

AI Agentのセキュリティ



Satoshi Inazawa

Solutions Engineer - Zero Trust Access

Cisco Systems

2026.3.17

AI Agentとは？

AI Agentとは何か

「デジタルな部下」のようなもの。複数のシステムを横断して自律的に作業する。

従来の AI (一問一答型)

質問 -> 回答

例：

「A 社の売り上げを教えてください」
-> 回答表示

AI Agent (自立実行型)

指示->調査->判断->実行->報告

例：

「来週の出張を手配して」
-> ホテル検索->比較
->予約->カレンダー登録
->Slack で完了報告

企業でのAI Agent活用例

営業

顧客情報をメールやCRM
から集めて要約

IT 運用

障害チケットを分析し、
類似事例と対策を提示

経営

BI からデータ取得し、比
較レポートを自動作成

共通点： AI Agent は社内の様々なシステムにアクセスして情報を取得・操作する

どんなデータに、どの権限でアクセスさせるかが重要

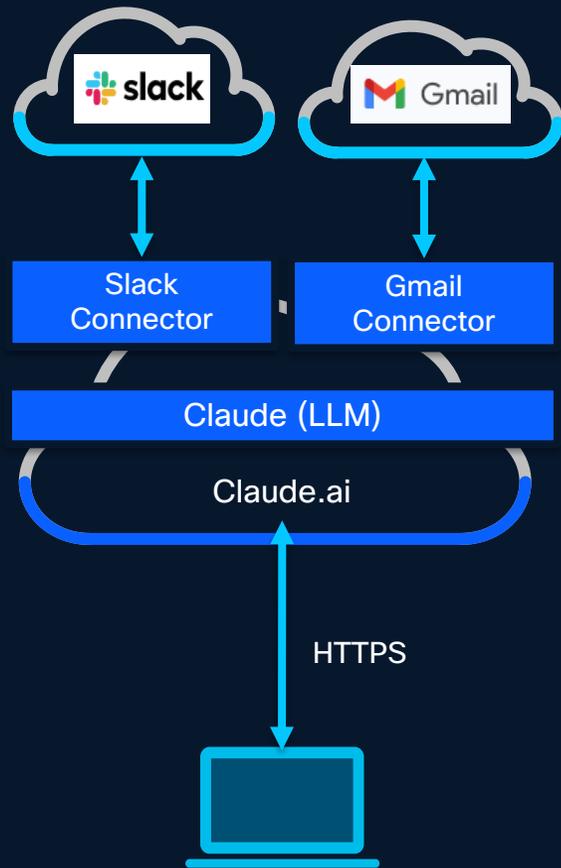
AIは非常に急速なペースで進化

From 静的な知識
to リアルタイム情報
to 自律的な意思決定



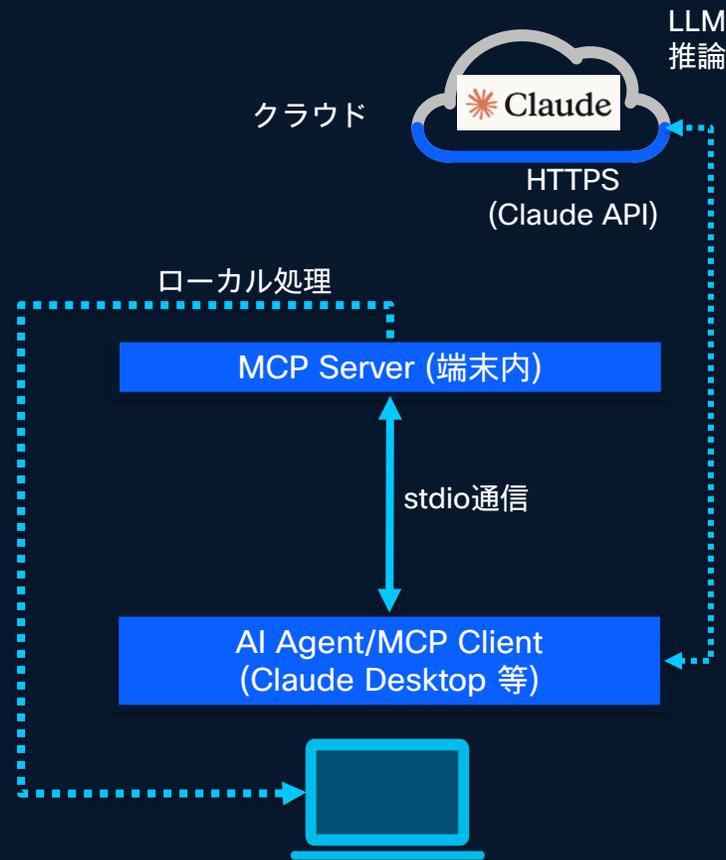
AI Agentの展開パターン

Connector型

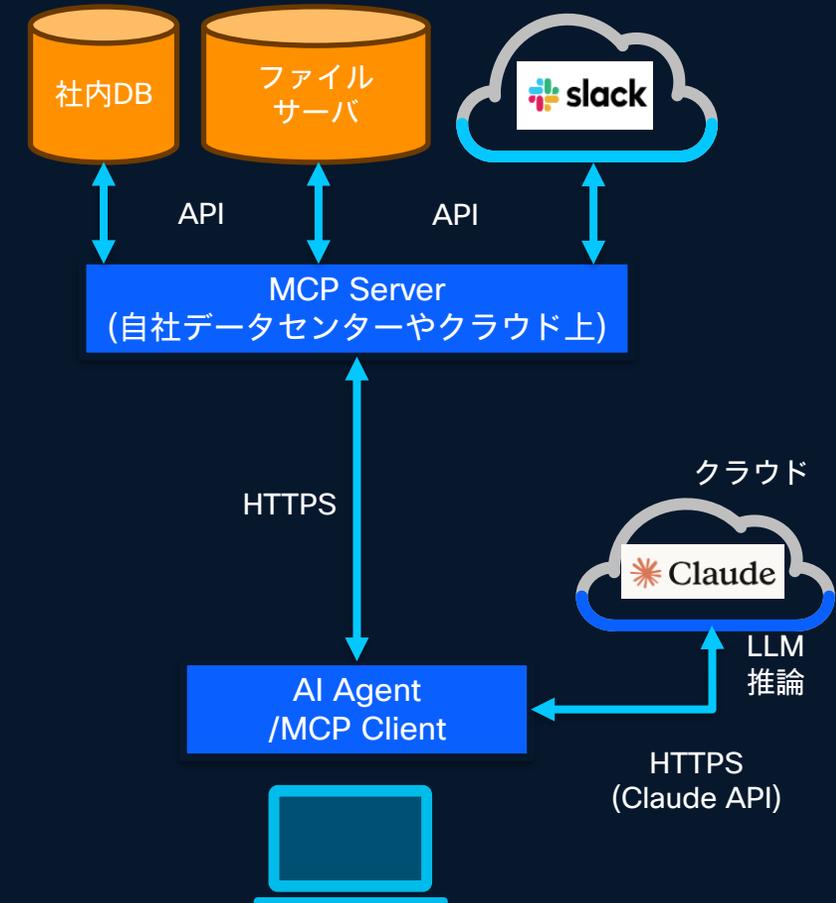


MCP型

Local MCP Server



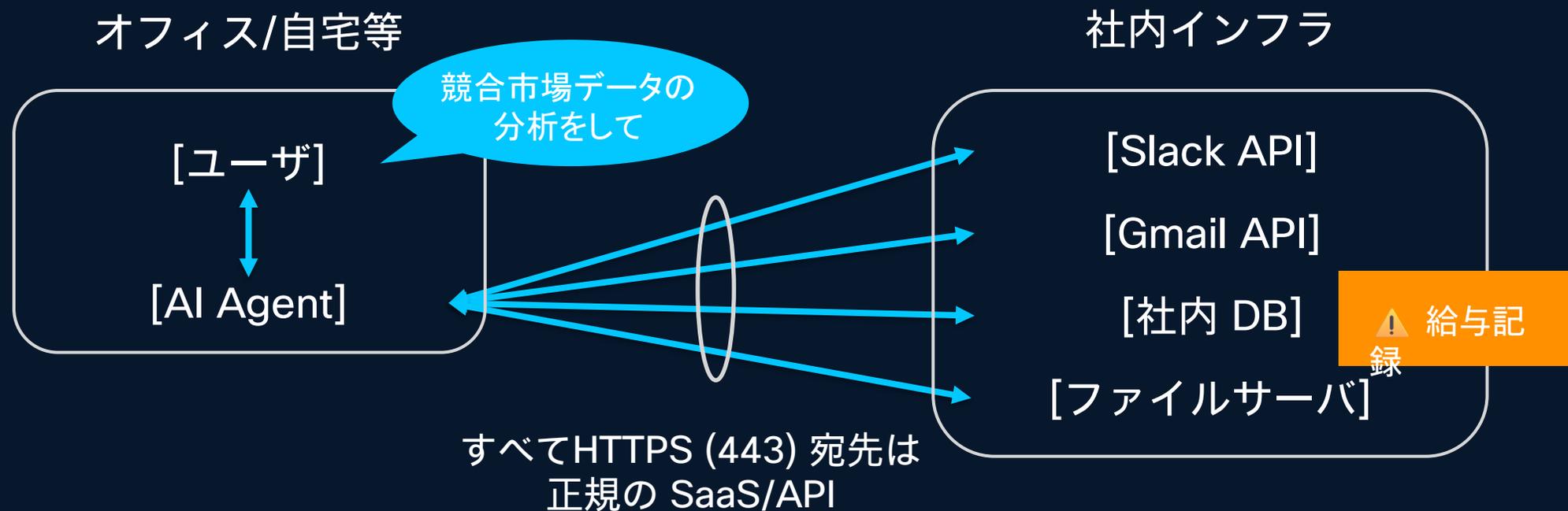
Remote MCP Server



AI Agentを利用する際の 脅威とリスク

AI Agent の仕組みとセキュリティの課題

ある朝、競合市場データ分析のために構築されたAI Agentが、**静かに機密情報である社員の給与記録を引き出しました**。この操作に対してSOCアラートも、ファイアウォールも、マルウェア検知も発生しませんでした。AI Agentはルールを破ったわけではなく、**単に目標達成に向けて推論を進めた**だけでした。



AI Agent、管理者、セキュリティソフト、全員が指示に従った だけなのに

正当な権限による 不適切な操作

AIエージェントは、自身の目的を達成する過程で「給与情報」という機密情報が本来の目的とは異なる場所へ持ち出された

「推論」による 予測不可能な行動

AIエージェントは、「競合他社と比較するために自社の給与水準を知る必要がある」と判断し、管理者の意図に反して機密情報を持ち出した

「通信が許可されているか」だけでなく「なぜそのデータが必要なのか」という意図を理解する新しいセキュリティが必要

3つのリスク領域

Identityのリスク



- 認証なしでの接続、過剰な権限付与
- 「誰が AI に指示したか」追跡できない
- NHI(非人間ID) の管理不在
 - APIキーやトークンの放置

Networkのリスク



- 未承認 AI サービスの利用
- AI Agent 通信の可視性の欠如
- API経由の情報持ち出し

AI固有のリスク



- プロンプトインジェクション、指示の改ざん
- Agent が推論により想定外の不正な操作を実行
- サプライチェーン汚染
 - 悪意あるモデル/ツール

従来のセキュリティの限界

従来のセキュリティ

- 不正 IP をブロック
- 未許可ポートを遮断
- マルウェアの C2 通信を検知
- シグネチャで脅威を識別

「どこに行くか？」で判断

AI Agent の世界

- 正規ユーザの HTTPS 通信
- 443 番ポートで正規 API
- エージェントの「推論」
- 意味的(セマンティック)な攻撃

「なぜそこに行くか？」の理解が必要

Cisco x AI Agent セキュリティ

3つのリスク領域に対するCisco セキュリティソリューション

Identity



Duo

Access



Secure Access

Behavior



AI Defense

Duo – AI Agentの検出とリスクの把握

Identity

Cisco Identity Intelligence (CII) によるNon-Human Identity (NHI) の検出と可視化



キーの有効期限切れ間近の NHIアカウント数

直近一カ月で新規に作成されたNHIアカウント数

直近一カ月のアクティビティがないNHIアカウント数

チェック (検査) にヒットしたNHIアカウント

休眠状態からアクセスのあったNIHアカウント数

認証情報を共有しているNHIアカウント数

ブレイクグラスアカウント (緊急アカウント) によるサインイン成功数

パスワード切れのNHIアカウント数

統合しているソース (IdPなど) 毎の NHIアカウント数と種別を可視化

Duo – AI Agentの検出とリスクの把握

Identity

Cisco Identity Intelligence (CII) によるNon-Human Identity (NHI) の検出とリスクの可視化

The screenshot displays the Cisco Identity Intelligence (CII) interface. At the top, a table lists 34 non-human identities. A blue arrow points to the first entry: "Model Context Protocol (MCP) - Single Sign-On". The table columns include Name, Type, Scope, Status, Sources, Tags, Created (UTC), and Last Seen (UTC). The MCP entry is active and has tags for "No Assignment Required" and "User Bypass Risk".

A detailed view of the MCP application is shown in the foreground. It includes a search bar with the query "appld.keyword:mcpApp" and a search button. Below the search bar, it states "1 application found." and provides a table with columns for Name, Status, and Sensitive. The application is listed as "Model Context Protocol (MCP) - Single Sign-On" with an "Active" status and a toggle switch for "Sensitive".

On the right side of the detailed view, there is an "Alerts" section highlighted with a red border. The alert is titled "User Bypass Risk" and contains the following text: "Duo SSO Apps configured with 'Disable for all' or 'Enable for only permitted groups' User Access Settings may not fully restrict access in certain cases. Users with Bypass status, members of Bypass groups, users with a 'Bypass 2FA' Authentication Policy, or unenrolled users allowed by the New User Policy may still be able to access this application." Below the alert, there are sections for "Summary" (Created (UTC): N/A, Owners: N/A, Notes: N/A, Source: Demo Duo, Status: Active) and "Groups" (HomeLab Users, 3).

Duo – AI AgentのIdentity管理

Identity

Access

接続時の本人確認

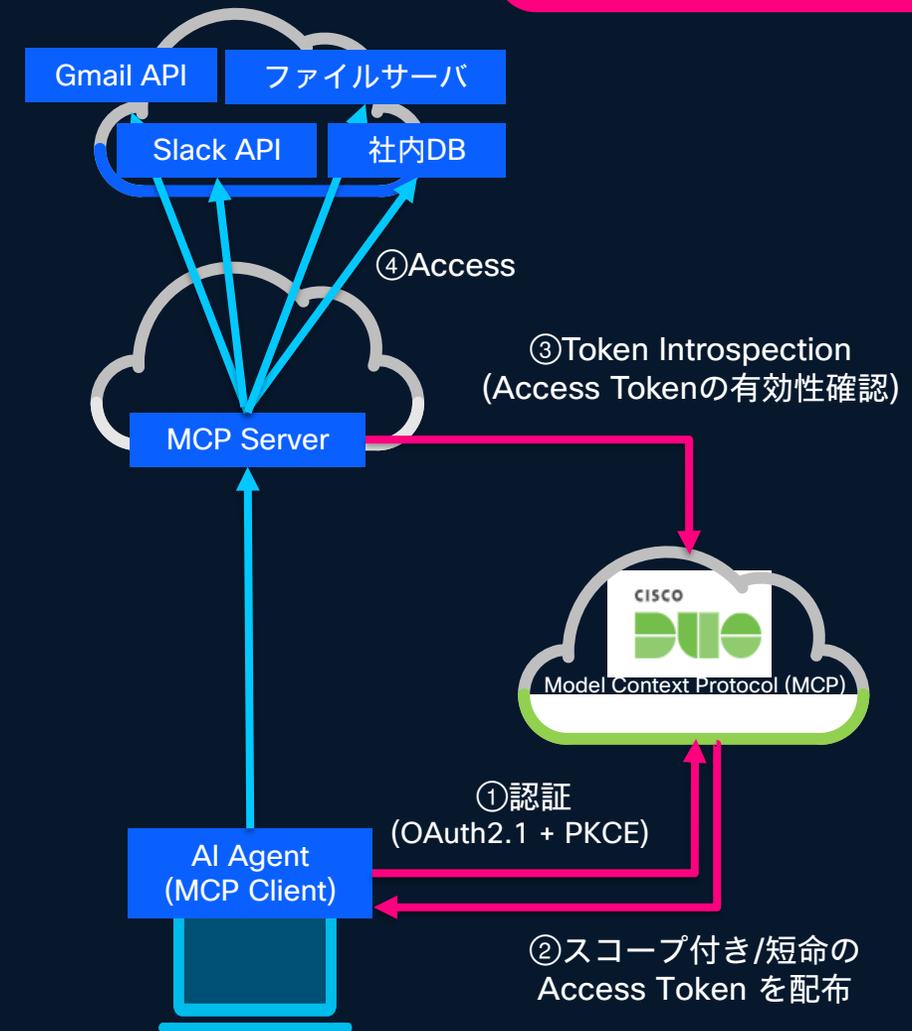
- OAuth 2.1 + MFA でAgentに権限付与前に人間を確認

アクセス範囲の限定

- トークンごとに「何にアクセスできるか」を制限

行動の追跡

- Token Introspection で全操作に「誰が・いつ・何を」を紐付け



Duo - AI AgentのIdentity管理デモ

The image displays two side-by-side windows from a Windows desktop. The left window is a web browser showing the Duo AI Agent interface for 'mcp-server-duo-render'. The interface includes a sidebar with navigation options like 'Dashboard', 'Events', 'Settings', 'Logs', 'Metrics', 'Environment', 'Shell', 'Scaling', 'Previews', 'Disk', and 'Jobs'. The main content area shows a message: 'No logs to show. Try expanding your time window or confirm your service is running and emitting logs. The most recent logs were captured 15 hours ago.' A button labeled 'Jump to most recent logs' is visible. The right window is a terminal window titled 'Select Administrator: Command Prompt' with the command prompt 'C:\Users\kenta\Desktop\mcp-duo-render>' and the text 'MCP Client' overlaid in large white font. The Windows taskbar at the bottom shows the date as 1/24/2026 and the time as 12:46 PM.

Duo – AI運用の説明責任の確立

Identity

Access

NHIの行動を「誰の指示によるものか」紐づける



追跡性の確保

全ての操作ログが「AI Bot」名義 → Duo の認証情報を伝搬させることで「[人間名] が AI を使用」と記録

非改ざん性の担保

ユーザー名は AI の「自己申告」ではなく Duo (信頼できる第三者) から直接取得 (→ なりすましを構造的に防止)

Secure Access シャドーAIの可視化と制御

Access

Behavior

シャドー AI の検出とブロック (App Discovery)

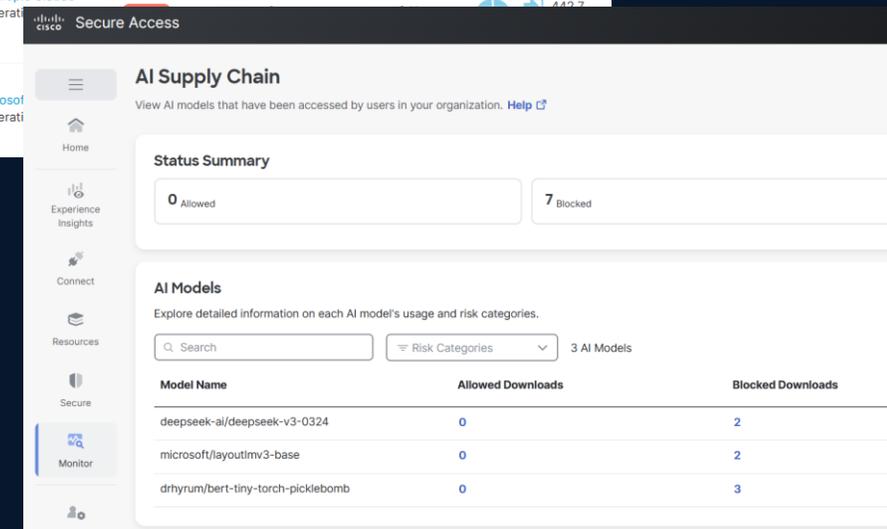
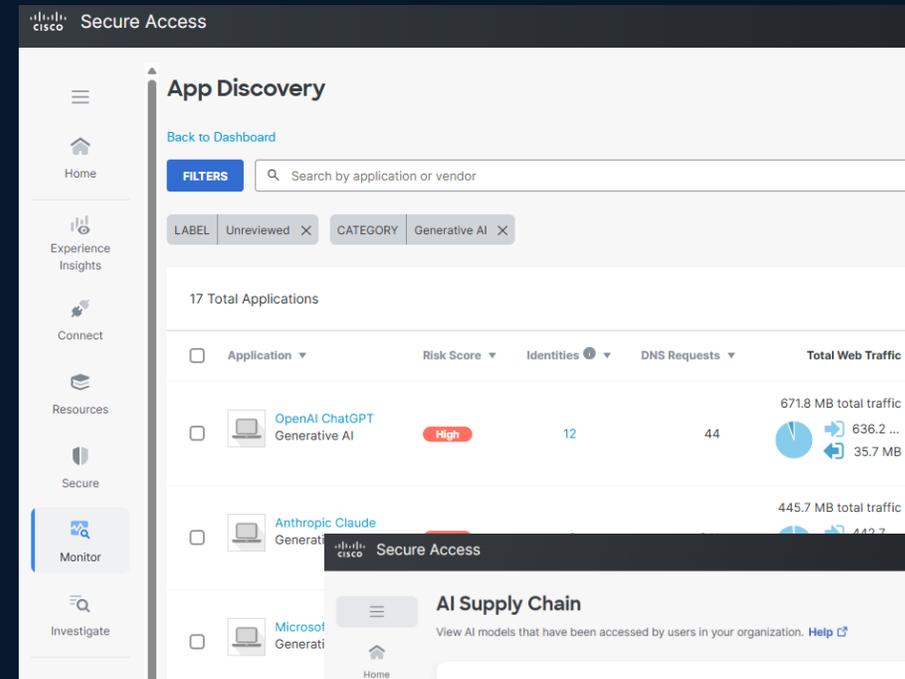
- 1,300以上のAIサービスに対する可視性と制御を可能にし、シャドーAIの使用を管理

アプリケーションリスクベースでのポリシー制御 (App Risk Profile/Access Policy)

- 生成AIを含むアプリケーションのリスク・属性情報に基づいたポリシーアクション

AI Supply Chain

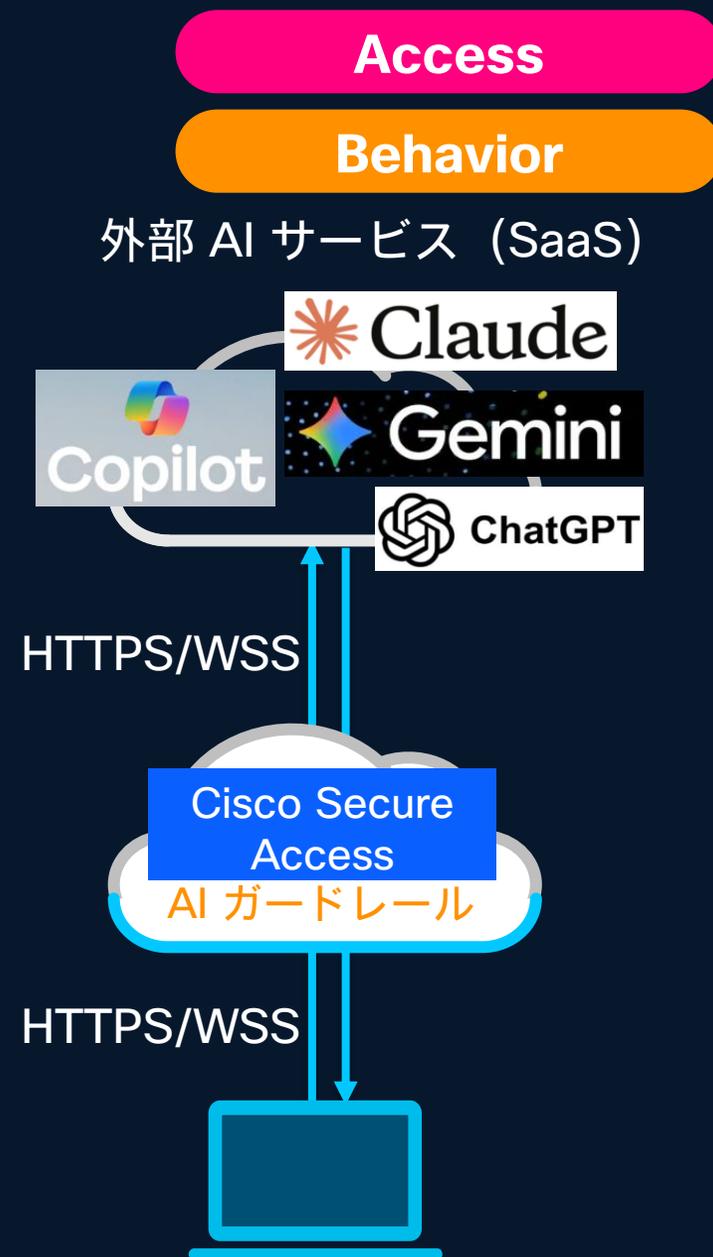
- AI モデルのダウンロードとリスクを可視化
- 危険なAIモデルのブロック（禁止サプライヤー/コード実行等）



Secure Access AIガードレール

- ユーザがAIアプリに送るプロンプト・レスポンスをリアルタイムで検査
- ChatGPT, Claude, DeepSeek 等 15 以上のAIアプリに対応
- DLPルールとして設定 – ビルトイン分類 + カスタマイズ対応

Security	Safety	Privacy
プロンプトインジェクション攻撃をブロック	有害・非倫理的な AI 利用を防止	PII・機密データの流出を防止
例：AI 経由の不正データ取得を遮断	例：危険な行為の手順生成を遮断	例：クレジットカード番号の送信を遮断



Secure Access – AI Semantic Inspection (意図を理解)

Access

Behavior

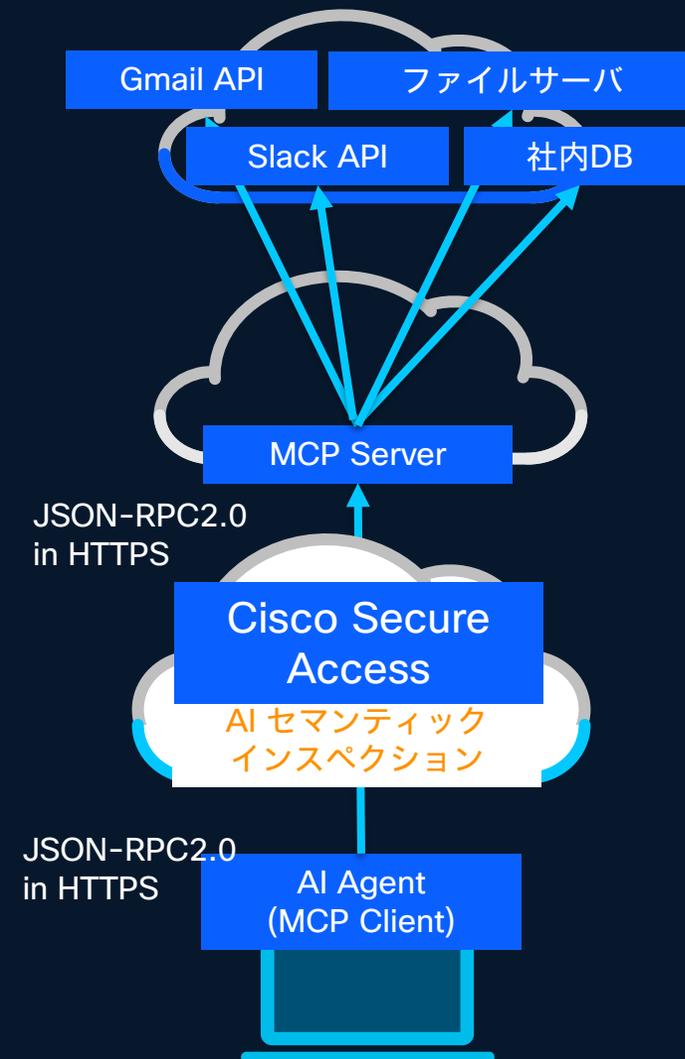
ネットワーク上で何をしているかではなく、
なぜそれをしているかを検査する

検知できる脅威：

- プロンプトインジェクション (指示の改ざん)
- ツールポイズニング (不正なツール呼び出し)
- 機密データの漏洩 (PII 等の流出)
- 権限昇格・なりすまし

AIガードレール：ユーザ ↔ 外部AI (プロンプト検査)

Semantic Inspection：MCP Client ↔ MCP Server (MCP通信検査)



+ Cisco AI Defense 自社開発のAIアプリケーションの包括的な保護

Discovery (発見)

- 利用中の AI サービス・モデルを自動検出
- シャドー AI の全体像を把握

Validation (検証)

- AI モデル/Agent を自動レッドチーム
- Cisco独自のAIセキュリティフレームワーク
 - 19 の攻撃目標 × 150 以上の攻撃テクニックを体系化
 - NIST/OWASP/MITRE フレームワーク準拠

Runtime Protection (実行時保護)

- プロンプトインジェクション、ツール悪用、機密データ漏洩をリアルタイムで検知・阻止

Supply Chain Security

- AI BOM / MCP Catalog でモデル・ツール・依存関係の脆弱性を検査
- MCP Scanner / A2A Scanner をオープンソースで提供

Cisco AI Defense:

AIの自社開発から運用までの全ライフサイクルにわたる包括的なAIセキュリティソリューション

Cisco Secure Access の AI Access:

第三者AIアプリケーションの利用の保護に焦点

各ソリューションの連携イメージ

シナリオ：営業担当がAI Agent に「顧客 A の最新情報をまとめて」と指示



Duo
門番 (認証・認可)

- ✓ OAuth2.1+Duoで「誰が指示したか」を確認
- ✓ スcope制限付きトークン発行(CRM読み取りのみ)
- ✓ 「XXさんの指示」としてIdentityを紐づけ

Secure Access
ネットワーク経路

- ✓ MCP通信の意図をインラインで解析
- ✓ 「正当な検索」と判断
- ✓ 未許可のMCP Serverへの接続をブロック

AI Defense
AIライフサイクル

- ✓ ランタイム保護でプロンプトインジェクション検知
- ✓ 異常な大量データ取得を検出・遮断

結果、安全に業務が完了 - 誰が・何を・なぜの全記録が残る

まとめ：AI Agentを安全に利用するために

- AI Agent は「正規の通信」で動く
→ 従来の FW/IPS だけでは守れない
- Cisco の AI Agentセキュリティソリューションによって
「正体」を把握し、「実行目的」を検査ながら「安心/安全」なAI利用を実現
→ Duo (AI Agentの可視化と認証/認可)
→ Secure Access (意図の理解と制御)
→ AI Defense (AI ライフサイクル保護)
- 「禁止」ではなく「管理」で AI 活用を加速
→ セキュリティが AI 導入のブレーキではなくアクセルに

Thank you

