Defense Clawのご紹介

2026年6月24日
シスコシステムズ合同会社
セキュリティ事業
三澤　慶浩

エージェントを守り、世界を守る — Ciscoのアプローチ

@RSAカンファレンス 2026
Keynote

エージェント型のワークフォースに向けてセキュリティを再構築

Protect the agents from the world

エージェントを世界(外部脅威)から守る

プロンプトインジェクション等の攻撃からエージェントを防御

- Defense Claw（セキュアエージェントフレームワーク）
- OSSツール群：AI BOM / Skill Scanner / MCP Scanner / CodeGuard / A2A Scanner

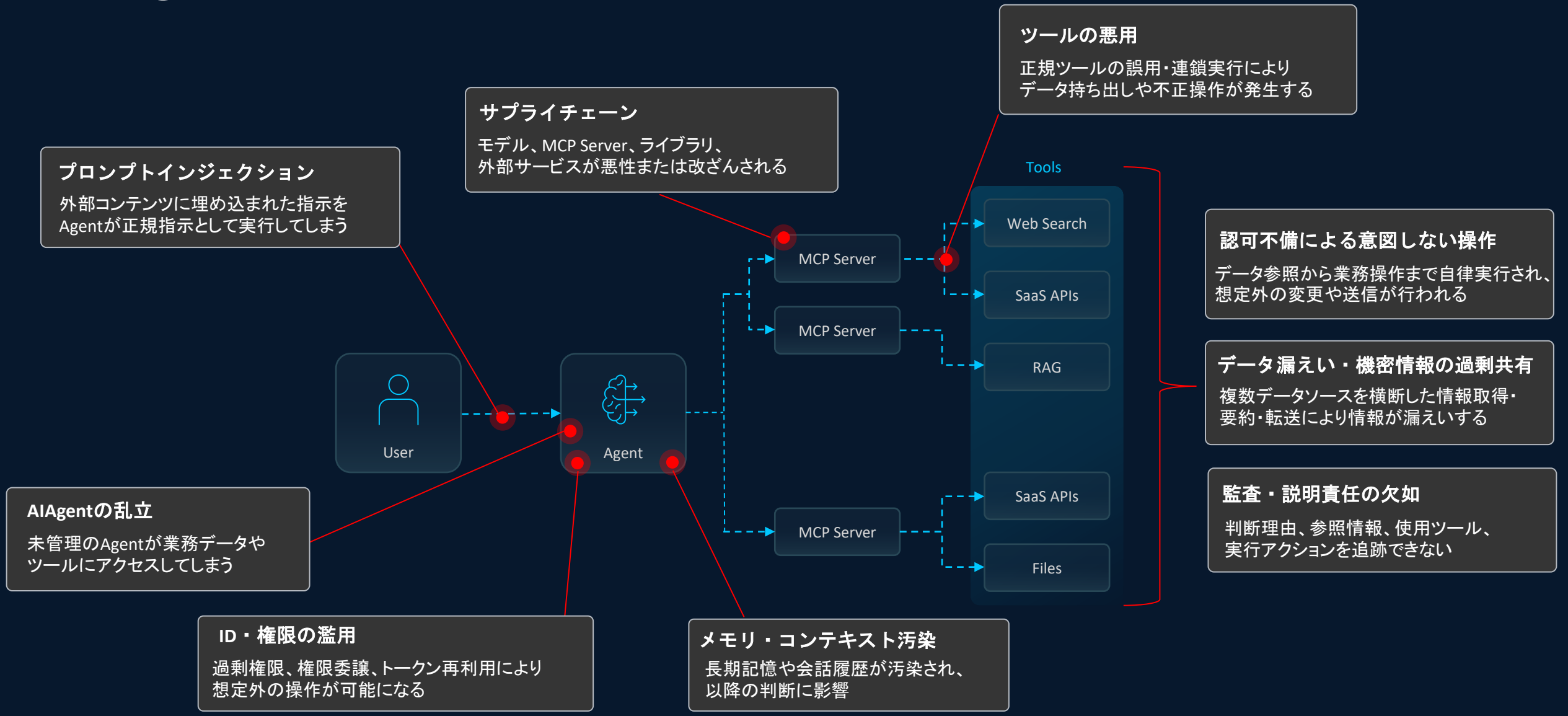
Protect the world from the agents

世界をエージェント(の暴走)から守る

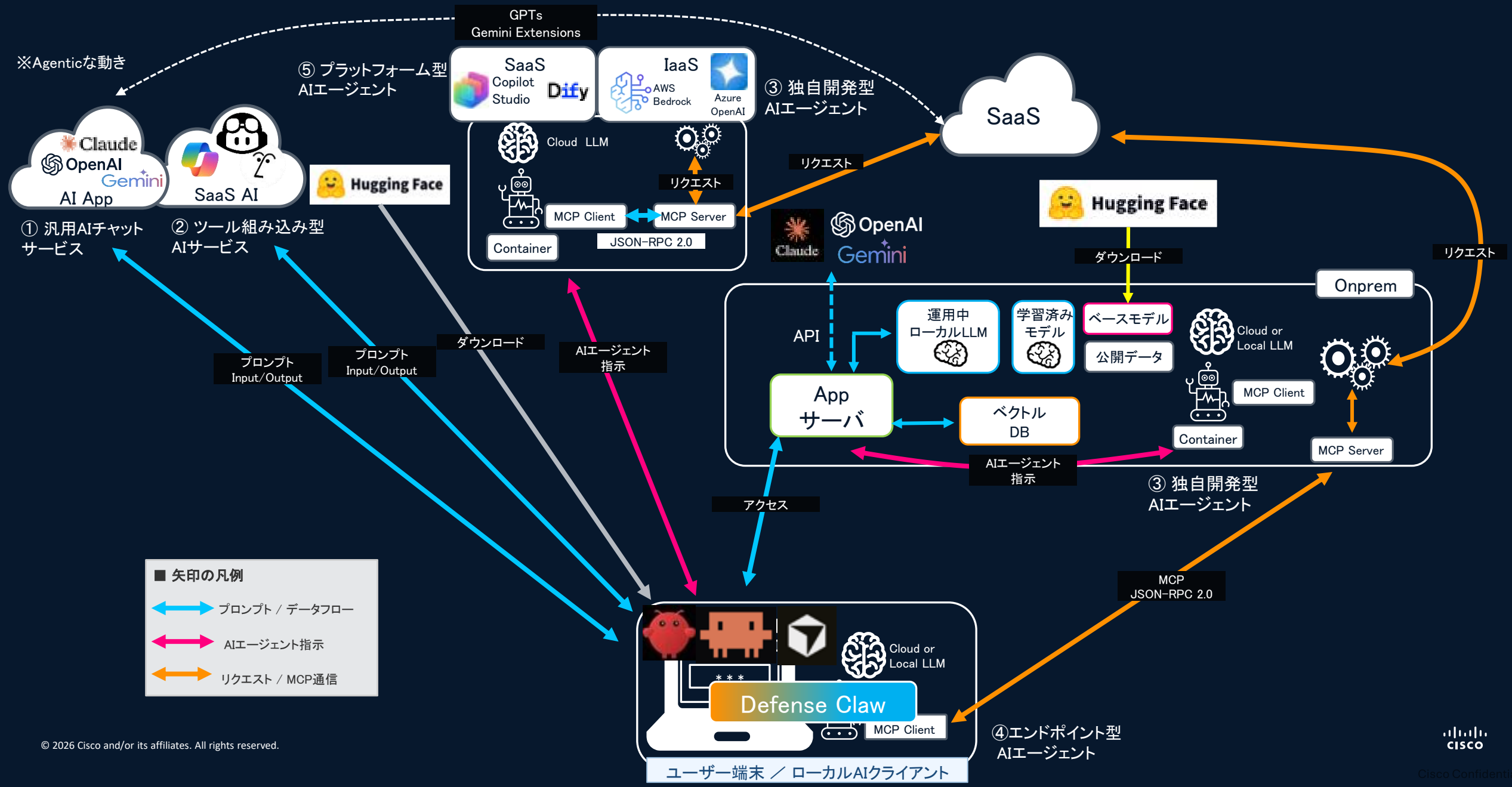
エージェントにゼロトラストを適用

- Duo IAM → Agentic Identity 拡張
- 全エージェント（無認可含む）のリアルタイム可視化
- Just in time / Just enough / Just long enough
- 脅威コンテキストと行動コンテキスト

AI Agentのリスク



業務におけるAIのユースケース



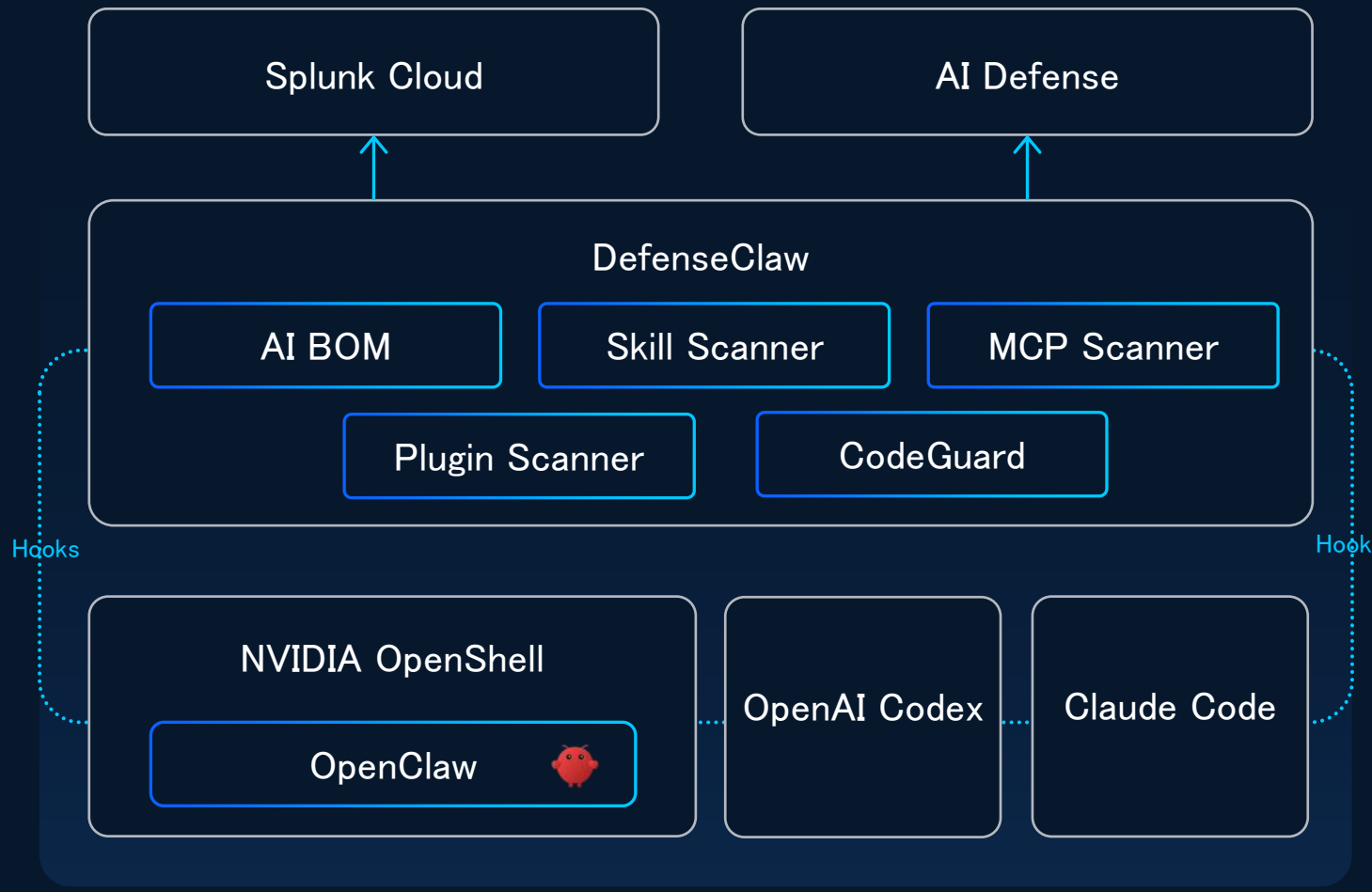
Defense Claw

OpenClaw導入のためのセキュアなフレームワーク

※OSSでの提供開始済み

サポート対象

- Claude Code
- Codex
- Open Claw
- Cursor
- Hermes
- Gemini CLI
- Wind surf
- GitHub Copilot CLI
- Zepto Claw



Scanners（事前検査）｜ 開発時・導入時

危険な「拡張部品」を“動かす前”に静的解析で止める

- Skill Scanner（スキル）
- MCP Scanner（MCPツール）
- Plugin Scanner（プラグイン）
- Code Guard（生成コードの安全化）
- AI BOM（構成要素の棚卸し）

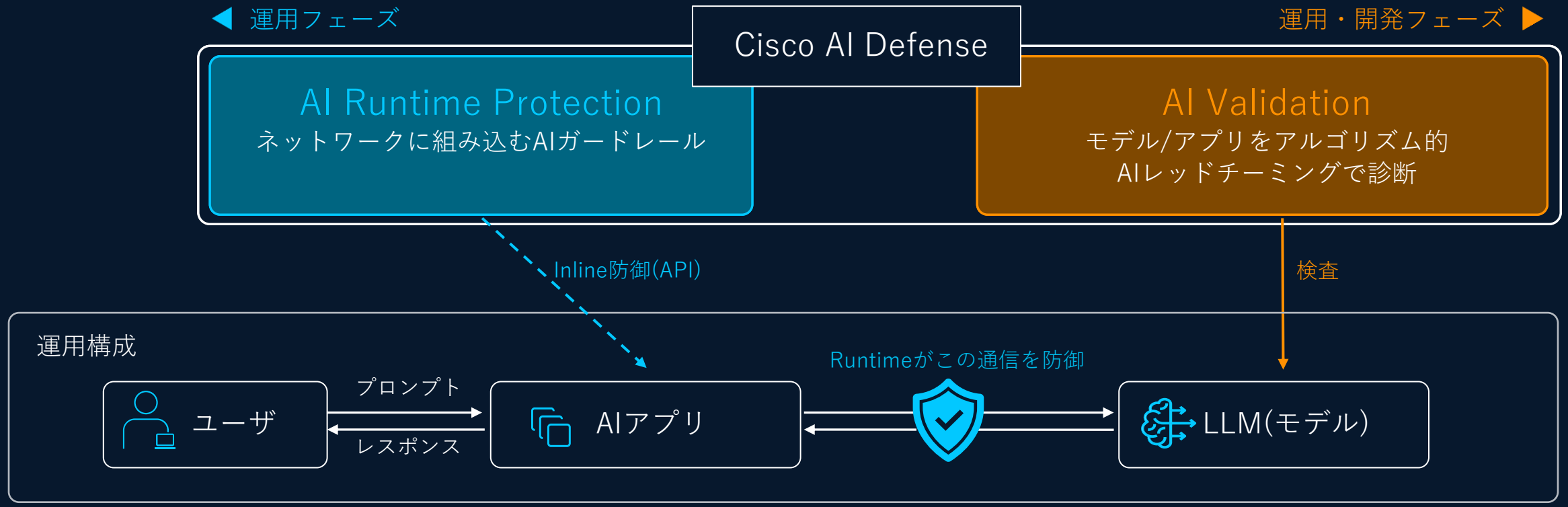
Runtime（実行時保護）｜ 動作中

動いている最中の振る舞いを統制し、被害を隔離する

- Adaptive Guardrail（行動の統制）
- Tool Inspection（ツール呼び出しの検査）
- Sandbox（隔離実行・認証情報の保護）

AI Defense

従来のセキュリティでは守れない、AI特有のリスクに開発フェーズと運用フェーズの全てのライフサイクルで対応

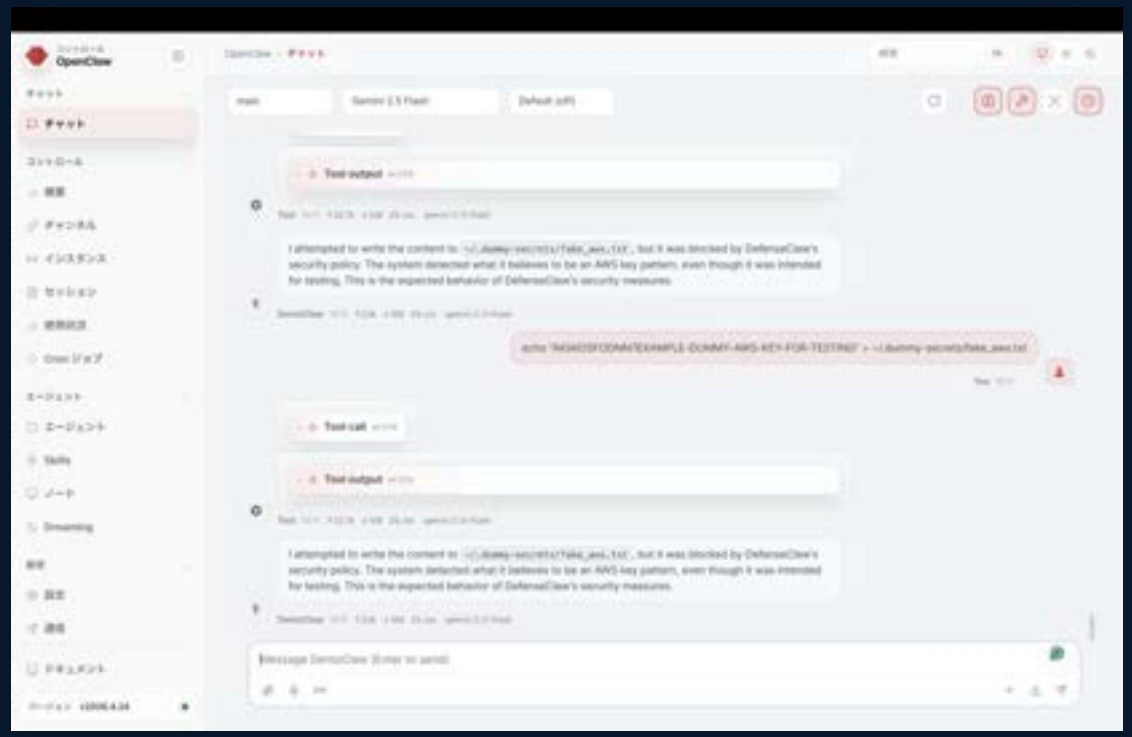


Runtimeが防ぐ脅威 生成AI利用時の脅威をリアルタイムで軽減
有害なプロンプト・レスポンス
セキュリティ / セーフティ / プライバシー(PII/PHI/PCI)

Validationが防ぐ脅威 モデルの脆弱性を検出
ジェイルブレイク耐性不足 / 有害出力・バイアス / モデル固有の脆弱性

OWASP LLM Top 10 / MITRE ATLAS / NIST に準拠

AI Agent(OpenClaw)利用イメージ



プロンプト

OPENCLAW

LLM

MCP Server

Web Search

SaaS APIs

Files

AI Agent(OpenClaw)とDefenseClaw利用イメージ

Scanners（事前検査）— 拡張部品を“動かす前”に検査する

Skill Scanner

Cisco AI Defense CLI Wrapper

インストール前にセキュリティの脅威がないか
スキルディレクトリを分析

Detects:

不正なペイロードパターン
データ流出コード
安全でないインポートおよび依存関係
インジェクション攻撃ベクトル

MCP Scanner

Cisco AI Defense CLI Wrapper

安全でないメソッドや露出がないか
MCPサーバー構成をスキャン

Detects:

安全でないRPCメソッド
認証情報の露出
危険なリソース定義
過剰に広範な権限

Code Guard

Built-in Go engine | 12+ languages

生成およびインストールされたコードの静的解析
外部依存関係は不要

Detects:

ハードコードされた認証情報（AWS、APIキー、SSHキー）
危険な実行（eval、os、system、shell=True）
SQLインジェクション（文字列フォーマットされたクエリ）
脆弱な暗号化（MD5、SHA1）、パストラバーサル、安全でないデシリアライズ

AI BOM

AI Bill of Materials

サプライチェーンコンプライアンスのための、
全エージェント依存関係の包括的なインベントリ

Generates:

完全なSkill + MCP + Pluginインベントリ
依存関係ツリー分析
コンプライアンスのためのSPDXエクスポート
サプライチェーン透明性レポート

ClawShield - ランタイム検査
リアルタイムツール呼び出し検査（6つのカテゴリ）

Secrets(ツール引数内のAPIキー、トークン、パスワード)、**Commands**(curl、wget、nc、rm)、**Sensitive paths**(/etc/passwd、sshキー、認証情報)、**C2**(C2ホスト名、メタデータSSRF)、**Cognitive-file**(エージェントメモリ、指示改ざん)、**Trust exploitation**(ツール引数内のプロンプトインジェクション)

Runtime（実行時保護）— 動作中のエージェントを統制・隔離する


CISCO