エージェントAIの脅威モデリングの調査

Survey of Threat Modeling in Agentic AI

宮内 由起　システム分科会　情報セキュリティ大学院大学

Abstract:
ISO/IEC defines an "AI agent" as "automated entity that senses and responds to its environment and takes actions to achieve its goals". The capacity to "perceive the environment" and operate "autonomously" toward goal achievement has given rise to novel threats that fall outside the assumptions of traditional information security measures. This study examines what distinguishes agentic AI security from conventional information security approaches, and surveys comprehensive threat modeling frameworks designed to address these emerging threats.

1. 研究背景と目的

- ◆ 2024年10月 Anthropicが初めてのエージェントAIサービスとなる「Computer Use」をリリース
- ◆ 「**MCP**(Model Context Protocol)」、「**A2A**(Agent-to-Agent)**Protocol**」の登場
- ◆ AIの民主化により、ローコード・ノーコードで容易にエージェントを自作できる環境が整備

- ● プロトコルの登場により**標準化が促進**され**サービス開発が加速**。
- ● 現在では**数千のMCPサーバ**が公開されている
- ● エージェントAIのサービスは驚異的なスピードで社会に提供され始めている。

ISO/IECは、「AIエージェント」＝「**環境を感知し、それに応答し、目標を達成するために 行動を起こす自動化されたエンティティ**」と定義。
NISTはこのエージェントの特性に対して、**現実世界のシステムや環境に影響を与える自律的な行動を取る**ことができ、
ハイジャック、バックドア攻撃、その他の脆弱性攻撃の影響を受ける可能性がある ことに警鐘を鳴らしている。

環境を認識し、目標達成に向けて**自律的**に動作する特性により、**これまでは想定されていなかった新たな脅威**が顕在化している。

研究目的

企業の様々な業務でエージェントAIが広がっていく状況を踏まえ、情報セキュリティ対策の検討に有効で、
これまで想定されてこなかった脅威にも包括的に対応可能な 脅威モデリングフレームワークを調査する。

2. 脅威モデリングフレームワークの比較 と、実際の脅威分析

従来のITシステムの脅威モデルフレームワーク **STRIDE**と、CSAが提案するエージェントAIの**MAESTRO**、
OWASPの**Top10 For Agentic Applications 2026**の比較と、実際のインシデント事例を当てはめ、脅威の網羅性を確認する。

MCP fURI脆弱性　Microsoftが提供するファイルをMarkdown形式に変換するMCPサーバにおいて、外部から指定されたURLを適切にチェックせずに読み込むSSRF(Server-Side Request Forgery)が発生。攻撃者はAIエージェントを操って、本来アクセスできない クラウドサーバー内部の機密情報を盗み出し、クラウド環境全体を乗っ取ることが可能になる。

STRIDE

Spoofing / **なりすまし**
Tampering / **改ざん**
Repudiation / **否認**(防止)
Information Disclosure / **情報漏洩**
Denial of Service / **サービス拒否**
Elevation of Privilege / **権限昇格**

- 情報漏洩：
  SSRFにより機密ファイルやクラウドの認証トークンなどが漏洩。
- 権限昇格：
  盗んだ認証情報でクラウドの管理者権限を取得。

CSA / MAESTRO

7	エージェントエコシステム	マーケットプレイス、複数のエージェントがやり取りする環境
6	セキュリティとコンプライアンス	システム全体を保護するセキュリティ制御とコンプライアンス (垂直に 他のレイヤにまたがる)
5	評価と観測可能性	エージェントの行動を監視、評価、デバッグするために使用されるシステム
4	デプロイメントとインフラストラクチャ	サーバー、ネットワーク、コンテナなど、エージェントとAPIをホストする基盤インフラストラクチャ
3	エージェントフレームワーク	MCP、A2Aなどエージェントの作成とやり取りを可能にするソフトウェアフレームワークとAPI
2	データオペレーション	ストレージ、処理、ベクトル埋め込みを含む、エージェントが使用するデータ
1	基盤モデル	エージェントが使用するAI基盤モデル

- **レイヤー3**：直接の欠陥。MCPプロトコルやサーバー自体の「入力検証」が不十分だった。
- **レイヤー4**：被害の舞台。攻撃の最終目標は、サーバが動いているクラウドインフラ(AWSなど)の支配。
- **レイヤー6**：最小権限の原則が守られていなかった
- **レイヤー7**：人気のあるツール(85kスター)に脆弱性があることで、エコシステム全体の信頼が棄損

OWASP Top10 For Agentic Applications

ASI01	エージェント目標ハイジャック
ASI02	ツールの誤用と悪用
ASI03	アイデンティティと特権の濫用
ASI04	エージェント型サプライチェーンの脆弱性
ASI05	予期しないコード実行 (RCE)
ASI06	メモリとコンテキストのポイズニング
ASI07	安全でないエージェント間通信
ASI08	カスケード障害
ASI09	人間-エージェント間の信頼の悪用
ASI10	不正エージェント

- **ASI01**: AIが攻撃者のプロンプト注入の指示に従って、意図しないSSRF攻撃を実行。
- **ASI02**: 正当なMarkdown変換機能が、「機密情報を読み取る道具」として悪用。
- **ASI03**: SSRFで盗んだクラウドの認証情報を悪用して権限を奪う。

3. まとめ・今後の計画

まとめ

STRIDEは、**シンプル**でどんなシステムにも適用でき、何が起きるかを明確に整理できる一方、**AI特有の「自律的に判断してツールを使う」という要素に対応できない**。
MAESTROは、**エージェントAIの要素を包括的にとらえ**、 レイヤーを越えた連鎖的な脅威を分析可能だが、包括的な分、**分類するのが複雑**。スキルが必要。
OWASP Top10は、やるべき対策に直結しやすく**具体策がわかりやすい**が、**全体像が見えにくい**。 インフラや学習データなど、リスト外のリスクを見落とす可能性がある

今後は、実際のインシデントケースについて 脅威モデリングフレームワークを活用しながら、エージェントAIの脅威分析の手法を調査する。