日本総研
The Japan Research Institute, Limited
先端技術ラボ

AIエージェントへの権限委任
～拡張される認可の仕組みとKYA（Know Your Agent）の必要性～

株式会社日本総合研究所 先端技術ラボ

2026年5月21日

執筆者：渡邊 大喜
＜お問い合わせ＞当社ホームページの問い合わせフォームよりご連絡ください。

はじめに

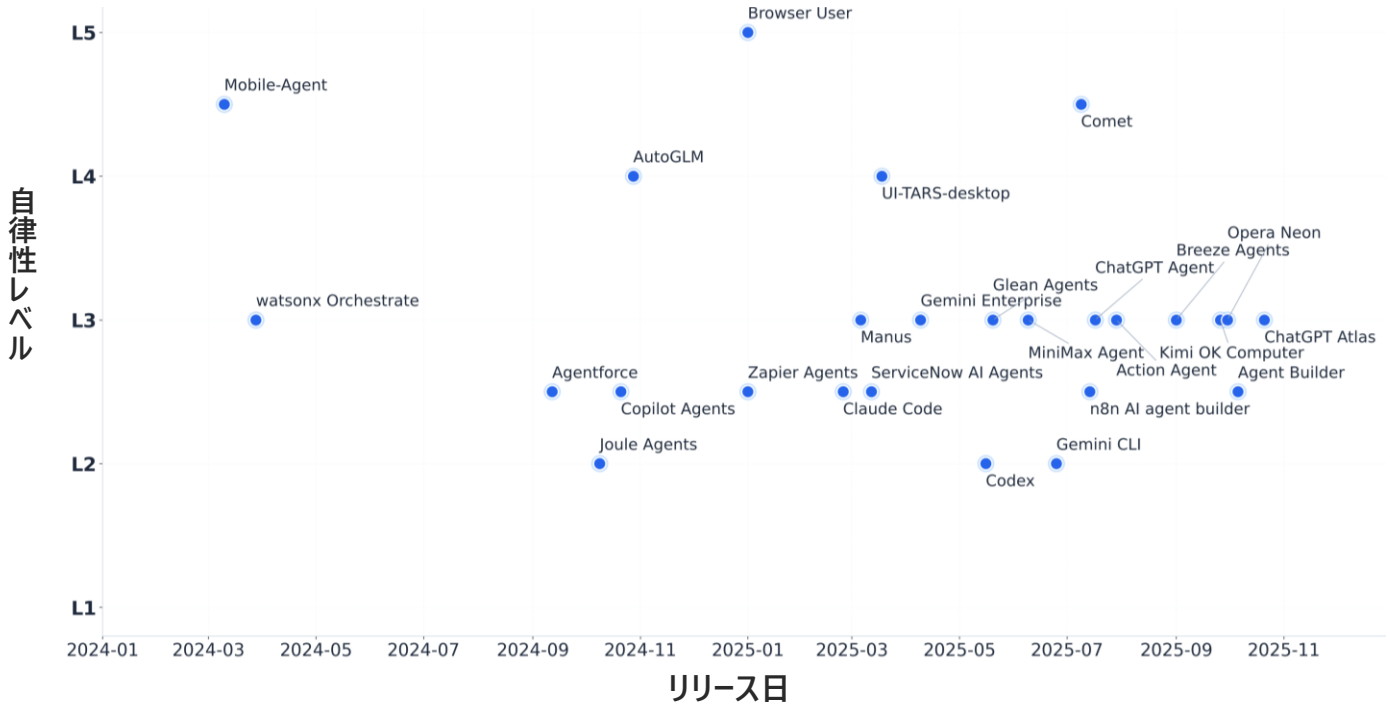
- 生成AIの進化により、AIエージェントが自律的にタスクを実行する時代が到来している。エージェントがAPIを呼び出し、メールを送信し、決済を行う世界では、「**誰が・何を・どの範囲で許可したか**」を技術的に保証する仕組みが不可欠となる。しかし、現行の認証・認可フレームワークは「人間が直接操作する」前提で設計されており、エージェントの自律性が高まるほど、その保証は難しくなる。
- 本レポートでは、このギャップを「**権限委任**」という観点から整理する。エージェントの自律性を中／高の二段階に分け、それぞれの課題を構造化する。中自律性については、OAuth拡張などを中心とする事前委任の技術的アプローチが出揃いつつあり、業界の標準化も並行して進められている。一方、高自律性については、既存の仕組みの延長では構造的な限界が浮上しており、業界横断の標準化はいまだ模索段階にある。
- 最後に、これらの整理を踏まえ、エージェントの身元・能力・権限・責任主体を取引前に検証するプロセス概念「**Know Your Agent（KYA）**」が、エージェントが経済主体として取引や契約に参加する時代に必要になると展望する。

INDEX

中程度の自律性を持ち、長時間の稼働が可能なAIエージェントが市場に広まっている

- 2024-25年にかけて、中程度の自律性（長期的な目標に対してタスク実行を主導できる）を備えた商業ベースのエージェントが登場。
- 米アンソロピックの同社製品Claude Codeに関する利用調査によれば、1ターン当たりのタスク実行時間は増加している。

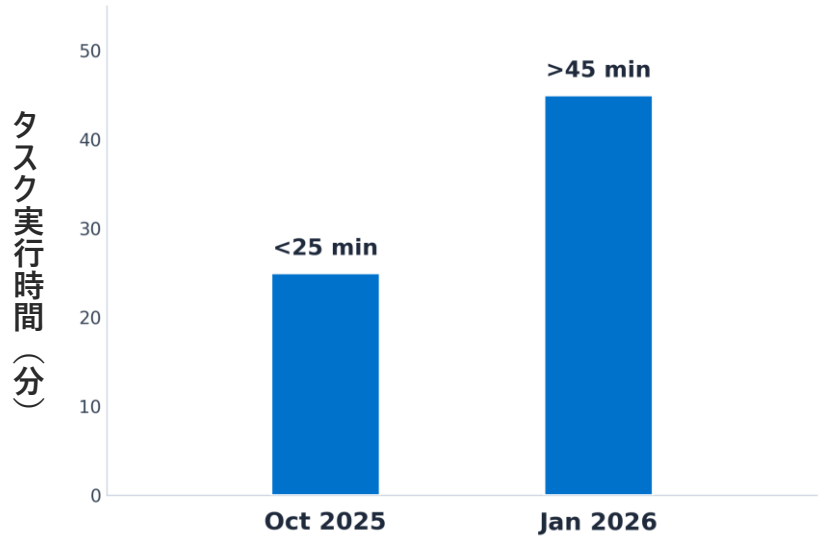
主要なAIエージェントのリリース時期（2024-2025）と自律性レベル*1



※出所を基に日本総研作成。なお自律性レベル*1は各エージェント製品に対して、幅広く対応している場合（例：L1-L3）には、中央値（例：L2）でプロットしている。
（出所）Staufer, Leon, et al. "The 2025 AI Agent Index: Documenting Technical and Safety Features of Deployed Agentic AI Systems." arXiv preprint arXiv:2602.17753 (2026).

*1: 自律性レベルは次の通り L1: ユーザーが指示を出し、ユーザーが意思決定 L2: ユーザーとエージェントが共同で計画、委任、実行を行う L3: エージェントが長期的な目標に対して主導権を発揮 L4: エージェントが障害に遭遇したのみユーザーが関与 L5: エージェントは完全に自律的に動作し、ユーザーが介入する機会が提供されていない

最長クラス*2のタスク実行時間の変化



※出所を基に日本総研作成。
（出所） McCain, Miles, et al. “Measuring AI Agent Autonomy in Practice.” Anthropic, https://www.anthropic.com/news/measuring-agent-autonomy. (参照：2026/4/15)

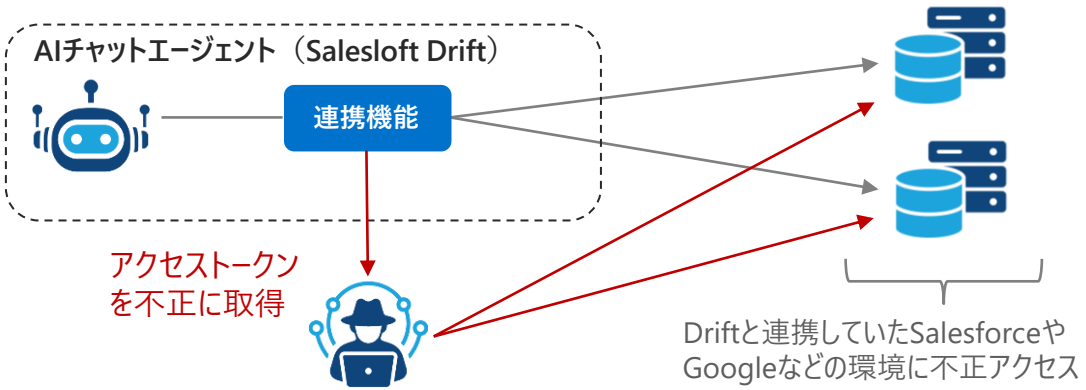
*2: Claude Codeセッションにおける、上位0.1パーセントの最長クラスのタスク実行時間（Claudeが1ターンあたりに作業する時間）

自律性の高いAIエージェントでは、権限管理の不備がインシデントにつながる

- AIエージェントのリスクは出力品質の課題（ハルシネーション等）だけではない。
- 自律性を考慮するあまり、外部連携やツール利用に対して過度な権限付与を行うことが、情報漏洩や不正アクセスの原因に。

委任した権限が悪用される

事例 **AIエージェント連携の権限が悪用され、連携先で不正アクセス**

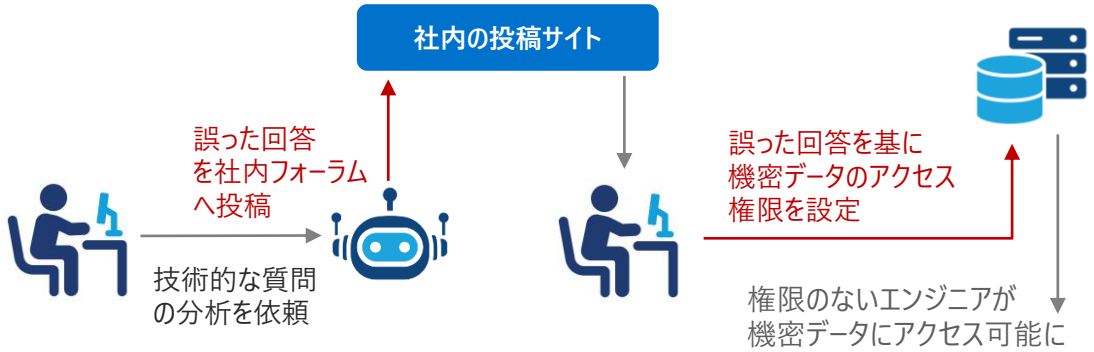


- Webサイト上で顧客対応を行うAIチャット「Salesloft Drift」は、見込み客の判定、営業への引継ぎを担うAIエージェント。
- DriftはWeb上の会話を営業活動や顧客管理に直結させるため、Salseforceなどの外部システムと連携している。
- 2025年8月、外部サービスと連携するためのアクセストークンが攻撃者によって悪用され、複数のSalesforce環境などに不正アクセスが発生*1。

*1: Google Cloud Threat Intelligence. "Widespread Data Theft Targets Salesforce Instances via Salesloft Drift." Google Cloud Blog, Sept. 2025. https://cloud.google.com/blog/topics/threat-intelligence/data-theft-salesforce-instances-via-salesloft-drift (参照：2026/4/30)

AIの助言が権限逸脱に拡大

事例 **AIエージェントによる投稿が、予期せぬ権限逸脱に拡大**



- Meta社では、社内の技術的な質問を分析し、エンジニアに助言するための内部技術アシスタントとしてAIエージェントを活用。
- AIエージェントは自律的に社内の技術フォーラム（投稿サイト）に回答する権限を持っていた。
- 2026年3月の報道によると、AIエージェントが公開で助言を投稿し、その内容を社員が実行した結果、本来アクセス権のない社員が機密データにアクセス可能となるインシデントを起こしていたことが判明*2

*2: TechCrunch. "Meta is having trouble with rogue AI agents." 18 Mar. 2026. https://techcrunch.com/2026/03/18/meta-is-having-trouble-with-rogue-ai-agents/ (参照：2026/5/8)

エージェント・アプリケーションのセキュリティは、「認証・認可」の設計が要となる

- ソフトウェアセキュリティ向上を目的とした米NPO団体OWASPは、AIエージェントに関するセキュリティ課題を公開[*1]（2025年12月）
- 各脅威の緩和策を横断的に分析すると、「エージェント固有のID」「最小権限」「スコープ付きトークン」が繰り返し登場。
  エージェントの身元を確認（認証）し、必要最低限の権限だけ与える（認可）ことが重要視されている。

ASI03：アイデンティティと権限の乱用 (Identity and Privilege Abuse)

エージェントがユーザーの認証情報をそのまま引き継ぎ、本来の想定を超えた操作ができてしまう。エージェントに固有の身分証明がないため、誰の権限で動いているのか追跡が困難になる。

対策例：エージェントごとに個別のIDを発行し、タスクごとに期限付きの限定的な権限を与える。操作のたびに認可を再確認する仕組みを導入する。

ASI02：ツールの誤用と悪用 (Tool Misuse and Exploitation)

エージェントが正規のツール（メール送信、ファイル削除等）を、曖昧な指示や操作により本来意図しない方法で使ってしまう。権限が広すぎると被害が拡大する。

対策例：ツールごとに「何ができるか」を細かく定義し、必要な時だけ権限を付与する。重要な操作には人間の承認を挟む。

ASI07：安全でないエージェント間通信 (Insecure Inter-Agent Communication)

エージェント同士がやり取りする際、エージェント間の認証や身元の確認が不十分だと、なりすましやメッセージの改ざんが起きる。

対策例：エージェント間で相互に身元を暗号学的に認証し、通信内容を暗号化するとともに、メッセージの整合性を確保し、意味的な改ざんを防止する。

ASI10：不正エージェントの侵入 (Rogue Agents)

不正なAIエージェントが侵入し、本来の意図や承認された範囲から逸脱するなどシステム内で有害的な行動を取る。エージェントの行動の整合性やガバナンスが失われる。

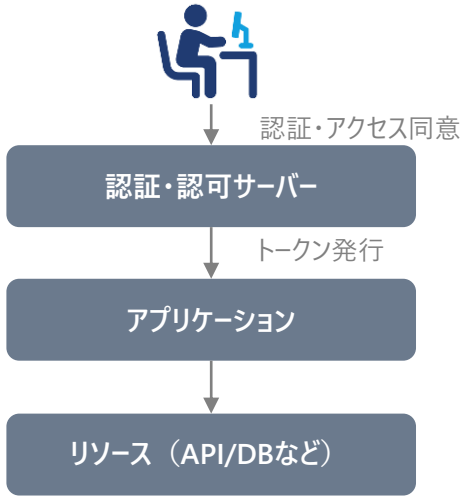
対策例：エージェントごとに暗号学的な身元認証を実装し、エージェントのライフサイクル全体を通じてその健全性（行動の整合性）を確保する。

*1: OWASP GenAI Security Project. “OWASP Top 10 for Agentic Applications for 2026.” 10 Dec. 2025. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/ (参照：2026/4/30) ※本稿では権限管理に直結する4項目（ASI02/03/07/10）のみ取り上げている

従来のアプリケーションでは「人間が操作する」前提で認証・認可が設計されていたが、AIエージェントの普及により、権限委任を前提としたモデルに変化すると考えられる

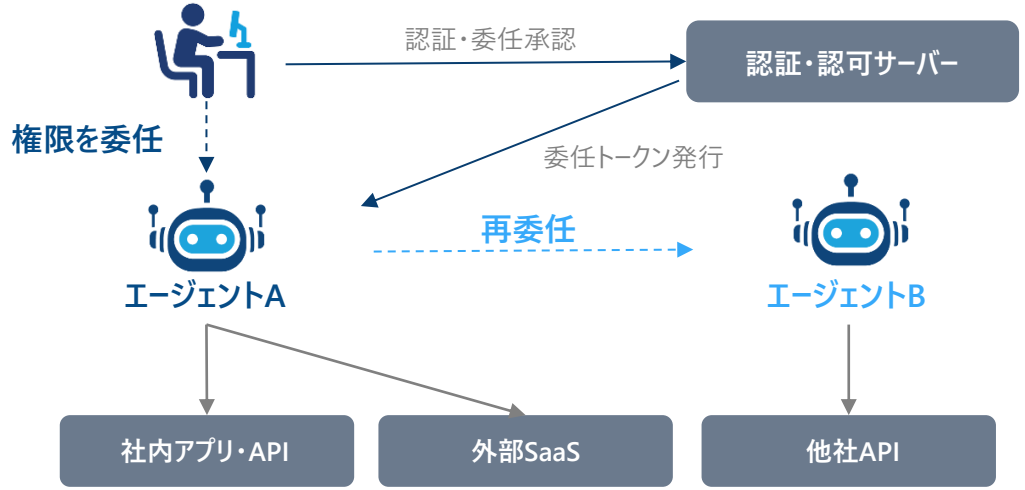
既存の認証・認可モデル

単一ドメイン・同期的・人間が常に操作



エージェント時代に予想されうる権限委任モデル

マルチドメイン・非同期・エージェントが自律判断

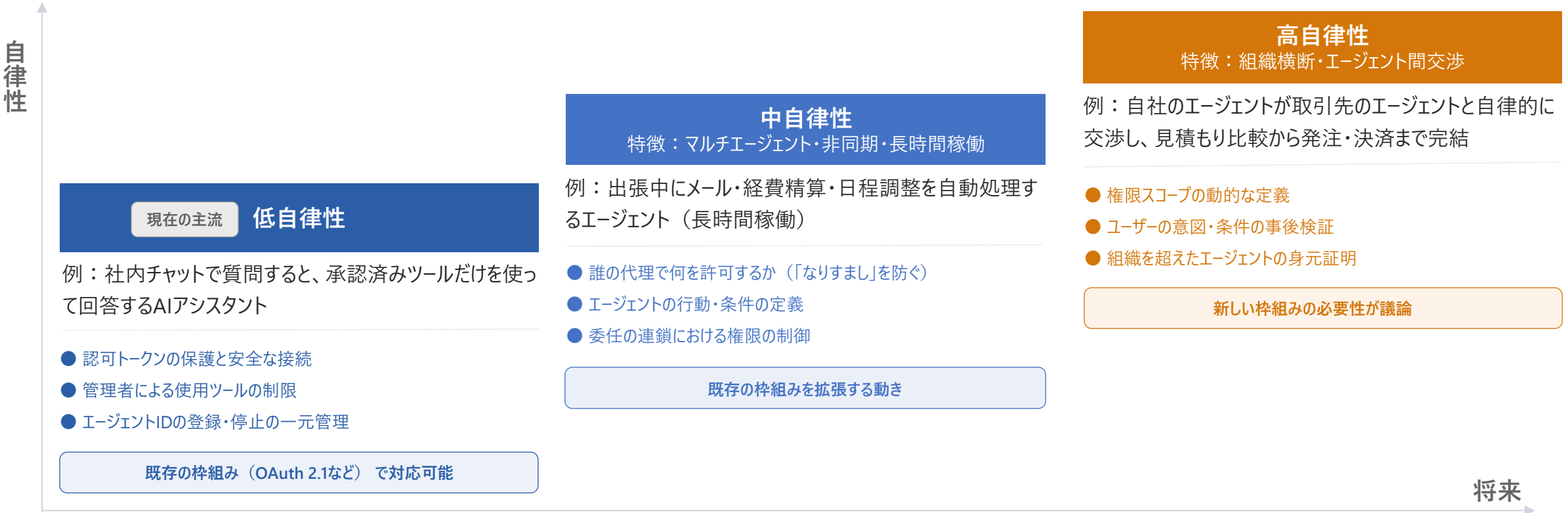


※概念図。具体的なアプローチについてP7以降で解説。

操作主体	ユーザーがアプリを直接操作。 各アクションに明示的な同意を行う。	AIエージェントが状況を判断し自律的に操作を実行。 人間の介在なしに複数ステップを連鎖的に処理。
権限の範囲	静的なスコープ権限（read, write等）で事前に定義。 変更にはユーザーの再同意が必要。	タスクの文脈に応じて動的に変化。 同一エージェントが異なるスコープを要求するケースも発生。
セッション	短時間・ユーザー操作ベース。 アプリケーションを閉じればセッション終了。	長期間・バックグラウンド実行。 ユーザー不在時も継続的に操作を行う。
監査追跡	ログにユーザーIDを記録。 操作主体と責任の帰属が明確。	別のエージェントへの再委任など操作主体と責任の帰属が複雑化。 委任チェーン全体の追跡が必要。

自律性の度合いが高まるほど、権限委任における認可やアイデンティティの課題が深刻化する

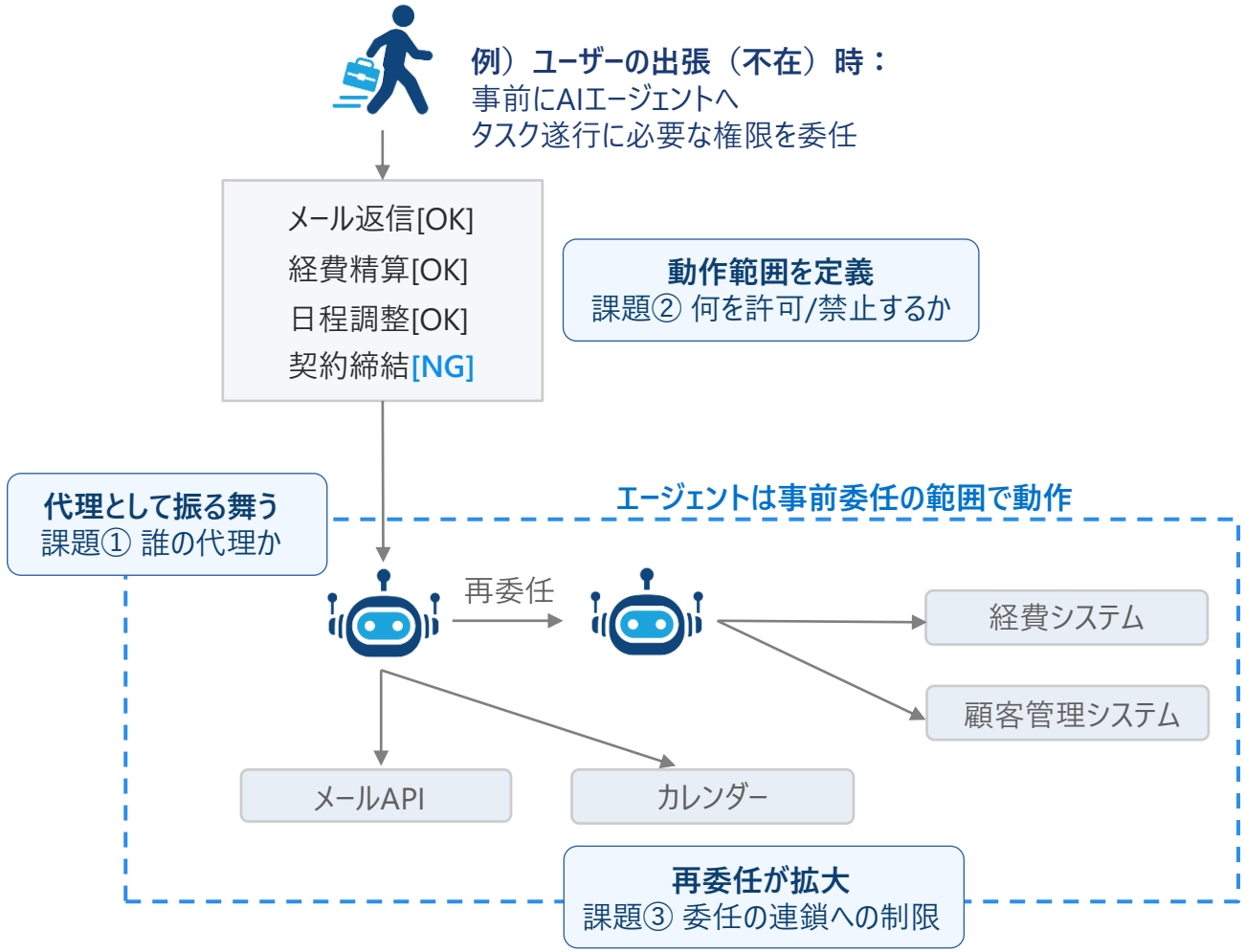
- 既存の認可フレームワークは、シンプルな（低自律性）AIエージェントには対応できる一方、高度な自律性を持つAIエージェントには権限委任のための新しい枠組みが必要との指摘[*1]がある。本レポートでは自律性の度合いに分けて課題を整理する。
- **中自律性については、主に認可面の課題が中心**となり、事前の権限委任を実現する技術的なアプローチが出始めている。**高自律性では、認可面に加えてエージェントのアイデンティティに関する課題が浮上**し、解決策はまだ模索中の段階である。



*1: OpenID Foundation. "Identity Management for Agentic AI: The new frontier of authorization, authentication, and security for an AI agent world." Oct. 2025. https://openid.net/wp-content/uploads/2025/10/Identity-Management-for-Agentic-AI.pdf (参照：2026/4/30)

中自律性エージェントは、「事前の権限委任」をどう実現するかが鍵。既存の仕組みを拡張する動き

- 中自律性エージェントはユーザー不在のまま長時間稼働するため、人間介在を前提とした従来の認可モデルが機能しなくなる。
- なりすましを防ぎ、エージェントの権限の範囲・条件・制約等を定義できる**事前委任の仕組みが求められる。**



課題①　誰の代理か（→P.8）

ツールの利用の際に「なりすまし」ではなく、
「代理」であることを正当に示せる必要がある。

課題②　何を許可/禁止するか（→P.9）

エージェントが行動できる境界を事前に定義し、
決定論的に動作範囲を制御できる必要がある。

課題③　委任の連鎖への制限（→P.10）

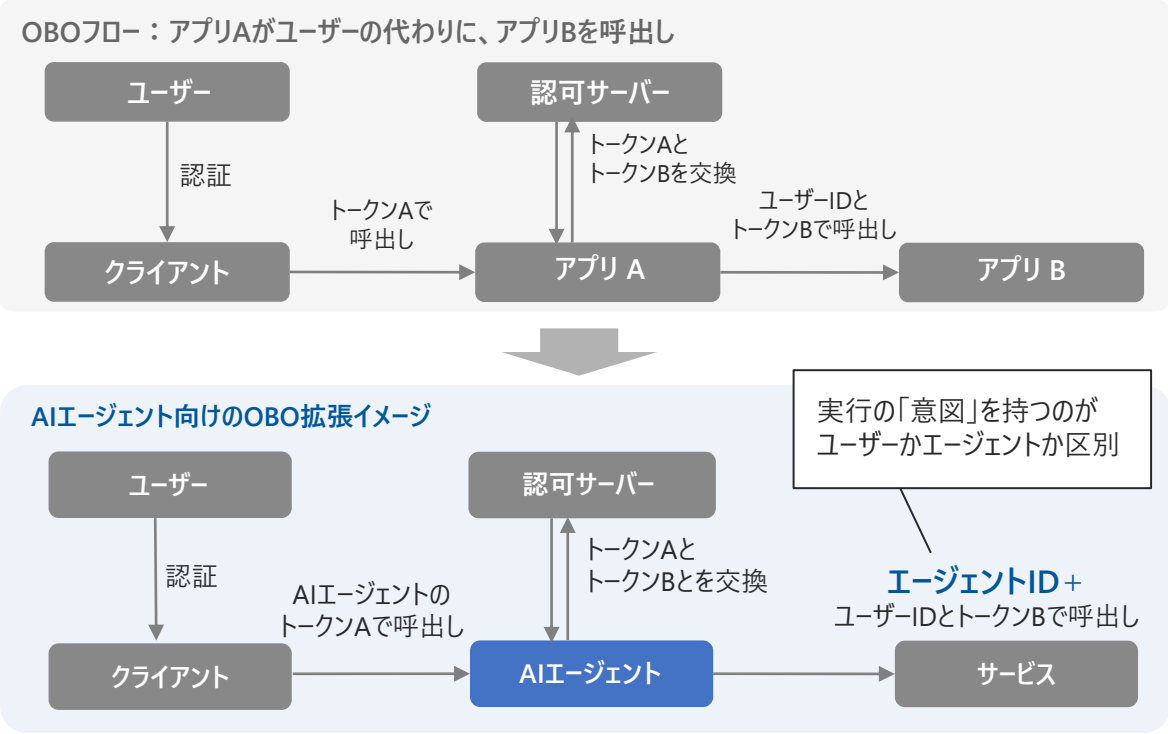
委任が連鎖し、連携が拡大した際に、
権限の範囲を絞ったり、停止の措置ができる必要がある。

誰の代理か：ユーザーの認可トークンを、エージェント用の代理トークンに交換する

- 「誰（ユーザー）の権限を、誰（どのエージェント）に委ねるか」をトークンとして表現する仕組みが必要。
- OAuth 2.0ベースのトークン委任フロー（On-Behalf-Of）を拡張し、エージェント向けの認可フローの実装や標準化が進む。

On-Behalf-Of（OBO）フローの活用

✓ サービス間での代理リクエストに用いられていた認可の仕組みである On-Behalf-Ofフローが、エージェント向けに拡張されつつある*



*: P.9のようなポリシーエンジンを用いた許可の一元管理に加えて、「誰の代理か（On-behalf-of）」を追えるような仕組みの導入が議論されている。

実装と標準化の動向

✓ エージェントIDの管理基盤にOBOフローの採用が始まっている

公表時期	製品	概要
2025.5	Entra Agent ID (Microsoft)	Microsoftが独自拡張したOBOフローを採用。既存のEntra IDでエージェントを管理できるよう対応。
2025.6	Cross App Access (Okta)	大手IDプロバイダーOkta社が開発。IETF標準のToken Exchangeをベースに拡張。ベンダー非依存のIdPで管理。

✓ IETFにおいてもOAuth 2.0を拡張する新たなドラフト提案が出てきている

ドラフト	概要
OBO for AI Agents (draft-oauth-ai-agents-on-behalf-of-user-02)	ユーザーが特定エージェントへの委任に明示同意し、トークンに委任経路を記録する認可フロー。
Identity Chaining / ID-JAG (draft-ietf-oauth-identity-chaining-08 / draft-ietf-oauth-identity-assertion-authz-grant-02)	IETF RFC 8693（Token Exchange）を拡張。企業IdPのIDアサーションから代理用認可グラントを導出し、異なる信頼ドメイン間でトークンを連鎖交換する。（上記のOkta Cross App Accessに採用）

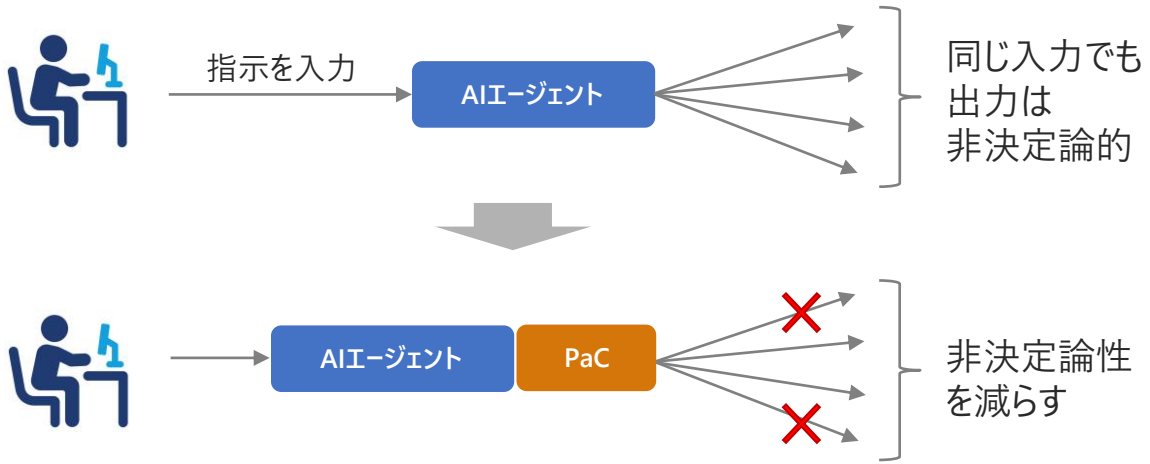
何を許可/禁止するか：エージェントの行動をコードで決定論的に制御する

- トークンの代理委任だけでは「何の操作を許可し、何を禁止するか」までは制御できず、粒度が粗い。
- エージェントの行動境界を宣言的なコードとして事前定義する仕組みとしてPolicy as Code（PaC）の採用が始まっている。

決定論的なガードレールの必要性

✓ ソフトウェアと異なり、AIエージェントは自律的に判断するため、同等の指示を行っても、ツールの利用有無や実行順など出力の振る舞いが異なる（非決定論的）

✓ 事前に、エージェントの行動境界を定義し、エージェントの外側から決定論的に制御するPolicy-as-Code（PaC）の導入が進む



Policy as Code（PaC）による制御

✓ PaCを実現する既存のポリシーエンジンとして、OPA、OpenFGA、Cedarなどがある。AIエージェント向けの製品への適用が進む

Cedar AWS Bedrock AgentCoreで採用[*1]（2026.3 GA）

ポリシーの矛盾・漏れを数学的に検証可能。OAuthスコープでは不可能な粒度。
例）エージェントのツールアクセスをCedarポリシーで制御：
Principal = エージェント（誰が）
Action = 操作（何をするか）
Resource = ツール/API（何に対して）
Context = 条件（いつ・どの状況で）

Cedarポリシー記述例

```
permit(
  principal == Agent::"travel-assistant",
  action in [Action::"read_email", Action::"book_calendar", Action::"submit_expense"],
  resource in Folder::"business-trip-2026"
) when { context.delegation_expires < datetime("2026-04-01T00:00:00Z") };
```

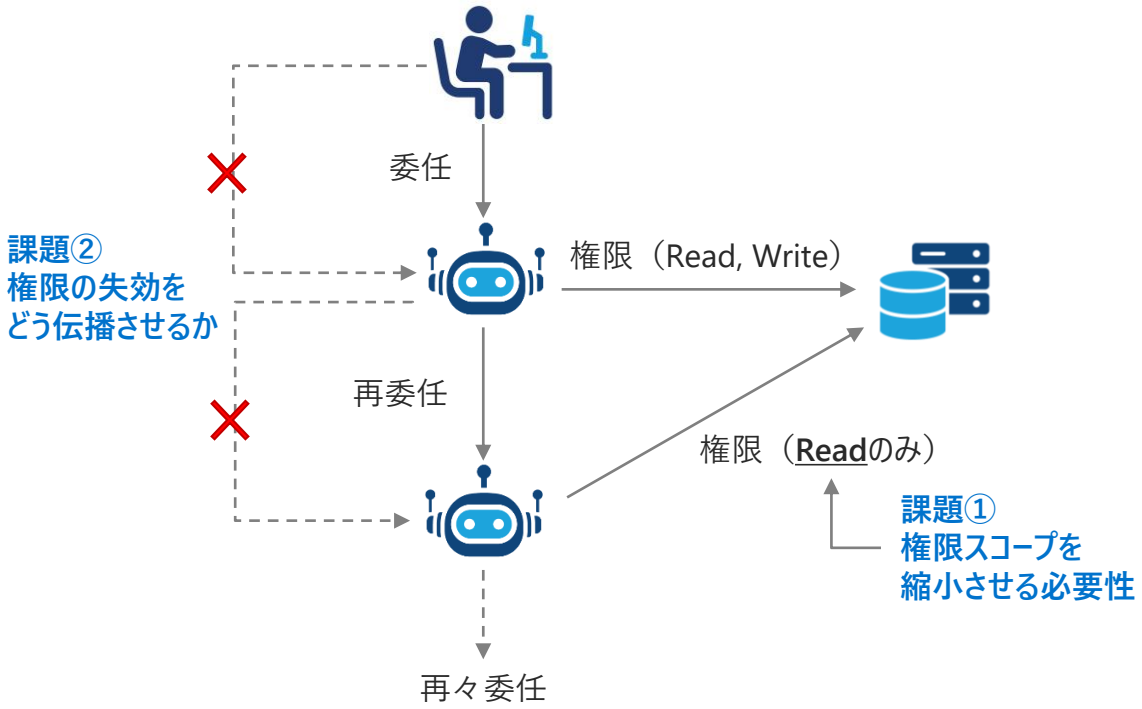
*1: AWS. "Amazon Bedrock AgentCore now includes Policy (preview), Evaluations (preview) and more." AWS What's New, Dec. 2025 (preview) / 3 Mar. 2026 (GA). https://aws.amazon.com/about-aws/whats-new/2025/12/amazon-bedrock-agentcore-policy-evaluations-preview/ (参照：2026/4/30)

委任の連鎖への制限：権限スコープ減衰と失効を実現する

- タスクに応じて自律性の高いエージェントでは、エージェント間でのやりとりの際に再委任が発生する可能性がある。
- 委任の連鎖がどこまで広がるか予測できず、無限に連鎖していく可能性もある。「権限スコープの減衰」や「失効の即時伝播」には既存の仕組みを活用できる。

連鎖する委任がもたらす課題

✓ 自律性が高まり、エージェントが他のエージェント（サブエージェント）にタスクを分解して依頼するケースでは、再委任が発生する
✓ 再委任においては、権限の縮小や失効の仕組みの検討が必要



減衰と失効を実現

課題① 権限スコープの減衰

権限が委任されると権限スコープを狭めるような設計が望ましい。委任の連鎖に応じて、段階的に権限スコープを縮小する「減衰」の仕組みが必要。

アプローチ：
- IETF RFC 8693（Token Exchange[*1]）には減衰の仕様が含まれており、トークン交換時に認可サーバーにダウンスコープされたトークンの発行が可能。
- 一部製品（Okta Cross App Access）で先行実装されている。

*1: あるトークンを別のスコープのトークンに交換するOAuth 2.0拡張

課題② 権限の即時失効

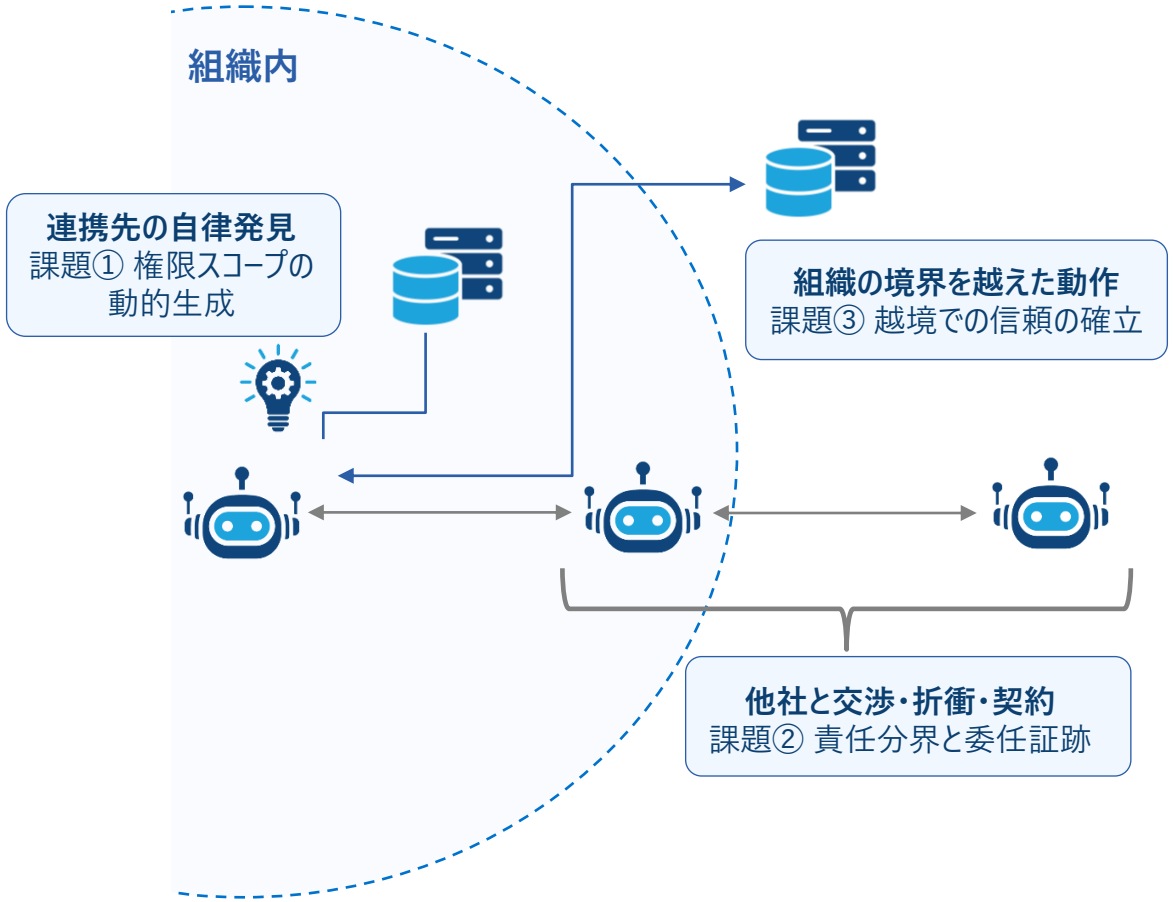
インシデントなどが発生して、上位のエージェントの権限を失効させた場合、再委任の数が増えるほど、即座に失効を伝播できない。

アプローチ：
- サービス間の失効イベントの伝播については、Shared Signals Framework[*2]などの標準化が完了し、実装も進んでいる。
- ただし、エージェント向けの委任チェーンへの適用は未確立である。

*2: 複数サービス間でセキュリティイベントを通知し合うOpenID Foundationの標準

高自律性エージェントでは、既存の仕組みによる権限委任の制御に限界が見えてきている

- 高自律性エージェントは、自ら連携先を発見し、他組織のエージェントと交渉するなど、組織の境界を越えて活動する。
- 権限委任において認可・アイデンティティの両面で構造的な課題が浮上し、**既存の仕組みでは制御しきれない。**



課題①　権限スコープの動的生成（→P.12）

エージェントが自ら連携先を発見するため、権限の範囲を事前に決めきれない。静的スコープに代えて、実行時にスコープを動的に生成する仕組みが必要。

課題②　責任分界と委任証跡（→P.13）

エージェントが経済主体として連携先（他社）と交渉・契約を行う際、意図・契約・実行の責任を後から検証可能にする監査証跡が必要。

課題③　越境での信頼の確立（→P.14）

異なる組織のエージェント同士が連携する際、相手の身元・能力・権限を検証可能にする信頼基盤が必要。

権限スコープの動的生成：静的なポリシーでは対応しきれず、動的なガバナンスへの転換が必要

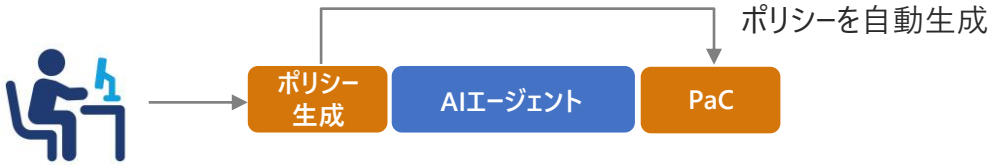
- 高自律性エージェントは自ら連携先を発見して操作するため、人間が事前に作成する静的なポリシーでは対応しきれない。
  静的な権限管理から、動的なポリシーベースのガバナンスへの転換が必要であるとの指摘がある[*1*2]
- 意図や文脈に応じてAIを用いて、システムが自ら認可のルールを生成するアプローチが提案されているものの、発展途上である。

動的なポリシー生成への転換

✓ 低～中自律性では、静的なポリシーを人が作成してきた

- Policy-as-Code（PaC）による静的なポリシー制御
- 管理者や操作者が事前にエージェントの行動範囲をポリシーとして定義

✓ 高自律性では、動的なポリシーをシステムが生成する



意図やリスクに基づくアプローチ

✓ 動的ポリシー生成の手法としてインテント（ユーザーの意図）やリスクを基にしたポリシー生成などが検討されている[*2]が、発展途上

インテントベースの認可

指示に対して、システムが最小特権のポリシーコードに変換。
例）
1．ユーザーが「出張の手配をして」と指示
2．AIや分類器により**最小権限の**ポリシーコードを生成
3．ユーザーの指示にポリシーを適用する

➡ **自然言語の曖昧さを、どのように正確なポリシーに変換するかは未解決**

リスクベースの動的認可

日常的でリスクの低い行為は自動許可し、高リスク操作は人間に承認を求める。
システムがリアルタイムでリスクを評価し、ポリシーを動的に生成して適用する。

➡ **リスクの定義自体が文脈依存であり、汎用的なリスクモデルが存在しない**

*1: Cloud Security Alliance. "Agentic AI Identity and Access Management: A New Approach." AI Safety Initiative, 11 Aug. 2025. https://cloudsecurityalliance.org/artifacts/agentic-ai-identity-and-access-management-a-new-approach (参照：2026/5/6)

*2: OpenID Foundation. "Identity Management for Agentic AI: The new frontier of authorization, authentication, and security for an AI agent world." Oct. 2025. https://openid.net/wp-content/uploads/2025/10/Identity-Management-for-Agentic-AI.pdf (参照：2026/4/30)

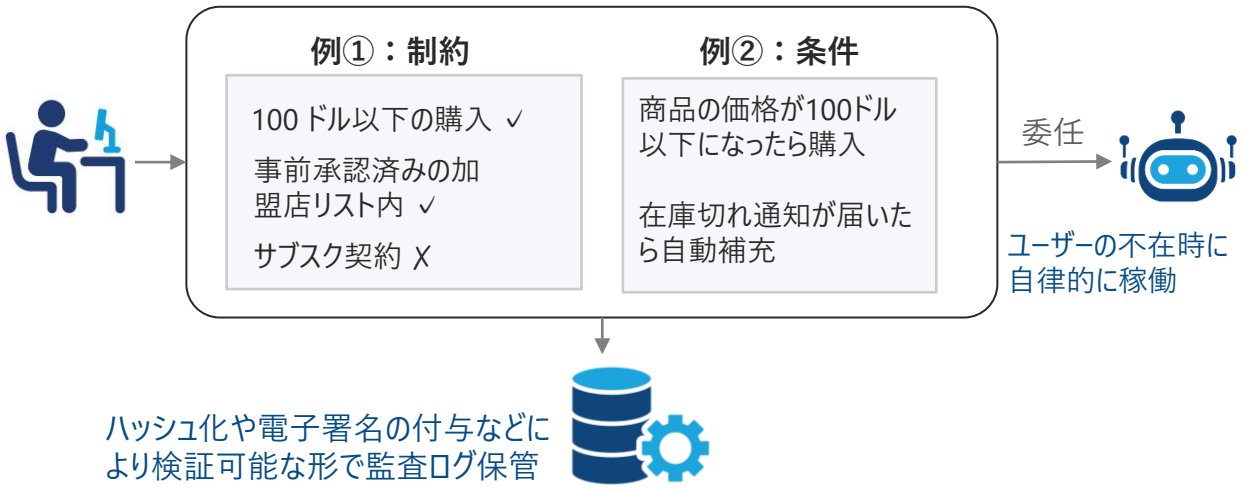
責任分界と権限委任の証跡：権限委任だけでは責任分界できず、暗号学的に検証可能な証跡が必要

- 商取引や支払いなどの責任分界が重要なユースケースでは、事前の権限委任だけでは責任の範囲を明確化しきれず、事後検証が可能となる仕組みが求められている。
- 委任の意図・条件を暗号学的な処理（電子署名やハッシュ化など）を施して保管し、事後検証できる仕組みが検討されている。

委任の証跡を残す

✓ 既存の権限委任だけでは、商取引などで責任分界できない
- ユーザーの意図・条件が、後から検証できる形で残らない
- エージェントの出力の非決定性で境界が曖昧
- 不可逆的な取引で責任の所在が確定しない

✓ 委任の制約・条件を検証可能な形で保管するなど、証跡を残す仕組みが必要



プロトコル提案や標準化の動向

✓ エージェントが購買や支払いを代行する「エージェンティック・コマース」の分野で購入意図を検証可能な形とする仕組みづくりが進行

プロトコル	提案元	概要
Agent Payment Protocol[*1]（AP2）	Google（60社以上と協力）	エージェントの決済行為に委任状(Mandate)と呼ばれる電子署名付き証明書を用いて、ユーザーの購買意図の証跡を残す
Verifiable Intent[*2]	Mastercard、Google	上記のAP2等と連携し、身元・意図・取引結果を電子署名付きのレコードにして、選択的開示により必要最低限の情報のみを検証する

✓ これらのプロトコルは標準化途上であり、決済特化など対象範囲は限定的。汎用エージェントへの拡張は今後の課題。

*1: Google Cloud. "Announcing Agent Payments Protocol (AP2)." Google Cloud Blog, 16 Sept. 2025. https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol (参照：2026/4/30)

*2: Mastercard. "How Verifiable Intent builds trust in agentic AI commerce." Mastercard News & Trends, 5 Mar. 2026. https://www.mastercard.com/us/en/news-and-trends/stories/2026/verifiable-intent.html (参照：2026/4/30)

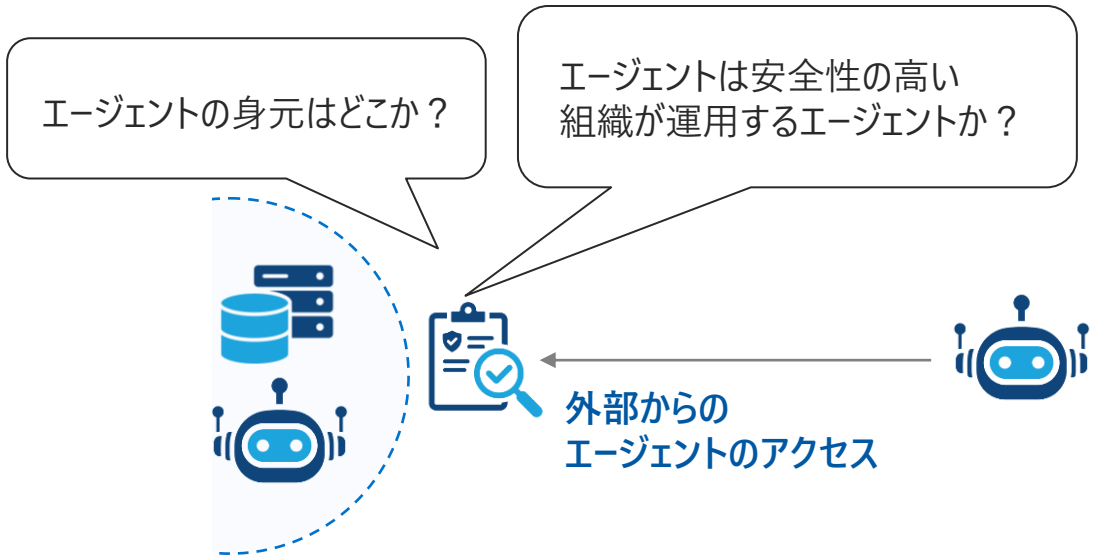
越境での信頼の確立：組織内の認証では対応できず、新しい身元検証が必要

- 高自律性エージェントは、組織の境界を越えて未知のサービスと動的に連携するが、越境時の身元検証に課題。
- 連携先のサービスは「このエージェントは何者か（エージェント・アイデンティティ）」を検証する仕組みが必要となるが、標準化は発展途上にある。

「良い」エージェントを見分ける

✓ ブラウザ型エージェントにおいて、人間に代わってWebインターフェースを利用するため、APIによる認可コントロールが機能しない例が出てきている。

✓ 「信頼のないエージェント」を排除し、正規のエージェント（安全性の高い事業者により運営されているエージェント）であることを示せるようにする動きが高まっている。



標準化の動向と残る課題

✓ 「Web Bot Auth」などWebインターフェース向けボット対策で、エージェントを見分ける取り組みは登場しつつある

Web Bot Auth　CloudflareとGoogleが共同提案するIETFドラフト*1

背景/目的：
・Web サイト側がアクセスしてきた Bot／エージェントの運営者を識別したい。
・身元確認済みの運営者からのアクセスを受け入れ、未確認のものは制限。
手段：
・　HTTPメッセージ署名でリクエストを署名し、運営者を検証可能に。

✓ エージェント・アイデンティティの基盤的な課題が複数残っており、統一的な標準化までは時間がかかる見込み。

- 業界横断のエージェント識別子の不在
- 信頼アンカ（安全な運営者を保証する仕組み）の未整備
- 越境時のアイデンティティ互換性（IdP・基盤間の相互運用）

*1: Meunier, Thibault (Cloudflare), and S. Major (Google). "HTTP Message Signatures for automated traffic Architecture." IETF Internet-Draft, draft-meunier-web-bot-auth-architecture, 2 Mar. 2026. https://datatracker.ietf.org/doc/draft-meunier-web-bot-auth-architecture/ （参照：2026/5/6）

展望１：AIエージェントへの権限委任は、OAuth拡張などを中心に「認可」の仕組みは整いつつある一方、「エージェント・アイデンティティ」の標準化が新たな鍵となる。後者については、複数の方針で検討が進むと予測される

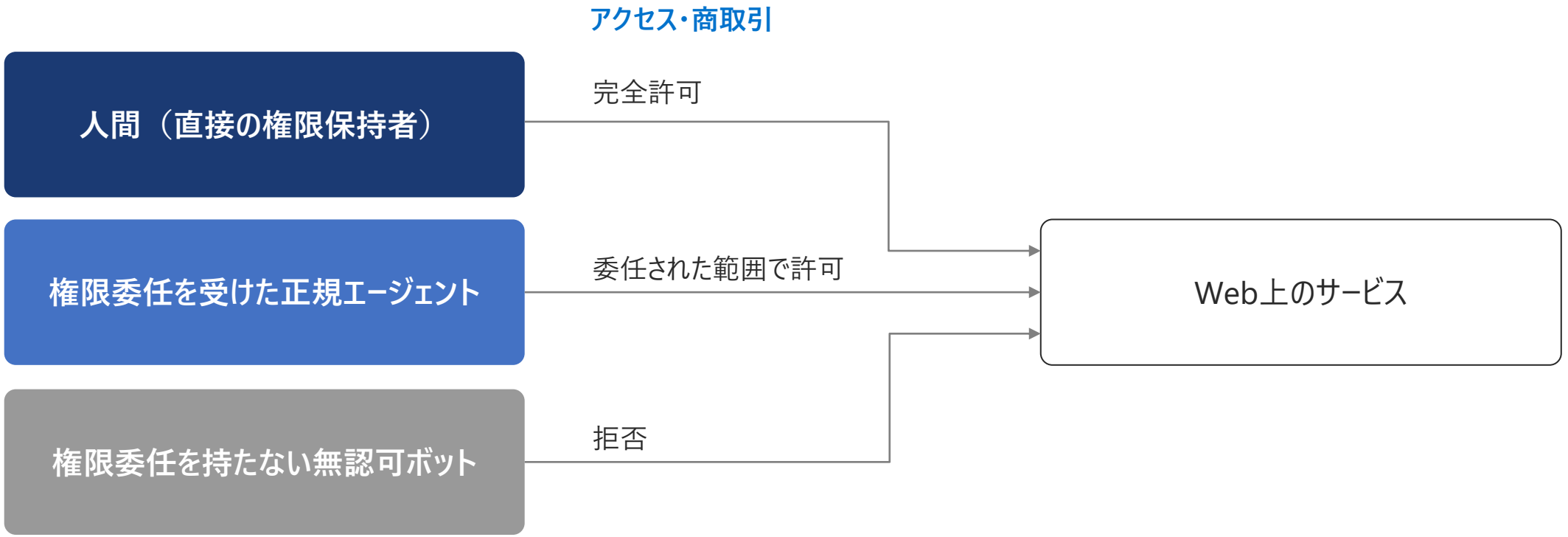
- 権限委任は、エージェント自身を識別する**アイデンティティ層（誰なのか）**と、その上に許可を表現する**認可層（何を許可するか）**の二段構造で成り立つ。
- 認可層は OAuth 2.0/2.1 拡張などを中心として標準化や既存フレームワークの活用が進む（P.7-10参照）一方、将来の高自律性エージェントを想定すると、**土台となるアイデンティティ層は議論が続く**と見られる。

アプローチ	主張の要点	強み	関連団体（例）
IdP中心（人間IDの延長）	エージェントを「人間ではないID（非人間ID）」としてシステムに登録し、人間と同じ仕組み上で既存のIdP（Identity Provider）を用いて管理する。	エンタープライズの既存の運用やガバナンス体制にそのまま乗せることができる。	IAM ベンダー大手（Okta/Auth0、Microsoft Entra等）、標準化団体（OpenID Foundation）
SPIFFE[*1]/WIMSE[*2]（インフラ層で証明）	エージェントを単なる「ソフトウェア（ワークロード）の一つ」としてとらえて、エージェントが動く「環境（サーバーなどインフラ）」が安全・正当であるかをチェックし、短い時間だけ使える「身分証明（クレデンシャル）」を発行。 *1: Secure Production Identity Framework For Everyone：クラウドネイティブ環境において、ソフトウェア（ワークロード）に短時間で失効する暗号学的な ID を自動発行する業界標準。 *2: Workload Identity in Multi-System Environments：SPIFFEの概念をクラウド単一環境だけでなく、組織を越えた複数の環境に拡張するためのIETFワーキンググループ。	暗号学的に強固であり、自動化しやすく、既に本番環境での実績がある。	OSS コミュニティ（SPIFFE/CNCF）、標準化 WG（IETF WIMSE）
VC[*3]/DID[*4]（分散IDを利用）	エージェントは特定のIdPに依存しない「持ち運び可能なクレデンシャル」を持つべきであり、第三者でも検証できる仕組みが必要である。 *3: Verifiable Credentials：暗号学的に検証可能な資格情報の仕様。発行者・所有者・検証者の三者モデルで、第三者でも改ざんの有無を検証できる。 *4: Decentralized Identifier：中央の認証機関に依存せず、所有者自身が制御できる分散型識別子の仕様。	組織を超えたIDのやり取りができ、第三者検証が可能であり、プライバシー保護も組み込みやすい。	標準化団体（W3C、DIF）

※上記に加え、AI エージェント専用プロトコル（MCP：Anthropic 主導、A2A：Google 主導）で、プロトコル自身の中で独自のアイデンティティ枠組みを定義する動きもある。

展望２：AIエージェントへの権限委任の標準化は、Web上のアクセス・商取引を、「人間／権限委任を受けた正規エージェント／無認可ボット」の三構造へと再編する可能性

- 「正規エージェント」として受け入れられるには、ユーザーや組織からの明示的な権限委任と、その検証可能性（誰が運営し、何の権限を委任されているかを暗号学的に証明できる仕組み）が前提となる。
- 今後、Webアクセス・API利用・経済活動へのエージェントの参加可否が、**権限委任の有無や検証可能性によって分化していく**可能性。Web Bot Auth（P.14）やVerifiable Intent（P.13）など、すでにこの方向性を見越した取り組みが始まっている。



展望３：AIエージェントへの権限委任の枠組みが整い、エージェントを用いた経済活動が活性化した社会では、「KYA（Know Your Agent）」の概念が重要になると予測する

- エージェントの所有者・権限などを取引前に検証するプロセス**「KYA（Know Your Agent）」が提唱され始めている。**
- エージェント同士の取引では、エージェントの能力差によって不利益な状況が生まれても、利用者は損に気が付きにくい。
  KYAによって、不適切な取引の排除だけでなく、より適切な取引戦略の立案も可能になると考えられる。

KYA（Know Your Agent）の提唱

✓ KYC/KYBのアナロジーとして、適切な取引相手のエージェントかを見極める「KYA（Know Your Agent）」が提唱され始めている。

銀行・カード

2026年1月、米FIS（Fidelity National Information Services）が**Visa、Mastercard**と提携。発行銀行向けに「**Know Your Agent (KYA) data**」を活用した取引認可・不正検知ソリューションを発表[*1]。

*1: Fidelity National Information Services (FIS). "FIS Launches Industry-First Offering Enabling Banks to Lead and Scale in Agentic Commerce." Press Release, 12 Jan. 2026. https://www.investor.fisglobal.com/news-releases/news-release-details/fis-launches-industry-first-offering-enabling-banks-lead-and/ (参照：2026/5/8)

KYC/KYB業界

KYC/KYB業界のプレイヤーがKYA関連の取り組みを開始：

- Trulioo：「Digital Agent Passport」KYAのフレームワーク開発[*2]（2025年8月）
- Sumsub：「Agent-to-Human Binding」検証済み人間 ID に紐付け[*3]（2026年1月）

相手エージェントを知ることは戦略面でも重要

✓ 2025年12月、アンソロピック社がAIエージェント同士の取引に関する実験を実施（参加者69名、186件、約$4,000相当の取引が成立）
✓ 参加者は、マーケットプレイス上でエージェントに商品の売買を委任。ただし、**互いのエージェントに能力差があることは隠されていた。**

強いモデル（Claude Opus 4.5）に代理されたユーザーは取引で優位に。
一方で、弱いモデル（Claude Haiku 4.5）に代理されたユーザーは
不利な取引をしていることに気がつかなかったという

（出所）Anthropic. "Project Deal: our Claude-run marketplace experiment." Anthropic Features. https://www.anthropic.com/features/project-deal (参照：2026/4/30)

相手のエージェントの能力・所有者・権限を検証する仕組み（KYA）があれば、取引を分析し、適切な販売戦略の立案に活用できる可能性

*2: Trulioo and PayOS. "Know Your Agent (KYA): An Identity Framework for Agentic Commerce." White Paper, 2025. https://www.trulioo.com/resources/white-papers/know-your-agent-an-identity-framework-for-trusted-agentic-commerce (参照：2026/5/8)

*3: Sumsub. "Sumsub's AI Agent Verification Introduces Agent-to-Human Binding to Establish Human Accountability in AI." PR Newswire, 29 Jan. 2026. https://www.prnewswire.com/news-releases/sumsubs-ai-agent-verification-introduces-agent-to-human-binding-to-establish-human-accountability-in-ai-302673467.html (参照：2026/5/8)

展望３（参考）：何がKYAされるのか、エージェントの検証対象

- 定義は定まっていないものの、KYAは、エージェントの身元・能力・権限・責任主体を取引前に検証するプロセスの概念であり、エージェントが安全な商取引を行うために重要になると展望する。
- 本レポートを踏まえて、KYAが行われるようになる際には、どのような項目がエージェントの検査対象となり得るかを示す。

エージェントID このエージェントは誰か？	標準化された検証可能な一意の識別子。なりすまし防止と監査の起点となる。 （→P.15の通り、エージェントアイデンティティに関する議論が開始）
運営者の認証 誰がデプロイ・運用しているか？	エージェントを動かしている運営者（事業者、開発者）の身元確認。 （→P.14の通り、運営者を識別する標準化が始まっている）
能力評価 何ができるか？	モデル世代やツール一覧、権限スコープを宣言・検証する。能力格差の可視化に必要。 （→P.17の通り、モデルの能力差は見えない格差を生む可能性）
権限の検証 どこまで取引・支払が許されているか？	ユーザから事前委任された権限の範囲・上限・有効期限を検証する。 （→P.13の通り、権限委任を事後検証できる仕組みが検討されている）
責任主体の紐付け 背後の人間／法人は誰か？	事故・不正発生時の責任の所在を明らかにする必要がある。AML/CFT対策にも不可欠。 （→P.17の通り、エージェントと責任主体を紐づける取り組みが始まっている）

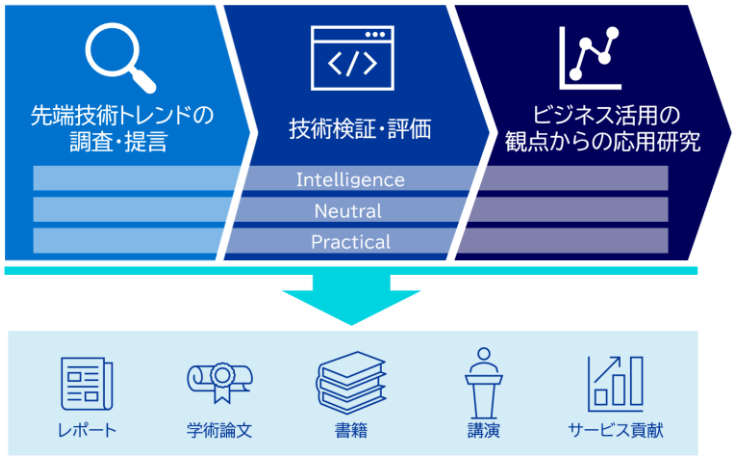
おわりに

- 本レポートでは、AI エージェントの自律性が高まる時代における認証・認可の課題を、中・高自律性に分けて整理した。中自律性については OAuth 拡張で対応可能な範囲が見えてきた一方、高自律性ではスコープ動的化・責任分界・組織を超えた信頼の三点で、業界横断の標準化はいまだ模索段階にある。これらの整理を踏まえ、エージェントが経済主体として振る舞う時代に必要となるKYA（Know Your Agent）の概念的必要性を提示した。

- 本論の核心は、**権限委任の枠組みが整うほど、「相手のエージェントを評価する仕組み」が必要となる**点にある。
これは、権限委任が単なる認可の問題から、信頼の問題へと移行していくことを意味する。中自律性で扱った権限委任は、結局のところ「自組織の」エージェントに何を委ねるかの問題であり、高自律性が広がるほど、「組織を超えて信頼を担保する」仕組みが必要となる。そこでは、「相手のエージェントをどう評価するか」が業界全体の論点として浮上し、**KYAは、その解決策の一つであると考えられる。**

先端技術ラボのご紹介

先端技術を活用したITサービスの創出に向けた技術の目利き役として、
「先端技術トレンドの調査・提言」、「技術検証・評価」、「ビジネス活用の観点からの応用研究」に取り組んでいます。



主な研究・取り組み領域

当社ホームページの 特集サイト では、IT分野における先端技術の調査レポート、及びメンバーのプロフィール詳細がご覧いただけます。
本レポート執筆者への取材や講演などに関するご相談は、当社ホームページの お問い合わせフォーム よりご連絡ください。

株式会社日本総合研究所

日本総研は、シンクタンク・コンサルティング・ITソリューションの3つの機能を有するSMBCグループの総合情報サービス企業です。

東京本社　〒141-0022 東京都品川区東五反田2丁目18番1号 大崎フォレストビルディング
大阪本社　〒550-0001 大阪市西区土佐堀2丁目2番4号

