資料４－５

海外動向の調査報告

2025年11月4日

三井物産セキュアディレクション株式会社

本報告のサマリ

- AI／AIエージェントのセキュリティに関して海外動向を調査しており、調査ドキュメントの一例を紹介します。
  - AI関連ガイドライン
    - Artificial Intelligence Risk Management Framework
    - OWASP Top 10 for LLM Applications
    - Guidelines for Secure AI System Development
  - AIエージェント関連ガイドライン
    - Agentic AI Red Teaming Guide
    - Agentic AI – Threats and Mitigations
    - Agentic AI Threat Modeling Framework: MAESTRO
    - Principles for Secure-by-Design Agentic Systems

- 直近の動向としては、既に海外ではAIエージェントのガイドラインがいくつか出始めており、具体的な対策も示されるようになってきています。

Artificial Intelligence Risk Management Framework

目的

AIに関連するリスクを効果的に管理し、信頼性（安全性・セキュリティ・プライバシー等）をAIライフサイクルに組み込むための指針を提供すること

発行機関、発行年

アメリカ合衆国商務省に属する研究機関であるNIST（米国国立標準技術研究所）が、2023年に発行

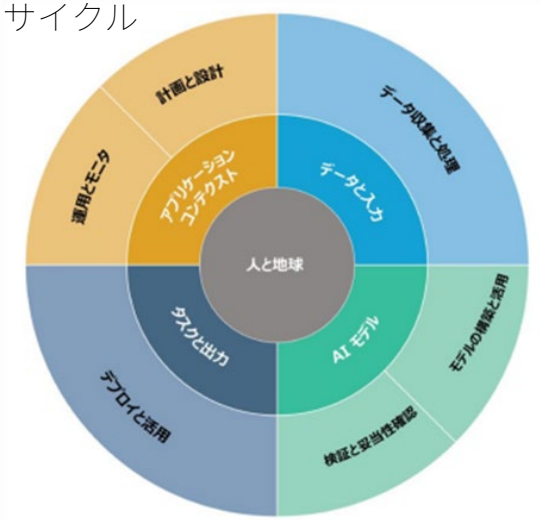
対象AI、想定読者

AIシステム全般を対象とし、AIリスク管理者・AIシステム開発者を中心とするステークホルダーを読者に想定

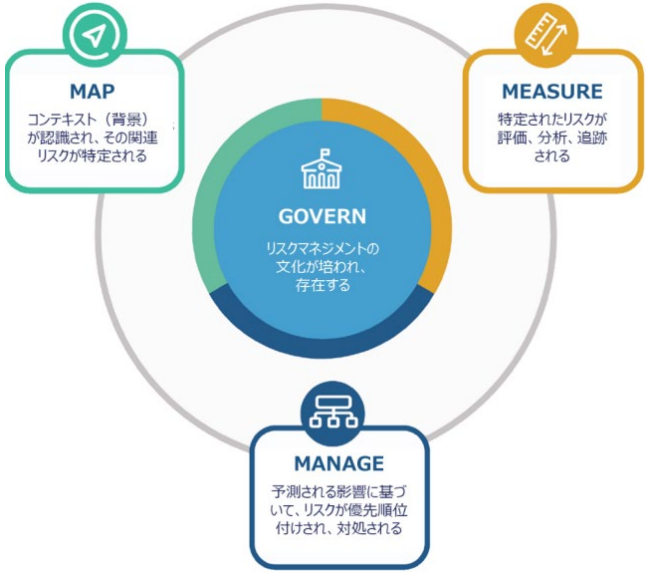
概要

- コア機能（GOVERN, MAP, MEASURE, MANAGE）により、継続的・反復的なAIリスク管理を実現
- プレイブック、ロードマップ、クロスウォーク等の補足文書で実装を支援（2024年にGenAI Profileも公開）
- AIシステムのライフサイクル毎に信頼性の観点からリスク低減策を提示

AIシステムのライフサイクル



RMFのコア機能



出典：https://aisi.go.jp/assets/pdf/NIST_RMF_Japanese_ED_updated_20240626.pdf

OWASP Top 10 for LLM Applications

目的

LLMアプリに特有の重大リスクを可視化し、設計・実装・運用全体での実践的な緩和策を提示すること

発行機関、発行年

Webアプリケーションのセキュリティ向上を目指す非営利団体であるOWASP（Open Web Application Security Project）が、2024年に発行

対象AI、想定読者

LLMアプリ全般（プラグイン等を含む）を対象とし、LLMアプリ開発者・セキュリティ担当者・運用者を中心とするステークホルダーを読者に想定

概要

- LLMアプリケーションのライフサイクル全体に跨る脅威を体系化し、脅威に対する対策例をマッピング
- 2023年の初版が発行後も、専門家コミュニティにより継続的に改訂（最新版は、2024年11月発行のver.2025）

LLMアプリケーションの脅威一覧

脅威	対策例
LLM01: Prompt Injection	期待される出力形式の定義、入出力フィルターの実装、システムプロンプトの強化、etc.
LLM02: Sensitive Information Disclosure	入力フィルターの実装、サニタイジング、アクセス管理、etc.
LLM03: Supply Chain	信頼できるデータソースやサプライヤーの利用、SBOM活用、etc.
LLM04: Data and Model Poisoning	データの来歴検証、データのバージョン管理、etc.
LLM05: Improper Output Handling	モデル出力のエンコード、Prepared statementsの利用、etc.
LLM06: Excessive Agency	拡張機能の最小権限化、ユーザー承認の要求、etc.
LLM07: System Prompt Leakage	システムプロンプトから機密データを分離、etc.
LLM08: Vector and Embedding Weaknesses	ベクトルDBのアクセス制御、データソースの検証、etc.
LLM09: Misinformation	RAGによる関連情報の取得、モデルの出力を信頼できる外部情報源と照合、etc.
LLM10: Unbounded Consumption	入力検証、レート制限、不必要な情報露出の制限、etc.

Guidelines for Secure AI System Development

目的

AIシステムを設計・開発・展開・運用の全段階において、安全かつ責任ある形で構築・維持するための実践的な指針を提供すること

発行機関、発行年

英国におけるサイバー攻撃の対応組織であるNCSC（英国国家サイバーセキュリティセンター）が他国政府機関等と共同で、2023年に発行

対象AI、想定読者

AIシステム全般を対象とし、AIシステム提供者を中心とするステークホルダーを読者に想定

概要

- 標準的なサイバー脅威に加え、AI特有の脆弱性を管理することを強調
- 各フェーズでの考慮事項と緩和策を提示し、セキュリティを"コア要件"として内在化
- 高速なAI開発に伴う後追い対策のリスクを低減

AIライフサイクルの各段階における緩和策例

設計

従業員のリスク啓発、システムに対する脅威のモデル化、e.t.c.

開発

サプライチェーンのセキュリティ確保、アセットの特定・監視・保護、文書化、技術負債の管理、e.t.c.

展開

インフラのセキュリティ確保、継続的なモデルの保護、インシデント管理手順の策定、e.t.c.

運用

システムの挙動およびインプットの監視、セキュアバイデザインに則ったアップデート、e.t.c.

Agentic AI Red Teaming Guide

目的

エージェンティックAIの特性に適合した継続的レッドチーミング手法を提供し、重大な欠陥を検出・是正すること

発行機関、発行年

クラウドセキュリティのベストプラクティスを推進する非営利団体であるCSA（Cloud Security Alliance）が、2025年に発行

対象AI、想定読者

エージェンティックAIを対象とし、エージェンティックAI開発者・セキュリティ担当者・運用者を中心とするステークホルダーを読者に想定

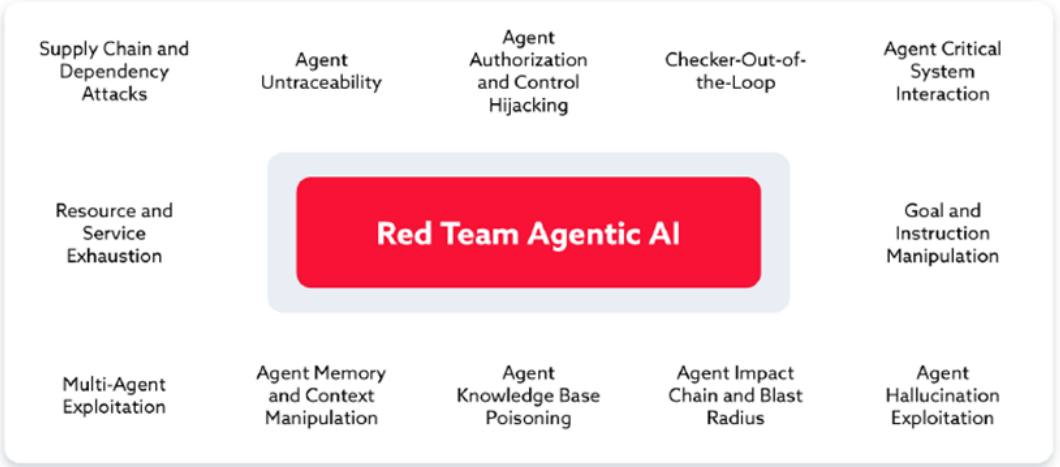
概要

- エージェンティックAIは自律的行動を伴うため、新たな脆弱性をもたらす
- そのため導入前後の継続的なレッドチーミングによって、予期せぬ結果や脆弱性を早期に特定することが不可欠
- そこで本書では主要な脅威を12個に分類し、それぞれに対する具体的なテスト手法を提示

レッドチーミングテストの手順

1. テストの準備
   シナリオの定義/環境構築など
2. テストの実行
   ログとメトリクスの取得など
3. テストの分析
   結果の評価、トリアージ
4. テストの報告
   レポート作成/低減策策定

脅威カテゴリーの一覧



出典：https://cloudsecurityalliance.org/artifacts/agentic-ai-red-teaming-guide

Agentic AI – Threats and Mitigations

目的

エージェンティックAI特有の新たな脅威の体系化と、脅威モデルに基づく実践的な緩和策の参考情報を提供すること

発行機関、発行年

Webアプリケーションのセキュリティ向上を目指す非営利団体であるOWASP（Open Web Application Security Project）が、2025年に発行

対象AI、想定読者

エージェンティックAIを対象とし、エージェンティックAI開発者・セキュリティ担当者・運用者を中心とするステークホルダーを読者に想定

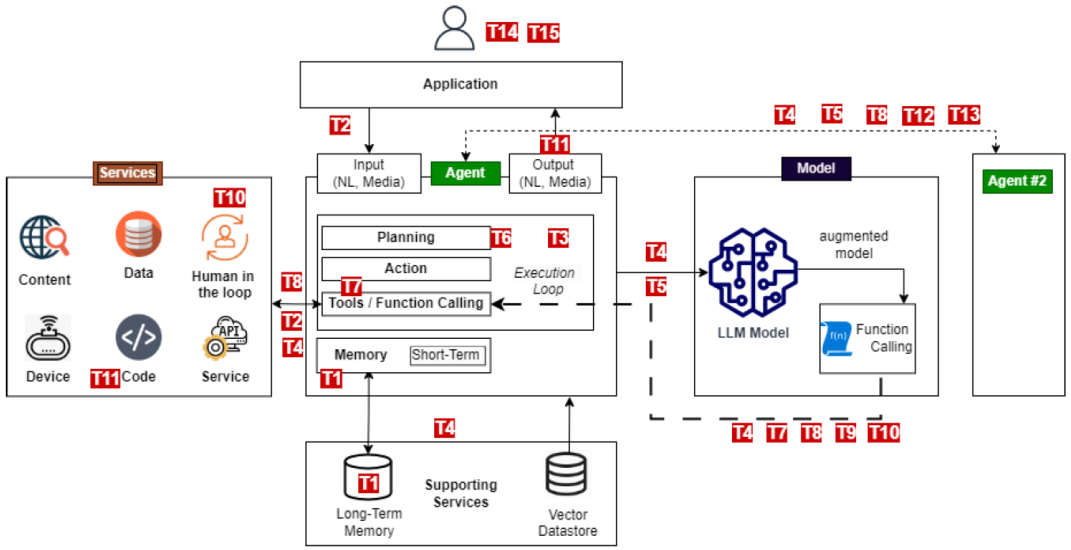
概要

- エージェンティックAIの自律性/相互作用が生む固有リスクを脅威モデル化
- 初版ガイドラインとして、用語/分類/脅威/対策の共通的な土台を提示
- OWASPの関連ガイドライン類（マルチエージェント脅威モデリング、Securing Agentic Applications）と連携

エージェンティックAIのシステム構成と脅威との対応関係

※図中のT1～T15の詳細は次頁で解説

Threat Model Summary:



出典：https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

Agentic AI – Threats and Mitigations（OWASP, 2025年発行）

エージェンティックAIの脅威一覧

脅威	概要	対策例
T1: Memory Poisoning	エージェントの短期・長期メモリに偽情報が注入されることで、エージェントの状況認識が汚染され、意思決定が操作される	・メモリ内容の検証 ・セッションの分離 ・メモリアクセスに対するロバストな認証機構 ・異常検知システム ・定期的なメモリ消去処理 ・メモリスナップショットの取得
T2: Tool Misuse	承認された権限内で連携しているツールが悪用され、意図しない操作が実行される	・ツールへの厳格なアクセス検証 ・ツールの使用パターン監視 ・エージェントへの指示内容の検証 ・明確な運用範囲の設定 ・AIツールの呼び出しを追跡する実行ログの実装
T3: Privilege Compromise	動的な権限継承等の権限管理の弱点をつかれ、権限昇格される	・詳細な権限管理 ・動的なアクセス検証 ・ロール変更の強力な監視 ・特権操作の徹底した監査 ・エージェント間でにおける特権委譲の禁止
T4: Resource Overload	AIシステムの計算資源等を意図的に使い果たされることで、サービス提供が妨害される	・リソース管理の制御 ・適応型スケーリング機構の実装 ・クォータの設定 ・システム負荷のリアルタイム監視 ・レートリミットの実装
T5: Cascading Hallucination Attacks	AIのハルシネーションを、自己強化メカニズムを通じて増幅・伝播される	・堅牢な出力検証 ・行動制約の実装 ・複数ソースによる検証 ・フィードバックループを通じたシステムの継続的な修正 ・AI生成知識の二次的な検証
T6: Intent Breaking & Goal Manipulation	エージェントの本来の目標が改変され、意図しない方向に誘導される	・計画検証のフレームワーク実装 ・境界管理 ・ゴールアラインメントのための動的保護機構 ・別モデルを用いたエージェントの行動監査 ・目標逸脱の検出
T7: Misaligned & Deceptive Behaviors	エージェントの目標達成を最優先するあまり、本来考慮すべき制約（倫理的制約等）が回避される	・有害なタスクにおける拒否能力の学習 ・ポリシー制限の適用 ・高リスク行動における人間の確認 ・ログ記録および監視 ・行動の一貫性分析 ・真実性検証 ・レッドチーミング

Agentic AI – Threats and Mitigations（OWASP, 2025年発行）

エージェンティックAIの脅威一覧

脅威	概要	対策例
T8: Repudiation & Untraceability	エージェントの非決定性や透明性不足により、AIの意志決定の追跡・監査が困難になる	・包括的なログ記録 ・暗号的検証 ・メタデータの拡張 ・リアルタイム監視
T9: Identity Spoofing & Impersonation	暗黙的な信頼関係や認証メカニズムを悪用することで、他のエージェントや人間になりすまされる	・包括的な身元検証フレームワークの構築 ・信頼境界の適用 ・継続的な監視 ・別のモデルを用いた行動プロファイリングによるAIエージェントの逸脱検出
T10: Overwhelming Human in the Loop	大量のアラートや複雑な情報により、監視者の判断能力を麻痺させる	・人間–AI相互作用フレームワークおよび適応型信頼メカニズム（人間の監督レベルと自動化の度合いを動的に調整） ・階層型の人間-AI協働（低リスクの場合は自動化し、高リスクの場合は人間の介入を優先）
T11: Unexpected RCE and Code Attacks	AIのコード生成・実行環境を悪用することで、不正実行・システム侵害が引き起こされる	・AIによるコード生成の権限制限 ・サンドボックス環境での実行 ・生成スクリプトの監視 ・特権を伴うAI生成コードに対する手動レビューを行う実行制御ポリシー
T12: Agent Communication Poisoning	マルチエージェント間の通信に対して偽情報を注入することで、システム全体の連携が混乱する	・メッセージ認証の導入 ・通信検証ポリシーの適用 ・エージェント間における相互作用の監視 ・複数エージェントの同意による検証
T13: Rogue Agents in Multi-Agent Systems	悪意ある単一のエージェントが、マルチエージェント環境に潜入することで、監視をバイパスする	・AIエージェントにおける自律性の制限（ポリシー制約、継続的な行動監視） ・制御されたホスティング環境 ・定期的なAIレッドチーミング ・入出力の逸脱監視
T14: Human Attacks on Multi-Agent Systems	エージェント間の信頼関係や依存関係を悪用することで、権限昇格されてしまう	・エージェント間における委譲の制限 ・エージェント間認証 ・エージェントの行動監視 ・複数エージェント間でのタスク分割の徹底
T15: Human Manipulation	AIに対するユーザーの信頼を悪用することで、ユーザー自身が有害行動に誘導されてしまう	・エージェントの行動監視 ・ツールアクセスの制限 ・エージェントによるリンク出力の制限 ・操作された応答の検出・フィルタリング（ガードレール、モデレーションAPI、別モデルの使用）

Agentic AI Threat Modeling Framework: MAESTRO

目的

エージェンティックAIの脅威を包括的に分析・評価するための指針を提供すること

発行機関、発行年

クラウドセキュリティのベストプラクティスを推進する非営利団体であるCSA（Cloud Security Alliance）が、2025年に発行

対象AI、想定読者

エージェンティックAIを対象とし、エージェンティックAI開発者・セキュリティ担当者・運用者を中心とするステークホルダーを読者に想定

概要

- 複雑なエージェンティックAIの構成要素を7階層に整理し、階層毎や複数階層にまたがる脅威や想定される設定不備を分析する観点を提示
- STRIDE/PASTA等の脅威モデリングのギャップを補完
- 実施手順は、①システム分解、②層別脅威特定、③クロスレイヤーの脅威特定、④リスク評価、⑤緩和策計画、⑥実装/監視

脅威を分析するレイヤー

レイヤー	概要
Layer 7: Agent Ecosystem	AIエージェントが現実世界のアプリケーションやツール、人間と相互作用するレイヤー。
Layer 6: Security and Compliance (Vertical Layer)	全てのレイヤーを横断し、AIエージェントの全ての操作にセキュリティ制御を施すレイヤー。
Layer 5: Evaluation and Observability	AIエージェントの監視と評価を行い、性能監視や異常検知を行うレイヤー。
Layer 4: Deployment and Infrastructure	AIエージェントが実行される基盤 (インフラ) を含むレイヤー。
Layer 3: Agent Frameworks	AIエージェントを構築するためのフレームワークのレイヤー。
Layer 2: Data Operations	AIエージェント向けにデータが処理・準備・保存されるレイヤー (RAGも含まれる)。
Layer 1: Foundation Models	AIエージェントの思考基盤となるLLM (またはその他のAI) のレイヤー。

Principles for Secure-by-Design Agentic Systems

目的

発展途上にあるエージェントシステムの安全利用のための基本原則を示し、リスクの抑制や説明責任性の担保につなげること

発行機関、発行年

Google等の数十社が運営・加盟するAIシステムのセキュリティ向上を推進する業界団体であるCoSAI（Coalition for Secure AI）が、2025年に発行

対象AI、想定読者

エージェントシステムを対象とし、エージェントシステム開発者・運用者・ AI リスク管理者を中心とするステークホルダーを読者に想定

※ユーザーに代わって自律的にタスクを実行するフレームワーク

概要

- 設計段階からセキュリティを備えたエージェントシステムを構築するための基本原則を提示
- ①人間による統治と責任共有 ②境界付けとレジリエンス ③透明性と検証可能性の３原則で整理

エージェントシステム構築の基本原則

1 人間による統治と共同責任

- 説明責任の明確化
- 実用的な制御と監視
- 上位原則のリスク許容度との整合

2 境界付けとレジリエンス

- 厳密で目的に特化した権限
- 堅牢な防御策
- 継続的な検証可能性

3 透明性と検証可能性

- 安全なAIサプライチェーン管理
- 包括的なテレメトリの生成
- リアルタイム監視、フォレンジック分析