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

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。

研究目的

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**が顕在化している状況下、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

従来のITシステムの脅威モデルフレームワーク **STRIDE**、**OWASPのTop10 For Agentic Applications 2026**、CSAが提案するエージェントAIの**MAESTRO**の比較と、実際のインシデント事例を当てはめ、研究目的への有効性を調査する。

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

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

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

MAESTRO

CSAが提唱するエージェントAI固有の課題のために設計された脅威モデリングフレームワーク。
AIエージェントの複雑さを捉えきれていない従来の手法に対して、構造化されたレイヤーごとのアプローチを提供。エージェントアーキテクチャの各レイヤー内の脆弱性、各レイヤーがどのように相互作用するか、およびAI脅威の進化する性質を理解することにFocusしている。

3. 今後の研究計画

- ・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。
- ・インシデントケースを調査することとともに、自分の環境にエージェントを実装し、実際の動きを確認しながらインシデントが発生する構造の理解を深める。