグゼクティブサマリー

AIエージェントの急速な普及は、認証、認可、ID管理における早急な対応を要する課題をもたらしている。現在のエージェント中心のプロトコル（MCPなど）は、認証と認可におけるベストプラクティスを明確にする必要性を浮き彫りにしている。将来を見据えると、高度で自律的なエージェントを求める気運は、拡張制の高いアクセス制御、エージェント中心のアイデンティティ、AIが行う作業の分類と整理、権限の委譲など、複雑で長期的な問題を突き付けている。このホワイトペーパーは、AIエージェントとアクセス管理が交差する位置にいる関係者向けである。本稿は、今日のエージェントの安全性を確保するために既に利用可能なリソースの概要と、将来広範囲に稼働するであろう自律システムに極めて重要な、基礎的な認証、認可、アイデンティティの問題に対処するための戦略的アジェンダを提示する。

今日のフレームワークは、シンプルなAIエージェントに対応している：

- **AIエージェントは従来のソフトウェアとは根本的に異なる。**外部サービスに対して自律的な行動をとり、決められた命令を単に実行するのではなく、リアルタイム、非決定的、柔軟な振る舞いを示す。
- **既存のOAuth 2.1フレームワークは、AIエージェントと共に使用される場合、** 同期的なエージェント操作（例えば、企業エージェントが内部ツールにアクセスしたり、利用者がAIツールを介してサービスにアクセスしたりする）を伴う**単一の信頼関係にあるドメイン内ではうまく機能する**が、クロスドメイン、高度に自律的、または非同期的な状況や、エージェントが同時に複数の人から委譲された権限を実行する必要がある場合は、うまく機能しない可能性がある。
- **モデル・コンテキスト・プロトコル（MCP）は 、**エージェントを開発する際に、言語モデルを外部のデータソースやツールに接続するための重要なフレームワークとして**積極的に採用されている 。**但し、他にも関数呼び出し（ツールの使用）やエージェント間（A2Aなど）通信プロトコルなどのアプローチがある。これらもサポートするべきである。
- **企業SSOとSCIMプロビジョニングは、エンタープライズエージェントの利用を可能**にし、エージェントのライフサイクルの一元管理、及びさまざまなAIエージェントのユースケースに対する権限とアクセスのガバナンスを促進するのに役立つ**。**
- **企業のセキュリティプロファイルは、AIを安全に導入するためのベースラインを提供する。**リスクを緩和するために、本稿では、AIエージェントが既存の アイデンティティ標準仕様の厳格で相互運用可能なプロファイルに準拠することを推奨する。IPSIE（Interoperability Profiling for the Secure Identity in the Enterprise）のようなワーキンググループはこれに関するガイダンスを提供しており、強固な管理が行き届いていることに確証を得られるようにすることで、組織が安心してAIを採用できるようにしている。
- **ユーザ中心の同意モデルは、消費者向けエージェントの土台である。**サードパーティの消費者サービス（電子メール、ソーシャルメディア、金融データなど）に接続するエージェントにとって、確立されたOAuth 2.1のユーザ同意フローは、権限移譲のための主要なメカニズムである。ユーザの信頼を得るには、透明性と明確なスコープ定義が重要である。

将来に向けての重大な課題：

- **エージェントのアイデンティティの断片化は避けるべきである。**ベンダーは、独自のエージェント型アイデンティティシステムを開発する可能性があるが、この場合、単発の統合を繰り返さざるを得なくなるため、開発者の速度が低下することになる。また、それぞれ異なるリスクと脆弱性を持つ複数のセキュリティモデルを作ることで、安全性を侵害することにもなる。
- **エージェントにユーザのように振舞わせるのではなく、権限を委譲するモデルに置き換えるべきである。**現在、エージェントはしばしばユーザと区別がつかないように行動し、説明責任のギャップやセキュリティリスクを生み出している。真の権限委譲は、エージェントがその代理元であるユーザを識別可能なまま、委譲された範囲を証明する、明示的な On behalf ofフロー（代理認証フロー）を必要とする。
- **人に出来るガバナンスと同意の量には限度がある。**エージェントの急増に伴い、ユーザは膨大な量の認可リクエストに直面することになり、確かめないで承認してしまうというセキュリティリスクが生じる。柔軟なエージェントの事前承認と範囲設定は、最小特権とは相反する。
- **再帰的委任はリスクを生む。**エージェントがサブエージェントを生成したり、他のエージェントにタスクを伝達したりすると、明確な減衰メカニズムがないまま、複雑な認可チェーンが形成される。
- **人間のチームに代わって行動し報告するエージェントにはサポートが不足している。**OAuthとOpenID Connectは個々のユーザの認可のために設計されたが、エージェントはグループ内の共有コードベースやチャットチャンネルで採用することができる。このようなマルチユーザ環境では、ユーザごとにさまざまな権限レベルが存在するかもしれないが、単一の設定の中では、すべてのユーザが共通の目的を共有している。共有エージェントをサポートする一般的なプロトコルは存在しない。
- **自律性を信頼するにはまだ、検証を自動化する仕組みが欠けている。**ヒューマンインザループ(人間がAIの意思決定プロセスに積極的に関与する）モデルを超えて拡張するには、エージェントの行動が継続的に運用上の目標と制約に合致していることを確認するための、新しいプログラムに従った手法が必要である。
- **ブラウザとコンピュータを使用するエージェントは、現在の認可パラダイムを破壊する。**視覚インターフェースを直接（またはMCPを経由してブラウザ・オーケストレータに）制御するエージェントは、従来のAPIベースの認可制御をすべて回避する。オープンなウェブをロックダウンから守るには、ウェブボットやウェブエージェントの強固な認証が必要になる。
- **エージェントの多面的な行動がアイデンティティを複雑にしている。**技術の進歩により、エージェントは独自に行動できるようになり、エージェントは独自の認証情報、権限、監査証跡を持つ必要がある。さらに、エージェントの性質をハイブリッドにすることで、独立した実行とユーザの代行を交互に行うことができる。

目次

Agentic AIのためのアイデンティティ管理:

AIエージェントの世界における認可、認証、セキュリティの新たな挑戦

Tobin South　Subramanya Nagabhushanaradhya　Ayesha Dissanayaka
Sarah Cecchetti　George Fletcher　Victor Lu　Aldo Pietropaolo
Dean H. Saxe　Jeff Lombardo　Abhishek Shivalingaiah　Stan Bounev
Alex Keisner　Andor Kesselman　Zack Proser　Ginny Fahs
Andrew Bunyea　Ben Moskowitz　Atul Tulshibagwale
Dazza Greenwood　Jiaxin Pei　Alex Pentland

ご協力いただいた方々

本書は2025年4月からTobin SouthがOpenID FoundationのためにArtificial Intelligence Identity Management Community Group、Stanford大学の Loyal Agents Initiative、および多くの独立した査読者や貢献者と協力して作成した。WorkOSには、著者の時間及びここに記載のテストに実質的に貢献し、完成を支援してくれたことに感謝したい。

多くの共同研究者、共著者、アドバイザー、査読者が本研究に貢献してくれたが、特に以下の人たちに特別に感謝する：Subramanya Nagabhushanaradhya, George Fletcher, Sarah Cecchetti, Ayesha Dissanayaka, Aaron Parecki, Pamela Dingle, Tom Jones, Aldo Pietropaolo, Lukasz Jaromin, Tal Skverer, Wils Dawson, Stan Bounev, Kunal Sinha, Bhavna Bhatnagar, Nick Steele, Abhishek Maligehalli Shivalingaiah, Victor Lu, Chris Keogh, Govindaraj Palanisamy (Govi), Pavindu Lakshan, Michael Hadley, Zack Proser, Cameron Matheson, Dan Dorman, Zack Alex Pentland, Jiaxin Pei, Dazza Greenwood, Ginny Fahs, Andrew Bunyea, Ben Moskowitz, Atul Tulshibagwale, Jeff Lombardo, Elizabeth Garber, and Gail Hodges.（敬称略）

1 AIエージェントは、これまでのエージェントと何が異なるのか？

AIエージェント特有のアイデンティティ、認証、認可、監査のニーズを理解するための最初のステップは、それらが何であり、なぜ異なるのかを正確に定義することである。

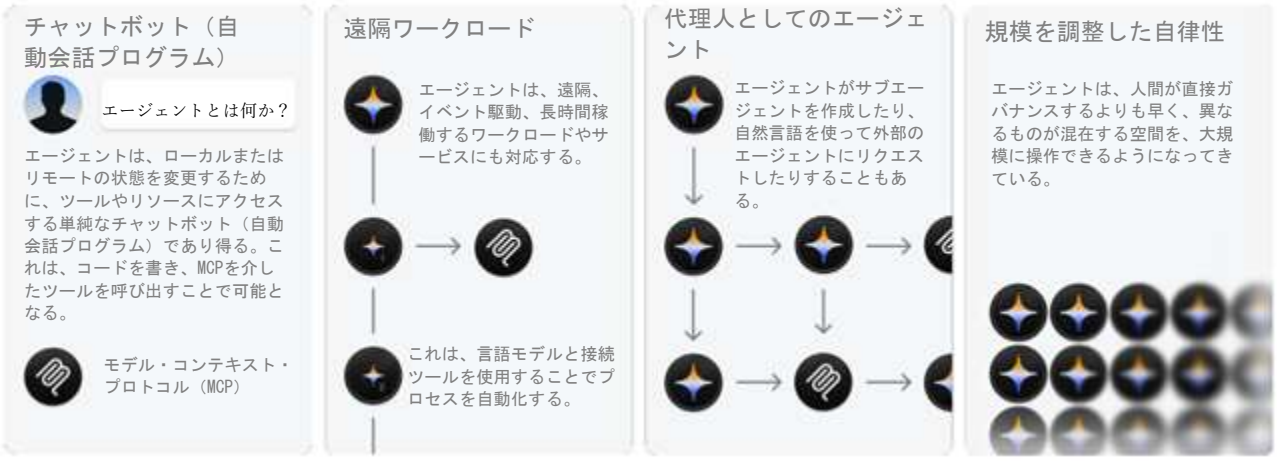

図1：自律性のレベルに応じた、AIエージェントタイプ別のツール（MCP等）使用例

1.1 AIエージェントの定義

AIシステムは世界を席巻しており、言語モデルを使ったチャット・インターフェースから、伝統的な機械学習技術を使った社内ワークフローの自動化まで、さまざまな形で提供されている。本稿でAIエージェントと呼ぶものは、**モデル推論時の「決定」をもとに、特定の目標に向けた「行動」を実行する能力を持つAIベースのシステムを指す。**これは、テキスト出力を生成する単純なチャットボットを超えている。あらかじめ定義されたルールや指示に従う従来のソフトウェアとは異なり、AIエージェントは前後関係から学習し、新しい状況に適応し、自律的に意思決定を行うことができる。従来のソフトウェアでは通常、新しい状況に対応するために手動での更新を必要としているが、AIエージェントは状況と論理的思考を駆使してパフォーマンスを向上させ適応する。これらは、APIだけでなく、モデル・コンテキスト・プロトコル（MCP）のような特殊なエージェント通信プロトコルや、ブラウザインタフェースを介して対話することができる。この動作の中には、標準的な遠隔ワークロードやアプリケーションに類似するものもあるが、その高度な柔軟性、非決定的性、そして文脈依存性により、明確に区別される。

一般的に「エージェント」とは、言語モデルを使用して外部リソースと対話する、識別され認可されたソフトウェアと定義される。本稿を通して、考慮すべきエージェントの例には以下が含まれる：

- MCPを介して特定のAIを中心とするツールを呼び出す、言語モデルベースのチャット型インターフェースのもの
- 言語モデルを使用し遠隔ワークフローを自動化したもの。システムへのインプットがツールやデータベース操作を引き起こすもの（反復的かつ非決定的な動作を持つ、ソフトウェアと言語モデルが一つのワークロードで統合されたもの）
- 半自律的に動作し思考を連鎖するエージェント。一連の段階的なステップを実行し、テキストによる「思考」プロセスと外部ツールやデータベースへの問い合わせを交互に繰り返すもの

多くのAIシステムが存在する中、本稿は主に大規模基盤モデルに基づくエージェントシステム

（例えば、ChatGPT、Claude、Gemini、LLaMa、およびこれらのモデル上に構築され外部ツールへのアクセスを可能にするシステム）に焦点を当てている。これは、この技術における現在の進歩と集中する投資の状況を考慮したものである。エージェントにはより包括的な定義も存在し、例えばウェブ検索に特化したAIシステムやコンピュータと直接対話するエージェントなどがある。これらは議論に値する重要なユースケースだが、明確な焦点を維持するために本稿で扱う範囲外である。

最後の決定的な違いは、（ユーザとコンピュータ間の）対話の形式そのものにある。従来のソフトウェア Webクライアント、デスクトップ、モバイルは、ボタンのクリック、フォームの送信、メニューの選択など、構造化され曖昧さのないユーザ入力で動作する。これらの行為は、明確で監査可能な行為の承認を意味する。対照的に、AIエージェントは、構造化されていないマルチモーダル(複数の情報源やモーダルを統合した）入力を解釈するように設計されている。ユーザは、メッセージのやり取りだけでなく、文書、画像、音声記録、ビデオファイルを通じても指示を与えることができる。例えば、ユーザはスキャンした請求書の画像を音声による命令でアップロードするかもしれないし、複雑な電子メールのスレッドを指示によって転送するかもしれない。これにより、従来のUIが持つような機械可読な同意信号がないデータから、意味を解釈し、権限の範囲を特定し、実行計画を作成するという、大きな負担がエージェントに課される。この指示の時点における曖昧さが、新たな認証・認可モデルが必要となる主な要因となっている。

1.2 エージェントは特定の認証と認可を必要とする

すべてのオンライン作業は認証と認可を必要とする。これは、エージェントベースのワークロードも例外ではない。しかし、エージェントは自律性を強めており、複数の外部ツールと順番にやりとりしながら、多段階のプロセスを実行できるようになっている。これにより、ユーザの同意、特定の規制への影響（ヒューマン・イン・ザ・ループの要件など）、エージェントのガバナンス、大規模で動的な状況におけるアクセス制御の粒度に、新たな懸念が生じる。エージェントは、サービスの観点からすると非常に特殊なワークロードの集合であり、それらをどう扱うべきかは重要な課題である。これについては本稿を通じて概要を示す。

さらに、AIエージェントと特定の外部リソースの数が増え、エージェント操作が複雑化するにつれて、AIエージェントのやりとりを手動またはきめ細かく認可することは不可能となる。スケーラブルかつ堅牢な信頼、ガバナンス、セキュリティのシステムを設計することは、このような大規模システムにおけるディスカバリ、登録、認可を実現する上で極めて重要となる。

エージェントの高い柔軟性と、非決定的で外部リソースに接続されているという性質は、エージェントのアイデンティティ、ガバナンス、認証、認可、最小特権、監査、およびその他の関連する側面に特有の課題を提示する。構造化されていない複雑な入力を解釈し、外部の一連の動的リソースとやり取りする必要性によって引き起こされるこの柔軟性は、エージェントのアイデンティティ、ガバナンス、認証、認可、最小特権、および監査に対する特定の課題を提示する。本ホワイトペーパーでは、このような課題とその解決策に焦点を当てる。

1.3 対象の読者層

このホワイトペーパーでは、AIエージェントのエコシステムを形成する3つの主要グループが直面する、異なるが相互に関連するセキュリティとアイデンティティの課題を取り上げる：

- **AIの実装者および設計者向け**：OAuth2.1のような基本的な標準仕様や、MCPのような新興の

プロトコルを使って、安全で相互運用可能なエージェントシステムを一から構築するための技術的なガイダンスを提供する。

- **企業およびセキュリティ責任者向け**：ガバナンス、リスク、統合の戦略について概要を説明し、エージェントのライフサイクルを管理し、コンプライアンスを確保するために、SSO（シングルサインオン）やSCIMのような馴染みのあるモデルを適用する方法について詳述している。
- **顧客データプラットフォームと製品責任者向け**：ユーザの信頼の基礎に焦点を当て、拡張可能な同意、実質的な権限委譲、エンドユーザにとって透明で安全なエージェントとのやり取りの設計といった中核的な課題に取り組んでいる。

本稿のセクション 2 では、エージェントのアイデンティティおよび認可の現状を要約し、さらに検討するための関連リソースを示す。

近い将来実現するであろう高度な能力を持つ自律システムの発展は、AIエージェントの認可と認証に関する新たな課題と疑問を生み出し続けるだろう。

セクション 3 では、このような高度な AI システムから発生する可能性のあるリスクと課題、リスクを 緩和するためにさまざまな アイデンティティおよび認証アプローチが果たし得る役割、およびこれらのシステムで発生し得る課題について検討する。

2　現在のユースケースに対応可能な解決策

AIエージェントが急速に開発・導入が進む今、その活動を承認し、説明責任を果たすこと、誰の代理として行動するのか、どのように認証、ガバナンス、監査、管理されるのか、どのような範囲と権限を付与されるのか、などを慎重に検討する必要がある。一般に、このセクションは、現在のAIエージェント実装に広く受け入れられている既存の仕様に基づいている。

2.1 エージェントとそのリソース

本質的に、外部サービス、データソース、またはツールとやり取りするAIエージェントは、クライアントアプリケーションとして動作する。エージェントが、複雑なワークフローを調整・連携する洗練された言語モデルであれ、より単純な自動化プロセスであれ、自身の運用境界を越えてリソースにアクセスしたり操作したりするリクエストは、従来のクライアントソフトウェアによるものと類似している。従って、あらゆるクライアントアプリケーションにおける認証と認可の基本原則は、AIエージェントにとっても同様に重要である。OAuth、OpenID Connect、その他の認証技術をウェブアプリケーションやモバイルアプリケーションに実装するために何年もかけて開発された豊富な知識と確立されたベストプラクティスは、エージェントベースのシステムを保護するための強固な出発点となる。

- AIエージェントは、多様なリソースへのアクセスを求める場合がある。これらには、API経由の構造化データ（顧客関係管理、在庫システム、財務データなど）、知識データベースや保存文書からの非構造化情報、計算サービス、あるいは他のAIモデルも含まれる。エージェントのフレームワークは極めて多様であり、エージェントがリソースとやり取りするメカニズムも異質である。エージェントは時にクライアントとしても、又サーバとしても機能するため、このような議論の根拠とするのは難しいかもしれない。一般的な目的においては、エージェントとは、リモートサーバに対してリクエストを行うクライアント側のワークロード（ユーザとのやり取りが同期的な場合も非同期的な場合もある）と見なすことができる。これらのサーバは、エージェント（および／またはエージェントが代理として操作している元のユーザ）を確実に識別し、実行が許容される範囲を規定しなければならない。

2.2 エージェント・プロトコル

特に大規模言語モデル（LLM）のコンテキストでは、エージェントは特定の「ツール」や「プラグイン」を利用する傾向が強まっている。これらのツールは、常にではないが、多くの場合、既存のREST APIのラッパーであり、その機能と必要な入力情報が明記されており、電子メールの送信、データベースへの問い合わせ、リアルタイムの情報の取得などの特定のアクションを実行するために、AIモデルが自動的に検出・実行できる。

AIエージェントの高性能化と自律性向上は、エージェントが遠隔サービスや他のエージェントとどのようにやり取りするかを標準化するためのプロトコル開発に拍車をかけている。多くのプロトコルが存在するが、Model Context Protocol（MCP） の採用が先行しているほか、Agent-to Agent Protocol（A2A） などのプロトコルも広く商業的に投資されている。要するに、MCPは、モデルベースのインターフェースと、外部ツールやデータリソースの多様なエコシステムとの接続を容易にするように設計されている。MCPの進化そのものが、強固な認可の議論の重要性を強調している。MCPの初期設計には認証が含まれていなかったが、その後のコミュニティからのフィードバ

ックと技術的な議論により、認証と認可に関する考慮が積極的に統合されるようになった。

2.3 MCP

MCPはクライアント・サーバ・アーキテクチャを採用し、AIモデルにリソース、プロンプト、ツールを提供する。リソースとは、モデルにコンテキストを与えるファイルやAPIレスポンスのような、アプリケーションが制御する読み取り専用のデータソースのことである。対照的に、ツールは、外部APIを呼び出したり、計算を実行したりといったアクションをAIに実行させる、モデル制御された機能である。AIアプリケーション（クライアント）はMCPサーバに接続し、これらのコンポーネントにアクセスする。この他にも、エージェントのユーザに入力を要求する機能（elicitation）や、利用者とのコミュニケーションのためのダイナミックUIなどが検討されている。通信には、Streamable HTTPやstdioのようなトランスポートを使用し、サーバがクライアントに更新をプッシュできるようにすることで、非同期操作をサポートする。

現在の進展状況、コミュニティの関与度、そして開発における参考例の多さを踏まえ、本節の詳細な検討の多くはMCPを中心に据える。ただし、ここで論じる原則や課題は、AIエージェントのプロトコル全般や、外部リソースとの安全な統合に幅広く適用できる。

2.4 認証

強固な認証は認可の重要な前提条件であり、AIエージェントによるリソースへの安全なアクセスの基盤を形成する。どのようなサービスでも、中心的な問題は認可である:このエージェントは、この時点で、このエンティティに代わってこのアクションを実行することが許可されていることを確かめる過程で、認証はその質問に答えるために必要な検証可能な「誰であるか」を示す。エージェントがユーザの代理として行動する場合、2つの異なる認証の課題に対処しなければならない：

1. **クライアント認証**：エージェント・ソフトウェア自身は、信頼できるクライアントとして認証されなければならない。これで、このエージェントが、自身が主張する正規のアプリケーションだと裏付けられ、多くの場合、ワークロード識別子によって立証される。
2. **ユーザ認証と委任**：まず人間のユーザを認証し、エージェントへ委譲したい権限の範囲を把握する必要がある。

進化の途にあるMCPは、これを正しく理解することの重要性を強調している。コミュニティはOAuth 2.1を標準的なフレームワークとして使用することにし、認可コードのインジェクション攻撃を防ぐためにPKCE（Proof Key for Code Exchange）[17]のような最新のセキュリティの実装を義務付けている。

アーキテクチャのキーとなる推奨事項は、MCPを実装するようなリソースサーバは、独自のロジック(処理内容や手順）を実装するのではなく、認証と認可の決定を専用の認可サーバまたはIdP（アイデンティティプロバイダー）に外部化すべきであるということである。このような重要な機能の分離は、最新のセキュリティアーキテクチャにおける基本的なベストプラクティスであり、 MCP仕様でも推奨されている。これは、企業環境のような単一の信頼関係にあるドメイン内で特に効果的である。一元化されたポリシーとアイデンティティ管理を可能にし、IT管理者が既存の企業内でのログインをシングルサインオン（SSO）経由で使用して、エージェントの設定と権限を管理できるようにする。この方法をとると、エージェントが使用する可能性のある個々のツールの許可を承認する代わりに、一元化した同意管理によりエンドユーザにも恩恵をもたらし、

アクセスはより広範なポリシーによって管理することができ、ユーザが使いにくさと同意疲れを感じるリスクを軽減する。

MCPのまだ完全に標準化されていない重要な課題は、MCPサーバがエージェントに代わりアクセスする外部サービスやAPIのプラットフォームから認証を受ける方法である。エージェントがMCPサーバを認証した後、そのサーバは最終的なツール（SalesforceやGitHubなど）に対して認証されたAPIリクエストを行う必要がある。現状では、カスタム実装に頼ることが多い。

最後に、安全なエージェント・セッションを維持するには、特に非同期操作の場合、認証ライフサイクル全体に注意を払う必要がある。長期的に実行されるタスクは、最初のアクセストークンの有効期限が切れる可能性があり、最小特権の原則を損なわず安全にトークンを更新するための戦略が必要となる。これには、トークンの適切な期限切れまたは無効化を実装すること、すべての認証イベントの監査ログを維持すること（アイデンティティを特定の認可にバインドすることで可能性がある）、危殆化したときや不要になったときにエージェントの認証情報を速やかに無効化できるようにすることなどが含まれる。

2.5 動的クライアント登録

MCPプロトコルは拡張性を確保するため、動的クライアント登録（Dynamic Client Registration）を採用している。これにより、あらゆるクライアントのサーバ登録が可能になり、認証に必要なクレデンシャルを取得できる。このモデルは、「多対多」のエコシステムにおいて使いにくさを感じないオンボーディングを提供する一方で、多数の匿名クライアントを生み出すという重大なセキュリティ上の欠陥をもたらす。認証されていないパブリック登録エンドポイントがあれば、クライアントは実際の開発者、組織、あるいは説明責任を負う当事者とリンクすることなく作成することができる。この結果、紙による証跡が完全に欠落し、エンドポイントを悪用（例：大量登録によるサービス拒否攻撃）する行為への入口を作り、堅牢なクライアントの識別と認証が不可能になる。どのような企業や高度なセキュリティの状況においても、これは高いリスクである。

クライアントをより堅牢なアイデンティティに結びつけるために、さまざまな方法が提案されてきた。例えば、Client ID Metadata[14]は、MCPのためのURLベースのOAuth Client ID Metadata Documentをサポートすることを提案しており、クライアントにHTTPS URLでメタデータを提供させる。サーバは、事前登録や動的登録なしに、信頼を確立するためにそれを取得し検証する。Client ID Metadataは、他の選択的ワークロード アイデンティティパラダイムと同様に、MCP以外のワークロードにも使用できる。

2.6 認可

認証に成功した後、認可はAIエージェントに許容される範囲で特定のアクションを命じる。今日のMCP仕様で定義されるように、認可はMCPクライアントとMCPサーバの関係を管理するものであることに注意することが重要です。MCPサーバがユーザの代わりに下流のAPIやリソースにアクセスする認可を得るメカニズムは、それぞれ異なる。

クライアントとサーバのやり取りの中で、認可は、役割ベースのアクセス制御（RBAC）、属性ベースのアクセス制御（ABAC）などの標準的なアクセス制御モデル、またはより詳細な方法論を活用する。言語モデルの典型的な非決定論的性質を考えると、最小特権はAIエージェントを展開する際に特に重要である[20]。

2.7 エージェントの非同期認証

多くのAIエージェントのワークフローは非同期的であるため、最初の権限付与ではカバーされなかったアクションについてユーザの承認を得なければならないという基本的な課題が生じる。エージェントが自律的に動作し、ユーザの最初の指示から数時間あるいは数日後にタスクを実行する可能性がある場合、機密性の高い操作にリアルタイムの認可を求めることは現実的でなくなる。Client Initiated Backchannel Authentication (CIBA)[7]は、認可リクエストをユーザの認証応答から切り離すことで、解決方法を提供する。CIBAは、帯域外メカニズムによってクライアントがエンドユーザの認証を開始することを可能にし、エージェントが認可を要求し、ユーザの承認を待っている間、処理を継続することを可能にする。

このアーキテクチャは、以下のような理由からAIエージェントに特に適している：エージェントは、ワークフロー全体を遮断することなく、高リスクの操作の認可を要求できる。ユーザは、認証用デバイスで通知を受け取り、都合の良いときに要求を承認または拒否できる。そして、システムは全ての認可決定の明確な監査証跡を維持出来る。CIBAは、Poll、Ping、Pushの3つの配信モードをサポートしており、それぞれが異なるエージェントのアーキテクチャに最適化されている。Pollモードでは、エージェントは定期的に認可状況をチェックし、一括処理したい状況に最適である。Pingモードでは、認可サーバがエージェントに対して、認証する判断が下され利用可能になったことを通知し、不必要なネットワークトラフィックを削減する。Pushモードでは、完全な認可結果がエージェントに直接配信されるため、可能な限り最速の応答時間を実現できる。ヒューマンインザループ （人間がAIの意思決定プロセスに関与する）を必要とする規制の下で運用されるAIエージェントに対して、CIBAは、ユーザ体験を低下させることなく、人間による有意義な制御を保証する標準化されたメカニズムを提供する。プロトコルのバインディングメッセージ(データの結びつきを伝えるメッセージ）のサポートにより、エージェントは、要求されたアクションに関する豊富な関連データを提供することができ、元のタスク開始から時間が経過していても、ユーザの同意に基づく認可の決定を下すことができる。

これらの帯域外の原則に基づき、モデル・コンテキスト・プロトコル(MCP)は、"URL モード "で拡張できる 、エージェントのユーザに入力を要求する機能(Elicitation).もサポートしている。(SEP-1036).このメカニズムにより、MCPサーバは、MCPクライアント自身を素通ししてはならない機密性の高いやり取りのために、ユーザをブラウザを介して外部の信頼できるURLに誘導することができる。これは、サードパーティのOAuth認可の取得、クレデンシャルの安全な収集、または仲介システムに機密データを公開することなく支払いフローを処理する場合に特に関連する。標準的なウェブ・セキュリティ・パターンと明確な信頼境界線を活用することで、URLモードによる情報引き出しは、明示的なユーザの同意または認証情報を取得するための安全な経路を提供し、その後、MCPサーバが委任されたアクションを完了するために利用できる。これにより、非常に機密性の高い認証やクロスドメインの認証であっても、必要に応じて人間のガバナンスを維持しながら、安全かつ非同期に処理することができる。

2.8 AIエージェントのアイデンティティ

AIエージェントのコンテキストにおけるアイデンティティの概念は多面的であり、単純にユーザになり替わるだけではない、認証、認可、監査可能性において重要な役割を果たす。

MCPを使用する現在の実装では、ユーザが操作するホスト（Claude DesktopやCursorのようなコーディングIDEなど）は、特定のMCPサーバとの通信を管理するためにMCPクライアントに接続する（ホストとMCPクライアントは異なる識別子を持つ可能性があることに注意）。OAuth 2.1とPKCE

の利用は、エージェントからMCPサーバへの認証フローを保護するために使用される。このMCPクライアントには、動的クライアント登録時に生成されたクライアントIDが付属しているが、これは堅牢な独立して動作するワークロードアイデンティティではない。クライアントIDメタデータは、システムユーザであるMCPサーバ（例えば、Claude DesktopがMCPを使用している）に対して、ある程度堅牢な識別子を提供することができるが、ワークロードアイデンティティやAIエージェントの識別子としてはまだ不十分である。

AIエージェント用の堅牢な識別子を作成するために、他のワークロードアイデンティティ標準仕様を利用することもできる。たとえば、Secure Production Identity Framework for Everyone（SPIFFE）とその実行環境である SPIRE は、そのモデルを提供する。SPIFFEは、標準化されたアイデンティティフォーマットと、暗号的に検証可能なアイデンティティドキュメント（SPIFFE Verifiable Identity Document：SVID）を、ワークロードの実行場所に関係なく自動的に発行するメカニズムを提供する。SPIREと統合することで、AIエージェントは、APIキーのような静的な共有秘密に依存することなく、他のサービスとの相互認証に使用できる、短命で自動的に循環されるアイデンティティをプロビジョニングすることができる。

SPIFFE/SPIREモデルは、管理されたインフラ内で検証可能なアイデンティティを確立するために強力である。しかし、ワークロードアイデンティティを証明する方法は基本的に、そのインフラに関する知識と管理に依存している。これは、そのようなインフラレベルの信頼が共有されていない異種のインフラ間の信頼の境界を越えて動作するように設計されているAIエージェントにとって、重要な課題となる。エージェントのアイデンティティは、ホスト環境を可視化できないサードパーティにも持ち運び可能で検証可能でなければならない。

しかし、SPIFFE/SPIREのような堅牢なワークロードアイデンティティ・フレームワークであっても、規模が大きくなると、エージェント型システム特有の要求に完全には対応できない場合がある。従来のワークロードのアイデンティティは、それが何であるかを確認するものであったが、エージェントのコンテキストの中では、その振る舞いも重要である。エージェントのアイデンティティは、リスクベースのアクセス制御を可能にするために、その基礎となるモデル、バージョン、および能力に関するメタデータで強化されなければならない。さらに、エージェントは組織の境界を越えて動作し、ユーザの代理として動作するように設計されているため、そのアイデンティティは移植性が高く、OAuth 2.1の権限移譲モデルと統合できるようにネイティブモードで設計されていなければならない。動的アクターに対して、より洗練され、管理可能で、相互運用可能なアイデンティティを求めるこのニーズは、まさに新興のエージェント固有アイデンティティソリューションが埋めようとしているギャップである。

よりガバナンスしやすく機能豊富なアイデンティティ・モデルに対するニーズの高まりが、商用のアイデンティティ・アクセス管理（IAM）市場に大きな変化をもたらしている。従来のサービスアカウントでは、AI エージェントの動的なライフサイクルと独自のガバナンスニーズに対して不十分であることを認識し、IDベンダーはAIエージェントを第一級（人間と同等）のエンティティとして扱い始めた。顧客登録時に割り当てられるアイデンティティに加え、ベンダーはAI専用のID管理ソリューション（Microsoft Entra Agent ID、Okta AIMなど他多数）を開発・展開している。これらのプラットフォームは、人間のユーザと同様のワークフローを使用して、エージェントのディスカバリ、承認、監査をサポートすることを目的としている。これらのアプローチは、従来のIAMサービスアカウントと類似しているが、エージェントが信頼の境界を越えて自律的にやり取りする必要性を満たす目的のために開発されている。これらのエージェントアイデンティティシステム間の相互運用性はまだ限定的であり、ベンダーは、共通する標準に収束する場合を除いて、独自の取り組み方を開発している。

エージェントの活動の大部分は、人間の代わりに行動することである。典型的なワークフローは、現在のところこれを対象としておらず、これらのOn-Behalf-Of（OBO）パラダイムへの対応については、将来トピックとしてセクション3「委任アイデンティティ」で取り扱う。

2.9 SSOとプロビジョニング

AIエージェントを企業内に展開するには、既存の企業システムと統合する堅牢なID管理インフラが必要である。フェデレーション機能を持つアイデンティティプロバイダを介したシングルサインオン（SSO）により、ユーザは既存の企業インフラのクレデンシャルを使用してエージェントプラットフォームや管理インターフェースにアクセスすることができる。このように確立されたヒトのアイデンティティのパターンを、ヒト以外の、エージェントアイデンティティにも適応を拡大する必要がある。

System for Cross-domain Identity Management (SCIM)[10]プロトコルは、ユーザのライフサイクル管理を自動化するための標準仕様であり、人事システムとアクセス権を同期させ、従業員の入社、役割の移動、または組織からの離脱に応じて、アクセス権を付与、更新、または取り消すためのものである。このライフサイクル管理は、エージェント自身にとっても同様に重要であり、エージェントには、作成、許可、そして最終的な無効化のための正式なプロセスが必要である。これに対処するため、SCIMプロトコルを拡張し、エージェントアイデンティティを正式にサポートするための実験的研究が進行中である。例えば、提案されている「クロスドメイン・アイデンティティ管理システム」：Agentic Identity Schema Draft（エージェントアイデンティティ定義書の草稿）[22]は、新しいエージェントアイデンティティのリソースタイプ(コンピュータを作動させるための資源の種類）を定義する。これによりエージェントは、独自の属性、所有者、およびグループへの所属を持つ、IAMシステム（アイデンティティ・アクセス管理システム）内の第一級（人間と同等）のエンティティとして扱われる。

拡張SCIMスキーマを使うことで、組織はユーザにしているのと同じサービスで、エージェントをプロビジョニングすることができる。これにより、IT管理の一元化が可能になり、エージェントのアクセス許可はその場限りのプロセスで管理されるのではなく、人間の従業員に使用されるのと同じ自動化されたポリシー主導のワークフローによって管理される。AIエージェントが急増し、複数の企業アプリケーションにアクセスする必要があるため、そのライフサイクルと同意フローを一元的に管理することは、セキュリティ、コンプライアンス、業務効率を維持するために不可欠となる。この目的のためにSCIMのような標準化されたプロトコルを活用することで、企業は使い慣れたガバナンス・モデルをこの新しいカテゴリのアイデンティティに適用することができる。

重要なのは、このライフサイクル管理が、堅牢で検証可能なデプロビジョニング[1]プロセスにまで及んでいることだ。エージェントの ″オフボーディング ″は重要なセキュリティ機能であり、エージェントのライフサイクルの信頼できる最終段階である。エージェントが停止されるとき、あるいはより緊急に、そのアイデンティティの侵害が疑われるとき、単に現在の認証情報を取り消すだけでは不十分である。エージェントのエージェントアイデンティティリソースに対する SCIM DELETE操作のような正式なデプロビジョニングシグナルは、アイデンティティ自体が恒久的に削

[1] （訳者注）不要になった、またはアクセス権限がなくなったアイデンティティに対し、登録削除、アクセス権の無効化などを行うプロセス

除されることを保証する。このアクションは、エージェントとそれに関連するすべての権限が不可逆的に削除されることを保証し、それによって潜在的な永続的脅威ベクターとしてそれを無力化するために、すべての統合されたシステム全体に伝播しなければならない。

2.10エージェントのアイデンティティと認可の運用

エージェントのアイデンティティと認可の原則は、確実に実施できる場合にのみ有効である。このための確立されたアーキテクチャパターンは、Policy Enforcement Point（PEP）をPolicy Decision Point（PDP）から分離する認可ロジック(誰がどこまでするかを決定する仕組み）の外部化である（NIST SP 800-162[9]を参照）。PEPは、リクエストの内容を検証するコンポーネント（APIゲートウェイ、サービス・メッシュ・サイドカー、ミドルウェアなど）であり、PDPは、定義されたポリシーに基づいて認可決定（「許可」または「拒否」など）を行う専用サービスである。このように関心を分離することで、アプリケーション開発者はビジネスロジックに集中でき、セキュリティとプラットフォームチームはポリシーを一元管理できる。

このモデルは、自律的な意思決定が爆発的に普及するエージェント・エコシステムにおいて重要である。エージェント自身は、それ自身の操作ロジックのための決定ポイントであるが、やり取りするインフラストラクチャは、そのアクションをガバナンスするための別の認可決定のポイントを必要とする。これらのアーキテクチャパターンは、エージェント特有のセキュリティモデルを実現するための具体的な制御ポイントを提供する。PEP は、委譲されたクレデンシャルを解析し、権限を委譲したユーザとその代理を務めるエージェントとを区別するのに理想的な場所である。

さらに、エージェントを認識していない既存のシステムとの連動では、ゲートウェイのような集中型PEPが差異を吸収する強力な変換層となる。最新の操作性に優れたエージェント・アイデンティティ・トークンを検査し、下流のサービスが期待するシンプルなAPIキーや旧式の認証方式に変換することができ、新興のエージェント・エコシステムと組織の既存のテクノロジー投資との間に重要な橋渡しをする。PEP と PDP 間の通信プロトコルを標準化する取り組みは、OpenID Foundation の AuthZEN（認可サービス）ワーキンググループで進行中であり、まさにこの種の外部化された認可決定のための相互運用可能な API を開発している。

2.11監査可能性のギャップを埋める

これらのアーキテクチャパターンの主な利点は、現在の多くのAIシステムを悩ませている重大な監査可能性（と実際の監査体制）のギャップを埋めることである。今日、ユーザの代理としてエージェントが行ったAPIコールは、ユーザが直接行った動作と区別がつかないようにログに記録されることが多く、説明責任と法的証拠の収集や分析にとってのブラックホールを作り出している。真の権限委譲を実装することで、PEPに提示される認証情報には、本人とエージェントで別々の識別子が含まれる。これにより、PEPは、あるアクションを誰が許可したかだけではなく、どの特定のエージェント実体がそれを実行したかも明確に記録する、充実した監査ログを生成することができる。このような豊富なコンテキスト・データ(適切に処理するためのデータ）を取得することは、デバッグ、コンプライアンス要件への準拠、そして最終的には、すべての行動がその起源まで追跡できる信頼できる自律システムの構築の基礎となる。

2.12ガードレールの設置

AIにおけるガードレール(想定外の事象を防ぐ制御手段）とは、AIシステムが安全かつ倫理的に、意図された境界の範囲内で動作することを保証するために設計されたメカニズム、ポリシー、制約を指す。ガードレールには、機密データへのアクセス制限、利用ポリシーの実施、有害コンテンツの出力監視などの技術的な制御を含めることができる。AIエージェントが責任を持って人間の価値観に沿った行動をとるよう導くことで、意図しない行動を防ぎ、リスクを減らし、信頼を維持する手助けをする。

これらのメカニズムは、従来のIGA（Identity Governance and Administration）に見られた原則の重要な拡張である。成熟したIGAプログラムでは、誰がどのリソースにアクセスできるかを設定するが、AIガードレールは、特にAIモデルとデータを交換する場合に、エージェントがそのアクセスをどのように使用するかに焦点を当てた、より専門的でリアルタイムの制御レイヤーを提供する。例えば、IGAはエージェントに顧客データベースへのアクセス許可を与えるかもしれないが、AIガードレールは、要約のためにLLMに送られる前に、個人を特定できる情報（PII）を自動的に保護するなど、アクションの時点でポリシーを実行する。これは、機密データのモデルへの漏洩や出力の非決定論的性質など、エージェントパラダイム（考え方の枠組み）特有のリスクに対処するものである。AIに特化したガードレールを強固なIGA基盤に統合することで、組織はエージェントの権限だけでなく、その行動も管理する多層的なセキュリティ戦略を整備することができる。

2.13エージェントーエージェント間通信

MCPはリソースにアクセスするための一般的な考え方の枠組みである。場合によっては、遠隔のMCPサーバへのツールコールが、外部のAIエージェントにアクションやレスポンスを要求するために使用される。エージェントが他のエージェントと構造化されたコミュニケーションを行えるように設計された新しいエージェント間プロトコル（A2A）がそのより広範な例である。同様のプロトコルは他にも数多く存在し、いずれもエージェント間のコミュニケーションとタスクの完了を可能にするというビジョンを持っている。A2Aは、認証がどのように実行されるべきかの輪郭を提供するが、上記で強調した多くの疑問は未解決のままである。A2Aが、認可が他のエージェントへのアクセスにとどまらず、下流エージェントのアクションの範囲やリソースの使用に関する制限の設定にまで及ぶ場合、複雑さが増す。これについては、セクション3で詳しく述べる。

2.14当面の解決策のまとめ

エージェントと今日の認証と認可のパターンは、単一のトラストドメインにわたって複数のツールを使用する同期エージェントでは機能するが、非同期またはマルチドメインのコンテキストでは機能しない。

今日のAIエージェント用の認証・認可ソリューションは、基本的なユースケース、つまり統一された信頼関係にあるドメイン内で複数のツールにアクセスする単一のエージェントに対して、効果的で配慮の行き届いたパターンを提供する。企業ユーザがAIアシスタントと対話し、企業のCRM（顧客関係管理）を照会したり、プロジェクト管理ツールを更新したり、社内の知識データベースからデータを取得したりする必要がある場合、PKCEを使用した既存のOAuth 2.1フローとMCPのようなプロトコルを組み合わせることで、堅牢なセキュリティが提供される。このようなシナリオでは、共有IDプロバイダ、一貫した認可ポリシー、および一元化された同意管理によるメリッ

トが得られる。エージェントはワークロードアイデンティティを受け取り、企業IdPを介して認証を行い、IT管理者が管理するスコープ化された権限で各種社内ツールにアクセスする。しかし、このような見事に解決出来るパターンは、ほんの一部にすぎない。例えば、企業エージェントがSalesforceの内部データと外部の市場調査APIの両方にアクセスしたり、エージェントが異なるセキュリティ・ドメインに存在する可能性のある他のエージェントにタスクを委任し始めたりする場合など、エージェントが信頼の境界を越えて動作する必要がでた瞬間、現在のフレームワークには大きなギャップがあることが明らかになっている。

このモデルは部分的に機能不全を起こし始めているのである。原因の一つは、SPIFFE/SPIREのような特定のインフラ制御に依存するアイデンティティメカニズムが、インフラの可視性や制御を共有していない組織間での拡張性を欠く点にある。次のセクションで示すように、再帰的な委任（エージェントがサブエージェントを生成する）、委任の連鎖をまたぐスコープの減衰、説明責任を維持する真のOBOユーザフロー、および異なるエージェントアイデンティティシステム間通信を試みる際の悪夢のような相互運用性の欠如によって、課題は指数関数的に増加する。単一組織の制御下において1つのエージェントを複数ツールに安全に接続することは可能であるが、自律エージェントがオープンウェブ全体でシームレスに動作するという広範なビジョンは、依然として大部分が未解決である。これは、次のセクションで論じる。

図2：トラストドメイン内（例：単一エンタープライズ）および信頼境界を越えたエージェント-ツール間通信（MCP）とエージェント間通信（A2A）のタイプ。外部通信と委任の増加に伴い、複雑性、セキュリティリスク、可観測性、アイデンティティ課題はすべて困難になる。

ベストプラクティス

AIエージェント、MCP、企業内アイデンティティ管理の融合で、いくつかの汎用的なベストプラクティスが生まれている：

- **標準プロトコルの使用**：認証にはOAuth 2.1[8]のような公開されているフレームワークを導入し、ライフサイクル管理にはカスタムの認証メカニズムではなくSCIMを導入する。
- **エージェントのやり取りの認証**：ほとんどのエージェントとリソース間およびエージェント間のやり取りは認証されるべきである。認証を必要としない動作環境もあるが、セキュリティの高い動作環境では決して匿名アクセスを許可すべきではない。
- **最小特権を厳格に適用する**：エージェントは、保護されたリソースへの広範なアクセス権を与えられるべきではなく、そのアクセス権は、情報の出力先のユーザのアクセス権と一致する方法で管理されるべきである。
- **エージェントのライフサイクル管理の自動化**：エージェントのデプロビジョニングを単一ユーザのライフサイクルに密結合するのではなく、SCIMのような標準仕様をプロファイリングすることによって、所有権移転のような複雑な出来事に対処する。
- **ガバナンスのために明確な監査証跡を維持する**：すべての認証の事象、認可決定、およびエージェントのアクションをログに記録する。これにより、将来のコンプライアンス要件を満たし、悪意のあるエージェントの行動を阻止し、不正行為の調査による検出を容易にする。
- **相互運用性を考慮した設計**：IPSIEや将来のMCP仕様のような新興の標準仕様に合わせて進化できる、適応可能なアイデンティティシステムを構築する。

3　自律エージェントのアイデンティティと認可の将来的な問題点

セクション2のソリューションは現在のニーズに対応しているが、AIの発展は、はるかに大規模で、より高度な自律性を持って活動するエージェントに向かっている。この飛躍的な進歩は、アイデンティティとアクセス管理にとって、複雑で、将来顕在化するであろう新たな課題を突きつけている。これらの問題は、単純なクライアント・サーバ認証を超え、何百万という人間以外のアクターが存在する世界におけるアイデンティティ、委任、同意、ガバナンスについて根本的な再考を要求している。

3.1 エージェントアイデンティティのアーキテクチャ思想

最も基本的な課題は、エージェントが誰なのか、何なのかを確立することである。今日、エージェントのアイデンティティは単なるクライアントID（識別子）であることが多く、従来のワークロードでは情報が少なく、差別化もできず、拡張性があり安全なエコシステムには不十分である。エージェントのワークロードが急増するにつれて、それらを堅牢かつ相互運用的に識別することが最も重要になる。識別子を超えて、エージェントはその性質、能力、ディスカバリ、ガバナンスを記述するメタデータや属性を持つことができる。これらの属性は、エージェントに与えられる具体的な権限やアクセス範囲にも影響を与える。

エージェントアイデンティティの標準仕様は、企業と消費者の両方が必要とするアーキテクチャパターンをサポートしている場合にのみ有効である。ベンダーが独自のシステムを開発するなど、すでに断片化リスクは存在している。エージェントが活動するために何十ものアイデンティティを必要とする未来を避けるために、検討に値するいくつかの重要なモデルが存在する：

- **強化されたサービスアカウント**：近い将来最も可能性の高い企業向けの実装は、既知のワークロードアイデンティティ概念の拡張である。エージェントはサービスとして扱われるが、そのアイデンティティトークンはエージェント固有のメタデータ(agent_model、agent_provider、agent_version等)で強化され、SPIFFE/SPIREなどの標準仕様または独自の拡張を介して検証される。
- **委任されたユーザのサブアイデンティティ**：ユーザに代わって直接行動するエージェントの基礎となるこのモデルは、ユーザのセッションに本質的にリンクされ、そこから派生するアイデンティティを作成する。これは "On-Behalf-Of"(OBO)フローの正式な実装であり、エージェントのアイデンティティはユーザの権限とは異なるが不可分である。
- **フェデレートされた信頼性と相互運用性**：エージェントは、中央のIdPなしで多様なドメインにわたって動作するために、相互運用可能なトラスト・ファブリック(信頼を統合する仕組み）を必要とする。このファブリックは、OpenID Federation（HTTPS ベースの識別子を使用）などの確立されたフレームワークや、X.509証明書を活用して構築することができ、異なるOpenIDとOpenID以外のアイデンティティシステム間の検証を可能にする。多様なアプローチが存在し活発な検討が進められている、発展途上の領域である。
- **主権とポータブル・エージェントアイデンティティ**：各エージェント・インスタンス(エージェントの実体）は、DIDs(分散型識別子）や現在標準化されている他のスキームを使用して、説明責任のためにグローバルに一意で、検証可能な識別子を割り当てることができる。自身の暗号鍵を管理することで、エージェントはピアツーピアのやり取りの中で自身のアイデンティティを直接主張することができ、よりオープンで分散化されたエコシステムを可能にする。

OpenID Connect for Agents（OIDC-A）[12]などの提案は、コアアイデンティティクレーム（アイデンティティを証明するための基本的な属性情報）、能力、およびディスカバリメカニズ ムを定義することによってこれを標準化することを目的としている。このような標準仕様は、相互運用性を確保し、強固な監査証跡と安全性に必要なきめ細かな識別を提供するために極めて重要である。より一般的には、エージェント[4]を特定するには、アクションを起こした特定の実体とそのシステムの特性を知る必要がある。

3.2 認可の委任と信頼の連鎖

エージェントがアイデンティティを持ったら、次は行動する権限が必要になる。現在、エージェントは、外部サービスからは不透明な方法（例えば、画面スクレイピングやブラウザの使用）でユーザに**なりすます** ことが多く、重大な説明責任のギャップとセキュリティリスクを生み出している。解決策は、明示的な**権限移譲**モデルに移行することである[20]。

なりすましから代理へ（On-Behalf-Of）

OBO（On-Behalf-Of）パターンは新しい問題ではないが、AIエージェントの普及により、大規模に対処すべき重大な課題となっている。このため、既存の標準仕様を用いて委任を実施する方法について、より規範的ガイダンスが行われるようになった。基本的なパターンは、"OAuth for AI Agents on Behalf of Users"[21]のような提案で検討されているような、正しいOBOフローである。これはなりすましとは決定的に異なる。というのも、アクセストークンには、権限を委譲したユーザ（例：subクレームにおいて）と、行為を許可されたエージェント（例：actまたはazp(クレームにおいて）という、2つの異なるアイデンティティが含まれるからである。これにより、最初の段階から明確で監査可能なリンクが作成される。

再帰的委任とスコープの減衰

エージェント・エコシステムの真の力は、再帰的委任から生まれる。あるエージェントが、より専門化された他のエージェントにサブタスクを委任することで、複雑なタスクを分解し効率的に実行する能力である。これは、洗練されたアプリケーションを構築するための基本的なパターンであり、主要エージェントが目標を達成するためにエージェントのネットワークを連携・管理することを可能にする。しかし、このモジュール性は、マルチホップの委譲の連鎖全体で認可を管理するという、セキュリティ上の大きな課題をもたらす。これは重大な推移的信頼の問題を引き起こす。というのも、連鎖の末端にあるリソースサーバは、リクエストを行う最終的なサブエージェントだけでなく、元のユーザに戻る権限移譲の経路全体を暗号的に検証できなければならないからである。このエンドツーエンドでの検証可能性の実現は、すべてのアクションをその起源にリンクさせる曖昧さのない監査証跡を作成するための基礎となる課題である。この課題は、ある組織のエージェントが別の組織のエージェントに委任する際など、委任チェーンが信頼ドメイン境界を越える場合に深刻化する。これには、異なるセキュリティシステム間で認可コンテキストを保持するための、堅牢なアイデンティティ連携モデルが必要となる。

これを解決するには、**スコープの減衰が**必要である。つまり、権限委譲の連鎖の各段階において、段階的かつ検証可能な形で権限の範囲を絞る能力が必要だということである。メカニズムの選択は、必要とされる信頼モデルと設計思想に依存する。一元管理されたエコシステムやハブ・アンド・スポーク(中核拠点から末端拠点へのネットワーク構造）エコシステムのために、OAuth 2.0 Token Exchange[3]は、エージェントがサブエージェントに代わって認可サーバにダウンスコ

ープ（より詳細な権限に制限）されたトークンを要求する、標準化されたオンラインアプローチを提供する。これはポリシーに基づいた管理を一元化し、リボケーションを単純化するが、遅延時間が発生する。対照的に、より分散化された動的なエージェントネットワークでは、Biscuits[19]やMacaroons[2]のような最新の能力に基づいたトークンフォーマットがオフラインでの減衰を可能にする。これらのトークンは、保有者が元の発行者に連絡することなく、より制限されたバージョンのトークンを作成することを可能にし、認証情報自体に権限と制約を埋め込む。このアプローチは、OCap[2]（Object-Capability）セキュリティモデルに支えられており、トークンの所持自体が権限の証明となる。OCapは強力できめ細かなセキュリティを提供する一方で、既存のWeb標準との統合や、オフライン環境での失効（リボーク）の課題は、なお活発に検討が続いている。

失効の課題

このような構造における重要な、そしてほとんど未解決の問題は、**失効**である。従来のOAuth 2.0では、ベアラートークンを取り消すのは、困難な課題だった。オフラインで減衰したトークンを使用する非集中型システムでは、問題は拡大する。ユーザが主要エージェントのアクセス権を失効した場合、その失効を、さらに委譲されている可能性のあるオフライン・トークンの連鎖の下に伝播させる、確実ですぐに実現できるメカニズムがない。

これを解決するために、いくつかの標準仕様ベースの方法が収束しつつある。OpenID FoundationのShared Signals Frameworkは、セキュリティに関する事象を伝達するためのプロトコルを定義しており、失効をほぼリアルタイムで伝達することを可能にしている。これらのメカニズムが一貫して実装されるようにすることが、Interoperability Profiling for Secure Identity in Enterprise(IPSIE)ワーキンググループで開発されているような企業プロファイルの主な目標である。これには、信頼のあるセッションの終了要件が含まれている。これについての新たなメカニズムは、OpenID Provider Commandsであり、OpenID Provider (OP)がRelying Party (RP)に「Unauthorize（無許可にする）」などの検証が可能なコマンドを直接送信し、特定のユーザカウントのセッションを終了させることを可能にするプロトコルである。このような標準仕様を活用することで、IDプロバイダ側でエージェントのアクセスを取り消したというユーザの決定を、エコシステム全体に確実に伝播することができる。

このように、タイムリーなシステム全体の失効を保証することができないため、事前のリスク軽減が不可欠となる。高速エージェントに不向きな時間ベースの有効期限のみに依存する代わりに、認証情報を実行回数で制約することができる。この方法では、エージェントに厳密に限定された数の操作を許可することで、失効が遅れたとしても、潜在的な影響は予測出来る程度に制限出来る。このテクニックは、このような複雑で自律的なシステムにおいて、最小特権を強制し、信頼を管理するための強力なツールである。

デプロビジョニングとオフボーディング

失効がエージェントのアクティブ・セッションの即時終了を対処するのに対して、デプロビジョニングは、エージェントのアイデンティティと関連する権限を恒久的かつ完全に削除することを意味する。この区別は非常に重要である。危殆化したエージェントのアイデンティティは、単に「失効」しただけで、その元となる登録情報および信頼関係を保持している可能性があり、休眠

---

[2] （訳者注）オブジェクト（機能やリソース）への参照そのものが「能力（capability）」であり、その参照を持っていることが権限の証明となるモデル

状態であるが潜在的として存在し続ける。デプロビジョニングは、危殆化またはエンド・オブ・ライフ・イベント（ライフサイクルの終了）に対する最終的な対応策であるが、その実施方法は、企業向けシステムと消費者向けシステムで大きく異なる。

企業では、エージェントは非人間従業員として扱われ、そのオフボーディングは構造化された検証可能なプロセスでなければならない。セキュリティアラートのような事象によって引き起こされると、エージェントのコアアイデンティティは中央のIdPで（理想的にはSCIM DELETEを介して）停止され、関連するすべての認証情報も無効になる。その後、IdPは、共有シグナルフレームワーク（Shared Signals Framework：SSF）のようなプロトコルを使用して、すべての連携済みドメインにデプロビジョニングシグナルを送信しなければならない。これにより、エージェントの識別子がすべてのアクセス・コントロール・リスト（ACL）の許可対象から削除され、特権の取り残しを防ぎ、エージェントが所有するステートフル(状態を保持する）リソースが安全に転送または廃止される。このプロセスに失敗すると、永続的なバックドア(秘密の接続窓口）が作られ、悪用可能なデータ資産が残され、深刻な監査やコンプライアンスの失敗につながる可能性がある。

消費者向けプラットフォームにとって、デプロビジョニングはユーザの信頼とプライバシーを守るための行為である。このプロセスは、ユーザがカスタムエージェントを削除したり、プラットフォーム権限を失効したりするときなど、ユーザによって直接開始される。

結局のところ、従来のアイデンティティ・ライフサイクル管理と原則を共有しながらも、根本的にAIエージェントのデプロビジョニングは異なり、より重大な課題を持つ。ライフサイクルが遅く、一元的に管理されるヒトのアイデンティティや、予測可能な範囲に限定されがちな従来のワークロードとは異なり、エージェントは、高速で、自律的で、領域横断的な運用のために設計されている。この組み合わせは、情報が漏れた場合、他に類を見ない形で危険をともなう。エージェントは、委任した人間の権限を行使するが、機械のスピードとスケールで動作するため、潜在的な侵害の爆発半径を大幅に増幅する。さらに、再帰的に権限を委譲する能力は、漏洩した単一の危殆化したアイデンティティがサブエージェントのエコシステム全体に連鎖的に障害を引き起こす可能性があることを意味する。その結果、堅牢なデプロビジョニングは単なる運用上のベストプラクティスではなく、安全性と信頼の基礎となる柱となる。すべての信頼の境界を越えて、不正なエージェントの存在を永久に消去する検証可能な高速機能がなければ、安全で統治可能な自律型エコシステムを構築することはできない。

3.3 レジストリと外部ツールへの動的接続

自律型エージェントの驚異的な能力は、ユーザの意図に基づいて新しいツールやサービスを動的に発見し、接続する能力である。このため、サービスディスカバリのための堅牢なインフラが必要となる。その代表例が MCPレジストリで、MCPサーバを公開し、クライアントが発見する方法を標準化するために設計されたオープンなカタログである。このようなレジストリは、発見可能性という重要な問題を解決する一方で、ID およびアクセス管理に、最初の接触における信頼の確立という重大な課題をもたらす。エージェントがレジストリを照会し、タスクを実行するために新しい未知のサーバを特定するとき、リソースサーバはエージェントと既存の信頼関係を持たず、逆にエージェントのユーザはサーバを信頼する根拠を持たない。

このことは、ユーザ体験とセキュリティ、特に(ユーザの操作なしで）裏側で動作する半自律的なエージェントにとって重要な問題を提起する：どのようにして、ユーザは、少し前までその存在を知らなかった可能性のあるサービスとのやり取りに認証をあたえ権限を委譲するのか？この場合、単純な事前の同意を超えて、非同期で帯域外の認可を処理できる構造が必要となる。前述し

たClient Initiated Backchannel Authentication (CIBA)のようなフレームワークは、エージェントがワークフローを一時停止し、接続を確立してデータを共有する前に、信頼できるデバイス上で明確なユーザ承認を安全に要求するメカニズムを提供するために不可欠となる。このダイナミックな信頼の確立は、事前に承認された、サービスの壁に囲まれた閉鎖的な環境内に限定されたものではなく、オープンで相互運用可能なエージェント・エコシステムを構築するための基礎的な要件である。MCPレジストリのこうした課題は、エージェント間のコミュニケーションや、サードパーティに対して自然言語で行われる外部からのリクエストにも同様に対応する。

3.4 拡張性の高いヒューマン・ガバナンスと同意

エージェントが急増するにつれ、その操作量は極めて膨大になり、人間が行うガバナンスに根本的な拡張性に関する課題が生じる。EUのAI法[5]のような規制の枠組みは、リスクの高いAIに対して「効果的なガバナンス」を義務付けているが（第14条）、自律的な動作のすべてに人間の承認を必要とすることは不可能である。一人のユーザが、何十人ものエージェントに毎日何千もの決定をさせる可能性があり、管理しきれないほどの許可プロンプトの殺到につながる。この**同意疲れ**は ユーザ体験を低下させるだけでなく、ユーザが反射的にリクエストを承認し始めるため、逆説的にセキュリティを低下させる。

拡張性のあるガバナンスの設計

これに対処するには、従来の双方向的な同意にとどまらず、ガバナンスのための新しいアーキテクチャパターンへの移行が必要である：

- **エージェント認可のためのPolicy as Code**[3]

  ユーザがすべての動作に対して ″承認 ″をクリックする代わりに、管理者またはユーザは、エージェントの動作範囲・限界（例えば、予算制限、データアクセスの階層、APIコールの速度）を設定する高レベルのポリシーを定義する。そしてIAM（アイデンティティ・アクセス管理）システムは、このポリシーをプログラムに従って実施する。

- **意図に基づく認可**

  ユーザは、自然言語で高レベルの意図を承認する（例えば、「今度の会議のために旅行を予約する」）。システムはこれを、特定の最小特権の許可の集合に変換し、その集合は(表には出ないが）裏で強制的に実行する。

- **リスクに基づいた動的認可**

  Policy Decision Point（PDP）は、エージェントが要求した動作のリスクをリアルタイムで評価することができる。日常的でリスクの低い行為は自動的に許可される。しかし、特異なリクエストは、明示的な、帯域外の人間による承認を要求するために、Client Initiated Backchannel Authentication (CIBA) フローを動的に自動実行する。

自然言語の範囲

ユーザは当然、平易な言葉（「レポートを手伝ってくれ、ただし機密データにはアクセスしないでくれ」）で意思を表す。これは柔軟ではあるが、セキュリティに必要な正確さに欠ける。その

---

[3] （訳者注）ポリシーをコード（機械が読み取り可能な形）で記述し、管理・運用する手法

解決策はハイブリッド・アプローチにある。つまり、ハイレベルな自然言語の指示を正式な機械可読のアクセス・コントロール・ポリシーに翻訳するためにAIを活用することである。ユーザは直感的な指示を承認するが、他方システムは監査可能で決定論的なリソース制約を実施し、たとえエージェントがユーザの意図を誤って解釈した場合でも、必ず動作が制限内に留まるようにする。

リスク軽減としてのガードレール

ガードレールは防御の重要な層として機能し、エージェントの望ましくない行動を防止し、あらかじめ定義された境界の順守を保証する。ガードレールは自律システムに内在するいくつかの重要な問題に直接取り組んでいる：

- **意図しない情報共有の防止** – ガードレールは厳格なアクセス制御を実施し、承認されたデータやリソースのみがAIモデルと交換されるようにすることができる。これにより、機密情報を保護しながら、データ漏洩や情報漏えいを防ぐことができる。
- **機密情報のマスキング** – ガードレールは、個人を特定できる情報や財務情報などの機密データを自動的に検出し、タスク実行のためにAIモデルと処理または共有する前にマスキングすることができる。
- **意図しない動作の制御** – 許容される動作と出力を定義することで、ガードレールは、たとえプログラムにバグがあったり命令を誤って解釈したりしても、エージェントが意図した範囲を超える動作を実行するのを止めることができる。
- **リソース消費の制限** – エージェントは時として、過剰な計算リソースを消費したり、APIコールをしすぎたりすることがある。ガードレールは、システムの過負荷と不必要なコストを防ぐために、レート制限とリソースクォータ制限を設定することができる。
- **コンプライアンスの維持** – 規制の枠組みや社内ポリシーによって、特定の行動や制限が求められることはよくある。ガードレールは、これらのコンプライアンス要件をプログラムに従って強制し、法的リスクや評判リスクを低減することができる。
- **倫理的な整合性の確保** – エージェントが有害な、偏向した、非倫理的なコンテンツを生成したり、差別的な動作に関与したりするのを防ぐために、ガードレールを設計することができる。

3.5 先進的な課題と広範な影響

これらの核心となる問題以外にも、いくつかの先進的な難題が目前に迫っている。

アイデンティティを動作とアウトプットに結びつける

単にエージェントを特定するだけでは不十分で、そのアイデンティティを、そのエージェントが実行する動作や、そのエージェントが生成するコンテンツに、反論の余地なく結びつけることができなければならない。これは、説明責任、否認防止、可監査性を確立するための基礎となる。Coalition for Content Provenance and Authenticity（C2PA）[6]のようなイニシアチブは、デジタル資産の改ざん防止メタデータを提供し、エージェント主導の動作について検証可能な監査証跡を作成するための貴重なノウハウを提供する。

プライバシーvs.説明責任

ユーザの代理として識別可能な形で動作するエージェントは、説明責任とプライバシーの間に高い緊張関係を生み出す。監査に必要なトレーサビリティは、ドメインをまたぐトラッキング(追跡・記録・分析）を可能にし、包括的で潜在的に機微な行動プロファイルを作ることができる。ゼロ知識証明や匿名クレデンシャルなどの暗号技術を活用した**選択的開示**メカニズムが、この問題の解決策となり得る。これらを用いれば、エージェントは「医療データにアクセスする権限がある」などの主張を、実体の身元を明かさずに立証できる。ただし、既存のアイデンティティ標準や規制要件と統合することは依然として大きな課題である。

プレゼンテーション層の問題：ブラウザとコンピュータ使用エージェント

OpenAIのOperatorのようなエージェントは、APIを呼び出すのではなく、ユーザインターフェース（ブラウザ、GUIs）を直接操作することによって動作する。これはセキュリティモデルを逆転させるもので、これらのエージェントはプレゼンテーション層で効果的に人間のユーザになりすまし、従来のAPIベースの認可コントロールをすべて回避する。これらの動作をユーザの動作と差別化することは不可能であり、現行の認証フレームワークに重大なギャップを生んでいるが、Web Bot Authのような取り組みは、このプレゼンテーション層のやり取りに特化したアイデンティティと認証のメカニズムを作成することで対処し始めている。

オープンウェブの保護とエージェントの差別化

最後に、エージェントの急増は、オンライン上でボットを検出するという古くからの問題を困難にする。特に、悪性ボットと、事業価値を生む正当なAIエージェントを確実に識別することが求められている。ウェブサイトがデータスクレイピングや不正利用を防ぐために、より積極的なボットブロッキングを導入するようになると[11]、ユーザの代わりにタスクを実行する正当なエージェントが、締め出されてしまう危険性がある。このため、信頼できるエージェントがウェブを閲覧するための標準化された方法が急務となっている。

IETFの提案である**Web Bot Auth** [13]は、エージェントがHTTPリクエストの中で暗号的に直接アイデンティティを証明できるようにするものであり、期待できそうな方法である。この方法は「エージェントのパスポート」として機能し、HTTPメッセージ署名を使用して、IPアドレスに関係なく、検証可能なアイデンティティをトラフィックに添付する。この取り組みは、 Browserbaseのようなエージェント・インフラストラクチャ・プロバイダーと、CloudflareやVercelのような主要なウェブ・セキュリティ・プラットフォーム企業との協業を通じて、大きな進展を見せている。

ここで重要なのは、ブラウザ中心の認証と、API ベースのエージェント（MCP 経由など）で議論されたワークロードアイデンティティモデルを区別することである。Web Bot Authは、エージェントプラットフォームがオープンウェブ上の責任あるアクターであることを証明するために、ウェブサーバに対して認証を行う。対照的に、API エージェントのためのワークロードアイデンティティ は、多くの場合、ユーザに代わって委任された認可フローの一部として、許可された特定の API エンドポイントに対してエージェントを認証する。前者は、パブリックなウェブアクセスに対する信頼のベースラインを確立するものであり、後者はプライベートなリソースアクセスに対するきめ細かな認可を行うものである。

Web Bot Authのようなプロトコルによる堅牢なエージェント識別は、識別され信頼されたエージェントにはアクセス許可が与えられ、匿名のエージェントにはアクセスが制限されるという、より微妙な二層構造のウェブを可能にするかもしれない。これは、オープンなウェブの将来と、人

間とエージェントのトラフィック（データの流れ）を区別する必要性についての深遠な問題を提起するものであり、技術的な人間性の証明[1]から商業的なアイデンティティ検証サービスにまで及ぶ課題である。

3.6 経済層：アイデンティティ、ペイメント、金融取引

自律型エージェントの有用性は、有料のデータソースへのアクセス、商品の購入、サービスの連携など、経済活動に従事する能力と結びついている部分もある。この能力は、従来の人間との直接のやり取りおよび取引時点での同意の上に成り立つ既存の電子商取引やAPIのセキュリティモデルに対し、根本的な挑戦をもたらす。このシフトは、認可を管理し、ユーザの意図を検証し、エージェント主導の商取引における説明責任を確保するための新しいプロトコルを必要とする。いくつかの新しい標準仕様がこの問題のさまざまな側面に対処しており、それぞれが異なるメカニズムと意図する用途を持っている。

FAPI：重要度の高いAPIの保護

銀行送金や証券取引など、大きな損害やリスクを伴う不可逆的な取引では、基盤となるAPIが最高のセキュリティ標準仕様に準拠しなければならない。OpenID FoundationによるFAPI 1.0および2.0仕様は、この目的のために確立されたセキュリティプロファイルを提供する。FAPIはエージェントに特化したものではなく、リスクの高いリソースサーバを保護するためにOAuth 2.1のフレームワークを強化するものである。（mTLSまたはDPoPによる）Sender-constrainedアクセストークン、より強力なクライアント認証、厳格な同意のログ記録の要件など、その義務化することで、基礎的なセキュリティを提供する。規制された、あるいは高額の金融業務に従事するいかなるエージェントシステムも、取引の完全性と否認防止を保証するために、この FAPI プロファイルによって保護された API とやり取りする必要がある。

Agent Payments Protocol（AP2）：商取引における検証可能な意思

APIセキュリティの上のレイヤーで動作する、グーグルの新しいAgent Payments Protocol（AP2）（エージェントによる決済手順）は、自律的な商取引におけるユーザの意思の把握と検証という特定の問題に対処するために設計されている。A2AやMCPのようなプロトコルの拡張として設計されている。その主なメカニズムは、マンデートという、ユーザの指示を監査可能な証明として機能する暗号署名されたデジタルアーティファクトである。AP2は、否認不可能な監査証跡を作成するための2段階のプロセスを定義している：

- **インテント・マンデート** – 利用者が高レベルの指示を与え、それを署名して記録する。これは、やり取り全体の監査可能な検証対象となる状況を提供する。
- **カート・マンデート**– エージェントが条件を満たすアクションを見つけると、当該購入についての利用者の承認を署名して記録する。事前に許可されたタスクの場合、インテント・マンデートの条件が正確に満たされていれば、エージェントはユーザに代わってこれを生成することができる。

この証拠の連鎖は、多くの場合、マンデートを利用者のアイデンティティに結びつける検証可能なクレデンシャル（VCs）を使用して署名され、認可と真正性という重要な問題に直接答える。

KYAPay：プログラムによって自動化されたオンボーディングのためのID連携トークン

エージェントが、既存のユーザカウントや支払い関係が存在しない新しいサービスと初めてやり取りしなければならない場合、異なる課題が生じる。KYAPayプロトコルは、アイデンティティ とペイメントの認可を単一の持ち運び可能なトークンと密結合することで、この「コールドスタート」（初期データが不足しているために、適切な予測が出来ない）問題に対処している。このプロトコルは「Know Your Agent」（KYA、エージェント確認）プロセスを定義し、従来のKYC（顧客確認）/KYB(事業体確認）をエージェント自体に拡張する。アウトはJSONウェブトークン（JWT）であり、Verified identity クレーム（認証に必要な個人情報）と支払い情報がバンドルされている。これによってエージェントは、新しいサービスとの不可分なやり取りの中で、プログラムに従ってオンボーディングと支払いを行うことができる。

3.7 パート3 結論

最後に、拡張性のあるエージェント・エコシステムを構築するには、運用管理全体に取り組む必要がある。エージェントのアイデンティティは静的なものではなく、安全な作成と登録から、権限の更新、最終的な検証可能なデコミッション（特定の要素を切り離す）処理に至るまで、強固な**ライフサイクル管理**が必要である。この管理されたアイデンティティは発見可能性のよりどころとなり、エージェントは未検証の環境で操作するのではなく、安全で認証された登録を通じて信頼できるサービスを見つけ、やり取りできるようになる。このアイデンティティを認識したインフラ全体が、今度は重要な**Policy Decision Point**として機能する。認可レイヤーは、包括的なガードレールを実装する理想的な場所となり、アクセス制御だけでなく、規制上の制約、安全プロトコル、責任あるAIの原則も実行する。重要なのは、このエコシステムが成功するためには、閉鎖的なプラットフォームになってはならないということである。

4　堅牢なエージェント認可のユースケース

エージェントアイデンティティとアクセス管理の将来の課題を具体化するために、このセクションでは、複雑度順に 6つのシナリオを概説する。それぞれのケースは、AIエージェントの特有の運用特性に直面したときの、従来のアイデンティティ・アクセス管理（IAM）フレームワークの明確な失敗事例を示し、新しいエージェント中心のソリューションの必要性を示している。

4.1 高速エージェントと同意疲れ

目の前に差し迫っている課題は、1つのトラストドメイン内におけるエージェントのアクションが非常に高速であることから生じている。デジタル広告予算の最適化を任された企業AIエージェントを考えてみよう。マーケティング・アナリストからの高度な指令：「クリックスルー率（クリックしてリンク先に移動）を最大化するために予算を再割り当てせよ」は、キャンペーンを一時停止し、入札を調整し、ほんの数秒で資金を移動するための何百もの個別のAPI呼び出しに変換される可能性がある。機密性の高い操作のたびに、ユーザを介在させた同期的な同意を前提とする従来のIAMモデルは成り立たない。ユーザは、管理しきれないほどの認可プロンプトに直面することになり、**同意疲れ**を起こしたり 、十分な注意を払うことなく反射的に要求を承認したりしてしまうことになる。このシナリオでは、高速エージェントの場合、アクションごとの認可を、事前認可の**ポリシーベースの制御**による、より強固なモデルに置き換える必要があることを証明している。エージェントは、リソースサーバによって強制される明確に定義された運用範囲（例えば、予算制限、承認された目標）の中で運用されなければならず、セキュリティの姿勢を対話的な同意からプログラムに従ったガバナンスへとシフトさせる。

4.2 非同期実行と永続的権限委譲

複雑さが次のレベルになると、長時間実行する非同期タスクを実行するエージェントが含まれる。例えば、エンタープライズ・プロセス(企業の業務プロセス）・エージェントは、数日から数週間にわたるワークフローで、新入社員のオンボーディングを担当するかもしれない。エージェントは、ITからハードウェアをプロビジョニングし、HR（人事関連）システムでアイデンティティを作成し、福利厚生プログラムにユーザを登録するために、複数の社内サービスとやりとりしなければならない。短命でユーザセッションにバインドされたアクセストークンに基づくIAMモデルは、このパターンと根本的に相容れない。エージェントは、IAMシステムの第一級主体（人間と同等に扱われる対象）であり、ユーザ主導でつくられたものとは区別され、長期間にわたって独立して認証される、耐久性のある**委任されたアイデンティティを**必要とする。さらに、ワークフローのある段階で例外的な承認が必要な場合（例えば、契約限度額を超える入社一時金）、エージェントは段階的に承認レベルが上げられる標準化されたメカニズムを持たなければならない。**CIBA (Client-Initiated Backchannel Authentication)** フローのようなプロトコルは、エージェントの動作全体を停止させることなく、エージェントが適切な人間の意思決定者に安全な帯域外の認可を要求することを可能にするソリューションを提供する。

しかし、この耐久性のあるアイデンティティは、攻撃価値の高い標的でもある。侵害された場合、エージェントは、長期間にわたって複数の企業システムにわたって悪意を持って行動できる持続的な脅威ベクターとなる。このリスクが上昇するプロファイルは、単純なトークン失効では不十分な対応であることを証明している。この場合、漏洩したアイデンティティを人事システム、IT インフラ、その他すべての統合サービスにわたって恒久的に無効化するために、即時かつ完全な**クロスシステム・デプロビジョニング** 機能が要求される。

4.3 クロスドメイン・フェデレーションと相互運用可能な信頼

エージェントのタスクが組織の境界を越える場合、サイロ化された企業固有のIAMの限界が明らかになる。ユーザに代わって、ユーザの銀行（Trust Domain A）、サードパーティの投資プラットフォーム（Trust Domain B）、信用調査機関（Trust Domain C）からのデータを集約しなければならない金融アドバイザリー・エージェントを考えてみよう。このシナリオでは、単一のアイデンティティプロバイダ（IdP）が、すべてのドメインにわたる ID と認可の両方の真実の情報源として機能することはなく、エージェントは、ドメイン B に対して、それがドメイン A に由来する正当な、ユーザに委譲された権限を持っていることを証明するという重要な課題に直面する。解決には、これらの境界線を超えて信頼を安全に移動できる相互運用可能な標準仕様の上に築かれた連携の構造が必要である。複雑なマルチホップのワークフローの場合、IETFのIdentity and Authorization Chaining Across Domains[18]では、OAuth 2.0 Token Exchange を使用して元のアイデンティティコンテキストを保持するパターンが定義されている。エージェントは、あるドメインから別のドメインにトークンを提示することができ、トークンは、元のユーザとクライアントのアイデンティティを保持する新しいトークンと交換され、完全な監査証跡が保証される。

より中央集権的な企業シナリオでは、Identity Assertion Authorization Grant[15]のドラフトが別のメカニズムを提供し、エージェントが信頼できるエンタープライズ IdPからのアイデンティティアサーションを使用してサードパーティ API のアクセストークンを取得できるようにすることで、アプリケーション間のアクセスを中央集権的に制御できるようにする。あるいは、共通の IdP が存在しない可能性のある、より分散化されたエコシステムでは、エージェントが、委任された権限を暗号的にカプセル化した検証可能な資格情報を提示できる。これらの各アプローチは、IAMを中央集権的な機能から、標準化された相互運用可能なトラスト・ファブリックへと進化させる必要性を示している。このトラスト・ファブリックは、異なるセキュリティ・ドメイン間で安全に受け渡しされ、理解される、検証可能で監査可能な権限委譲の証明を可能にするという基本原則に基づいて構築されている。

4.4 動的エージェントネットワークにおける再帰的委任

将来のアーキテクチャに目を向けると、主要なエージェントは、リアルタイムで発見され接続される、サードパーティの専門エージェント群のネットワークに委任することで、タスクを構成する必要があるかもしれない。これは、エージェントがその権限のサブセットをサブエージェントに渡さなければならない（ロシアのマトリョーシカに似ている）**再帰的委任という**重要な課題を導入している。この現実に対応するIAMモデルは、最小特権の原則（**スコープ減衰として**知られるプロセス）を強制するために、各段階で許可が徐々に絞られるマルチホップ委任連鎖をサポートしなければならない。例えば、広範なデータ分析権限を持つ主要エージェントは、特定のデータ収集タスクをサブエージェントに委任することができるが、サブエージェントには自身のアクセス権限と運用予算のほんの一部しか与えられない。このようなネットワークにおける信頼は分散型であり、各段階で証明されなければならない。そのためには、トークン所有者がオフラインでより制限されたバージョンのトークンを作成できる、**BiscuitsやMacaroonsの**ようなトークンフォーマットが必要になる可能性が高い。権限はクレデンシャル自体に埋め込まれ、検証可能になるため、中央の認可サーバへの絶え間ないコールバックの必要性がなくなり、安全で分散化された共同作業が可能になる。

4.5 サイバーフィジカル(現実世界と仮想空間が融合した）エージェントの安全システムとしてのIAM

IAMの究極の課題は、その行動が世界において直接的かつ不可逆的な結果をもたらす可能性がある自律的なエージェントを統制することである。都市の配水ネットワークを管理するエージェントや、自律型配送ドローン群にとって、認可はもはやデータへのアクセスを制御するためのものではなく、システムのセーフティ・ケースの基本的な要素となる。委任される権限は、安全に動作できる範囲を明確に定めた、機械が読み取れる複雑なポリシーとして表現する必要がある（例えば、「貯水池の水位をXとYの間に維持し、圧力Zを決して超えない」）。エージェントのアイデンティティは、フォレンジック分析を可能にし、否認防止を保証するために、その行動に反論の余地がないよう結び付けられなければならない。この安全とされた範囲を超える重大な結果をもたらす決定については、エージェントは高保証で監査可能なエスカレーションパスを人間のオペレーターに対してトリガーし、現実世界に影響を及ぼす行動の最終的な判断者が人間であることを保証しなければならない。これらのサイバーフィジカルシステムにおいて、IAMは従来の役割を超え、中核的な安全性および政策実施層となる。

4.6 複数のユーザを代行するエージェント

OAuthは、1人の人間に代わり、その人間の権限の一部を使用して、保護されたリソースへアクセスするための仕組みとして設計された。エージェントはチームの一員として使用されることが徐々に多くなり、エージェントのアウトプットは、複数のユーザがアクセスできるコードベースやチャットチャンネルに書き込まれることもある。エージェントが1人のユーザのためだけに行動する場合、他のユーザがアクセスできないMCPやA2Aを介して関連データを収集し、その情報を出力に含めることができる。例えば、CFO（最高財務責任者）がチャットチャンネルでエージェントに質問に答えさせることを想像してみてほしい。CFOは給与データにアクセスできるかもしれないが、チャネル(経路）のすべてのメンバーがアクセスできるわけではない。その結果、エージェントはCFOの許可の下で行動できるため、機密である給与情報を開示することができる。チャネル内のすべてのユーザが共通して持つ公約数のスコープを適切に扱うための標準的な方法はない。属性ベースのアクセス制御（ABAC）ときめ細かな認可はこれに対処するために有効だが、どのような実装にも複雑さと課題が現れる。

5 結論

AIエージェントが単なるツールから自律的なアクターへと急速に進化していることは、デジタルアイデンティティの環境にとって重要な転換点を示している。本稿で概説したように、この業界はゼロからスタートしなければならないわけではない。既存の基礎的なフレームワークは、今日のエージェントを保護するための堅牢で即座に適用可能なソリューションを提供している。責任を分離し、最小特権を適用し、明確な監査証跡を確保するというベスト・プラクティスは、次世代のエージェントシステムを構築するための基盤である。

しかし、真に相互接続された自律的なエージェント・エコシステムの未来は、実装者にこの基盤の先を見据えることを求めている。それは、**なりすましよりも真の委任、同意疲れよりも拡張性のあるガバナンス、独自の環境での孤立よりも相互運用可能な信頼**によって定義される、アイデンティティと権限の先駆的な新時代を招く。再帰的委任、スコープ減衰、検証可能な大企業規模の利用に耐えうるセキュリティプロファイルを解決することが、今回の中心的な任務である。

これは行動への呼びかけである。この未来をうまく切り開くには、業界全体が一致団結し、協力し合うことが必要であり、そのための明確な道筋がある：

- **開発者や設計者にとって** 喫緊の課題は、権限委譲やエージェント・ネイティブ・アイデンティティの新たなモデルを取り入れた柔軟性のあるシステムを設計しながら、既存の標準仕様の安全な基盤の上に構築することである。これは、IPSIEが開発したような企業向けプロファイルに沿うことで、自社のソリューションが安全で相互運用可能であり、企業での採用が可能であることを保証することを意味する。
- **標準化団体にとっての** 課題は、これらの新しい概念を形式化するプロトコルの開発を加速し、 将来のエコシステムが独自の断片的なアイデンティティシステムの継ぎ接ぎではなく、相互運用性のある基盤の上に構築されるようにすることである。
- **企業にとって**不可欠なのは、エージェントをIAMインフラ内で第一級（人間と同等の）アイデンティティとして扱い始め、プロビジョニングから安全で検証可能なデプロビジョニング、ガバナンス・ポリシー、明確な権限系統に至るまで、強固なライフサイクル管理を確立することである。

シンプルなクライアントの認証から、自律的なエージェントのための信頼できるアイデンティティの確立への道のりは、単なる技術的な上位バージョンへの更新ではなく、オンラインでの信頼の管理方法の根本的な進化である。これらの課題をイノベーションの機会として受け入れることで、私たちは一丸となって、AIエージェントの計り知れない可能性が安全かつ責任を持って解き放たれ、すべての人のためになるエコシステムを構築することができる。

主要用語と略語(頭字語)

用語	定義
人工知能（AI）	ユーザの意図や指示に知的に反応できるシステム。通常、言語モデルに支えられている。
AIエージェント	特定の目標を達成するために、モデル推論時の「決定」に基づいて自律的に「行動」できるAIベースのシステム。
認証	アイデンティティを検証するプロセス。エージェントの場合、これにはエージェント・ソフトウェア自体の認証（クライアント認証）と、権限を委譲する人間のユーザの認証の両方が含まれる。
認可	認証されたエンティティが、アクセスまたは使用することが許可される特定の行為およびリソースを決定するプロセス。
アイデンティティとアクセス管理（IAM）	適切な主体（ユーザまたはエージェント）がテクノロジー・リソースに適切にアクセスできるようにするためのポリシーとテクノロジーのフレームワーク。
OAuth 2.1	アプリケーションがユーザアカウントへの限定的なアクセスを取得できるようにする最新の認可フレームワーク。APIへのエージェント・アクセスを保護するための基本的な標準仕様。
Model Context Protocol（MCP）	AIモデルを外部のツール、データソース、リソースに接続し、エージェントがアクションを実行できるようにするための主要プロトコル。
Agent-to-Agent（A2A）プロトコル	AIエージェントがタスクを完了するために他のエージェントと構造化された通信を行うことを可能にするために設計された通信プロトコル。
シングルサインオン（SSO）	複数の独立したソフトウェアシステムに単一系列の認証情報でユーザがログインできるようにする認証スキームで、企業でエージェントプラットフォームへのアクセスを管理するために使用されることが多い。
SCIM（System for Cross-domain Identity Management）	ユーザとエージェントの両方について、異なるシステム間で ID ライフサイクル管理（作成、更新、デプロビジョニングなど）を自動化するための標準プロトコル。

用語	定義
ワークロードアイデンティティ	(AIエージェントのような）ソフトウェア・アプリケーション自体に割り当てられた、一意で検証可能なアイデンティティで、APIキーのような静的な秘密を使用することなく、他のサービスとの認証を可能にする。
SPIFFE / SPIRE	強力で検証可能な、自動的に循環されるワークロードアイデンティティをソフトウェアサービスに提供するためのフレームワーク(SPIFFE)と実行環境(SPIRE)。
委任されたアイデンティティ	エージェントが、指示を開始したユーザとは別に保有する第一級（人間と同等）のアイデンティティ。これにより、エージェントは長期にわたる非同期処理を独立して認証・実行できる。
委任された権限	ユーザが明示的にエージェントに、特定の限定された範囲で自分の代わりに行動する許可を与える認可モデル。
On-Behalf-Of (OBO) Flow	権限委譲を実装するための技術的なパターンで、権限を与えたユーザとアクションを実行するエージェントの両方の識別子を含むアクセストークンが得られる。
なりすまし／なり替わり	エージェントがユーザと見分けがつかない方法で行動し（例えば、ユーザのクレデンシャルを直接使用する）、説明責任と監査証跡にギャップが生じる高リスクの状況。
Client Initiated Backchannel Authentication (CIBA)	エージェントがユーザの認可を非同期かつ帯域外で要求することを可能にするOpenID標準仕様。長時間実行されるタスクや、リスクの高いアクションが明示的かつ非中断的な承認を必要とする場合に最適。
再帰的委任	エージェントがサブタスクとその権限のサブセットを他のエージェントに委譲し、多段階の認可連鎖を構築するプロセス。
スコープ減衰	最小特権の原則を強制するために、再帰的な権限委譲の連鎖の各段階において、徐々に権限を狭めていくプロセス。

用語	定義
失効	現在のセッションを終了するために、エージェントの動作中のクレデンシャルを直ちに無効にすること。
デプロビジョニング	エージェントのアイデンティティおよび関連するすべてのアクセス権をすべてのシステムから正式かつ恒久的に削除することで、ライフサイクルの最終段階、または危殆化への反応。
同意疲れ	高速エージェントからの過剰な認可プロンプトに圧倒されたユーザが、適切なレビューなしに、反射的にリクエストを承認し始めるセキュリティリスク
Policy Enforcement Point (PEP)	エージェントからの受信リクエストを検証し、アクセス権限の判断結果を実際に適用するコンポーネント（例：（APIゲートウェイなど）
Policy Decision Point (PDP)	定義されたポリシーに基づいて認可の決定（許可または拒否など）を行う集中管理するサービスで、その後、PEPによって強制的に施行される
トラストドメイン	単一の機関（ある企業の ID プロバイダなど）がユーザとサービスの認証と認可を担当する、別個のシステムまたは環境エージェントはしばしば、複数の信頼関係にあるドメインにまたがって活動する必要がある。
ガードレール	AIエージェントが安全かつ意図された範囲内で動作するように設計された技術的な制約またはポリシー（機密データのマスキングやリソース消費の制限など）
Web Bot Auth	正規のAIエージェントがHTTPリクエスト内で暗号的にその身元を証明することを可能にする新興のプロトコルで、ウェブサイトが悪意のあるボットと区別するのに役立つ

参考文献

[1] Steven Adler、 Zoë Hitzig、 Shrey Jain、 Catherine Brewer、 Wayne Chang、 Renée DiResta、 Eddy Lazzarin、 Sean McGregor、 Wendy Seltzer、 Divya Siddarth、 Nouran Soliman、 Tobin South、 Connor Spelliscy、 Manu Sporny、 Varya Srivastava、 John Bailey、 Brian Christian、 Andrew Critch、 Ronnie Falcon、 Heather Flanagan、 Kim Hamilton Duffy、 Eric Ho、 Claire R. Leibowicz、 Srikanth Nadhamuni、 Alan Z. Rozenshtein、 David Schnurr、 Evan Shapiro、 Lacey Strahm、 Andrew Trask、 Zoe Weinberg、 Cedric Whitney、 and Tom Zick. “Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online”、 2024年

[2] Arnar Birgisson、 Joe Politz、 Ankur Taly、 Mohsen Vaziri、 and Moses Liskov. Macaroons、 2014年URL https://github.com/rescrv/libmacaroons

[3] Brian Campbell、 John Bradley、 Nat Sakimura、 and Dave Tonge. “OAuth 2.0 OAuth 2.0 token exchange. Technical Report”、 RFC 8693、 RFC Editor、2020年、URL https://www.rfc-editor.org/rfc/rfc8693

[4] Alan Chan、 Noam Kolt、 Peter Wills、 Usman Anwar、 Christian Schroeder de Witt、 Nitarshan Rajkumar、 Lewis Hammond、 David Krueger、 Lennart Heim、 and Markus Anderljung. “Ids for ai systems”、 2024年、URL https://arxiv.org/abs/2406.12137.

[5] European Parliament and Council of the European Union、” Regulation (eu) 2024/1689 of the european parliament and of the council on harmonised rules on artificial intelligence (EU AI 法) ”、2024年、OJ L、 2024/1689、 12.7.2024.

[6] Coalition for Content Provenance and Authenticity、 “C2PA specification、” C2PA Specification v1.3、2024年 URL ： https://c2pa.org/specifications/specifications/1.3/specs/C2PA_Specification.html

[7] OpenID Foundation OpenID Connect Client-Initiated Backchannel Authentication Flow - Core 1.0. OpenID Foundation Specification、2020年、URL https://openid.net/specs/ openid-client-initiated-backchannel-authentication-core-1_0.html

[8] Dick Hardt and Aaron Parecki、 ”OAuth 2.1”、 IETF Internet-Draft draft- ietf-oauth-v2-1-13、 2024年、URL https://datatracker.ietf.org/doc/html/ draft-ietf-oauth-v2-1-13

[9] Vincent C. Hu、 David Ferraiolo、 Rick Kuhn、 Adam Schnitzer、 Kenneth Sandlin、 Robert Miller、 and Karen Scarfone “Guide to attribute based access control (abac) definition and considerations. Technical Report”、 NIST SP 800-162、 National Institute of Standards and Technology(米国国立標準技術研究所) 2014年、URL https://nvlpubs.nist. gov/nistpubs/SpecialPublications/NIST.SP.800-162.pdf

[10] Phil Hunt、 Kelly Grizzle、 Morteza Ansari、 Erik Wahlstroem、 and Thomas Hardjono. “System for cross-domain identity management: Protocol. Technical Report”、 RFC 7644、 RFC Editor、 2015年、URL https://www.rfc-editor.org/rfc/rfc7644

[11] Shayne Longpre、 Robert Mahari、 Anthony Chen、 Naana Obeng-Marnu、 Damien Sileo、 William Brannon、 Niklas Muennighoff、 Nathan Khazam、 Jad Kabbara、 Kar- tik Peris

etla、 Xinyi Wu、 Enrico Shippole、 Kurt Bollacker、 Tongshuang Wu、 Luis Villa、 Sandy Pentland、 and Sara Hooker. “The data provenance initiative: A large scale audit of dataset licensing & attribution in ai”、 2023年、URL https://arxiv.org/abs/2310.16787

[12] Subramanya Nagabhushanaradhya、”OpenID Connect for Agents (oidc-a) 1.0: A standard extension for LLM-based agent identity and authorization”、 2025年、URL https: //arxiv.org/abs/2509.25974

[13] Mark Nottingham. “Web bot auth”、 IETF BOF Request bofreq-nottingham-web-bot-auth、 2024年.URL https://datatracker.ietf.org/doc/ bofreq-nottingham-web-bot-auth/

[14] Aaron Parecki. “OAuth client id metadata document”、 IETF Internet-Draft draft- parecki-oauth-client-id-metadata-document、 2024年.URL https://datatracker. ietf.org/doc/draft-parecki-oauth-client-id-metadata-document/

[15] Aaron Parecki、 Karl McGuinness、 and Brian Campbell. “Identity assertion authorization grant”、 IETF Internet-Draft draft-ietf-oauth-identity- assertion-authz-grant、 2024年、URL https://datatracker.ietf.org/doc/ draft-ietf-oauth-identity-assertion-authz-grant/

[16] Justin Richer、 John Bradley、 Michael Machulak、 and Phil Hunt. “OAuth 2.0 dynamic client registration protocol”、 Technical Report RFC 7591、 RFC Editor、 2015年、URL https://www.rfc-editor.org/rfc/rfc7591

[17] Nat Sakimura、 John Bradley、 and Narendra Agarwal. “Proof key for code exchange by OAuth public clients”、 Technical Report RFC 7636、 RFC Editor、 2015年、URL https: //www.rfc-editor.org/rfc/rfc7636

[18] Arndt Schwenkschuster、 Pieter Kasselman、 Kelley Burgin、 Michael J. Jenkins、 and Brian Campbell. “Identity and authorization chaining across domain”、 IETF Internet Draft draft-ietf-oauth-identity-chaining、 2024年.URL https://datatracker.ietf.org/doc/draft-ietf-oauth-identity-chaining/

[19] Biscuit Security. “Biscuit”、 Biscuit Security Framework、2024年、URL https://www. biscuitsec.org/

[20] Tobin South、 Samuele Marro、 Thomas Hardjono、 Robert Mahari、 Cedric Deslandes Whitney、 Dazza Greenwood、 Alan Chan、 and Alex Pentland. “Authenticated delegation and authorized ai agents”、 2025年、URL https://arxiv.org/abs/2501.09674

[21] Thilina and Ayesha Dissanayaka. “OAuth for ai agents on behalf of users”、 IETF Internet-Draft draft-oauth-ai-agents-on-behalf-of-user、 2024年、URL https:// datatracker.ietf.org/doc/draft-oauth-ai-agents-on-behalf-of-user/

[22] Erik Wahlstroem. “System for cross-domain identity management: Agentic identity schema”、 IETF Internet-Draft draft-wahl-scim-agent-schema、 2024年、URL https://datatracker.ietf.org/doc/draft-wahl-scim-agent-schema/