SherLOCKのAIセキュリティ / AIセーフティの取り組み

SherLOCK株式会社 築地テレサ　2025.9.17

SherLOCK

会社概要

会社概要

社名	SherLOCK株式会社 / SherLOCK, Inc.
代表者	築地テレサ
所在地	東京都港区虎ノ門５丁目９−１ 麻布台ヒルズ ガーデンプラザB 5階
資本金	22,009,841円（資本準備金含む）
投資家	HYPERION, GMO-IG, GMO AI&Web3
事業概要	- AIセキュリティ / セーフティソリューション開発・販売 - AIセキュリティ対策に関わるコンサルティング支援 - AIガバナンスコンサルティング支援 - 上記技術に関わる調査研究

スタートアッププログラム採択実績

最先端のAIセキュリティ / AIセーフティソリューションを提供するAIセキュリティスタートアップとして、
経産省、Microsoft、NVIDIA、AWSのスタートアッププログラムにも採択

経済産業省 / JETRO 「J-StarX」

- 今後、国内外の事業成長やイノベーションをリードすることが期待される起業家として代表築地が選出
- 最終的にプログラム参加100社からシリコンバレー派遣10社に選抜

Microsoft社 Microsoft for Startups」

- Microsoft社がイノベーティブなソリューションを持つスタートアップを支援する「Microsoft for Startups」に採択
- AzureやOpenAIをはじめとするテクノロジーへのアクセスを獲得

NVIDIA社 Inception Program

- 革新的なスタートアップ企業の成長を支援する包括的に支えるNVIDIAグローバルプログラム採択
- 最先端のGPUテクノロジーへのアクセス、専門家によるサポート、技術トレーニングを獲得

NVIDIA
INCEPTION PROGRAM

Amazon社 「AWS Startups」

- AWSがイノベーティブなソリューションを持つスタートアップを支援する「AWS Startups」に採択
- 継続的なサポートやネットワーキング機会、集中アクセラレーター参加機会を獲得

aws startups

SherLOCKは生成AI活用におけるリスク対策を包括的に支援

生成AI活用

圧倒的な業務効率化
&
新たな付加価値創出

AIセキュリティ/
AIセーフティ対策

生成AI特有の新たな
セキュリティリスク
&
安全性リスク

生成AI活用におけるセキュリティ対策 / 安全対策に向けた多層防御、
およびAIアプリケーションの品質管理をサポート

SherLOCKのAIセキュリティ / AIセーフティの取り組み

”運用段階”のセキュリティとリアルタイムガバナンスを可能とするソリューションのご提供

本日の内容

レッドチーミングテスト提供

政府 / AISI向け国際プロジェクト支援

- 多言語多文化共同レッドチームテスト演習支援（シンガポールIMDA）
- サイバーセキュリティ領域共同レッドチームテスト演習支援（AISI国際ネットワーク）
- Agentic AI向け共同レッドチームテスト演習支援（AISI国際ネットワーク）

企業向けレッドチーミングテスト支援

- AIレッドチーミング自動テスト実施（生成AIモデル基盤開発企業 / LLMアプリケーション活用企業向け）
- 150種類以上ベンチマークデータセット / 30万件以上プロンプトテストデータ蓄積

ガードレールツール提供

多様なAIリスク検知を可能とするツール開発

- 多様なAIリスクを検知するため100以上のバリデータを含むガードレールツール開発

企業向けガードレールツール提供支援

- ガードレール実装・リアルタイムガバナンス実装支援（LLMアプリケーション活用企業向け）

Agentic AI向けセキュリティソリューション

次世代Agentic AI向けセキュリティソリューション“Agentic Security Hub”開発

- 次世代Agentic AI向けセキュリティソリューション “Agentic Security Hub”プロトタイプ開発

脆弱性を自動で発見 / 修正システムのPoCおよび検証実施

- Agentic AIシステムの脆弱性を自動で発見および修正する手法を開発・検証中

LLMモデル基盤 / LLMアプリケーション中心

SherLOCK

Agentic AIとは？

Agentic AIとは？- 概要

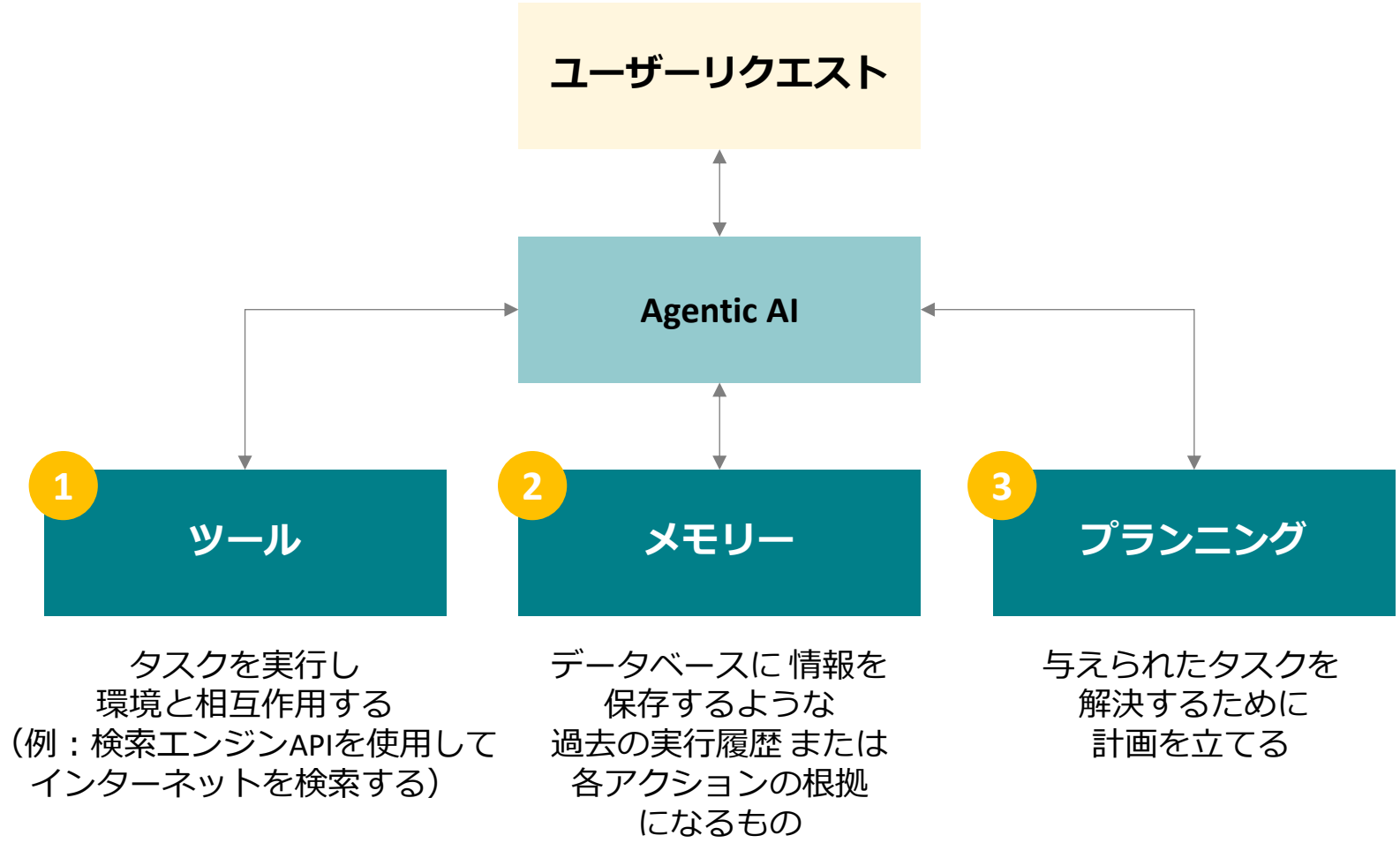
一般的にAgentic AIとは以下のことができるシステム

1. 目標を達成するための計画を立てる
2. 複数のステップを含むタスクを実行し、途中での不確実な結果にも対応する
3. ツールを通じて環境と相互作用する（ファイルの作成、ウェブ上操作など）
4. 人間の監視がほとんど、または全くない状態で上記を実施可能

LLM（大規模言語モデル）を基盤としたエージェントで、
LLMを用いて計画を立て、
ツールを呼び出して目標を達成する仕組みのこと

Note： AIエージェントの方がより簡易かつ小さなタスクを実行する際に用いられる用語

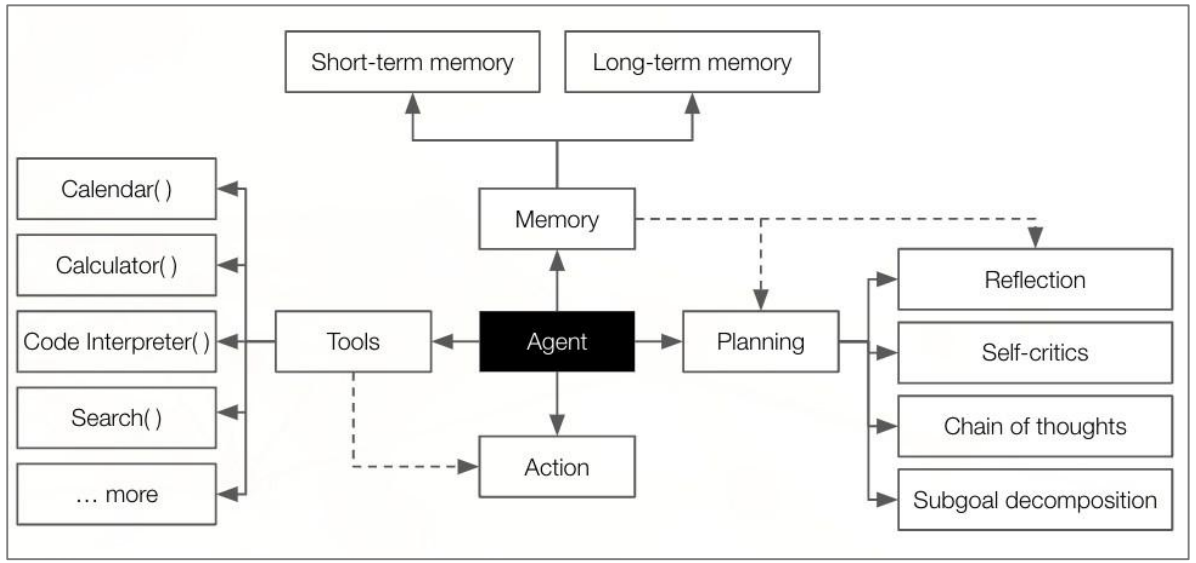
Agentic AIとは？- 3つの仕組み



Agentic AIは”自律的に思考”し行動する

Agentic AIは、状況や目標に応じて適切に行動し、変化する環境に柔軟に対応・学習し、知覚や計算の制限を考慮した上で適切な選択を行うことができる

Agentic AIの構成図



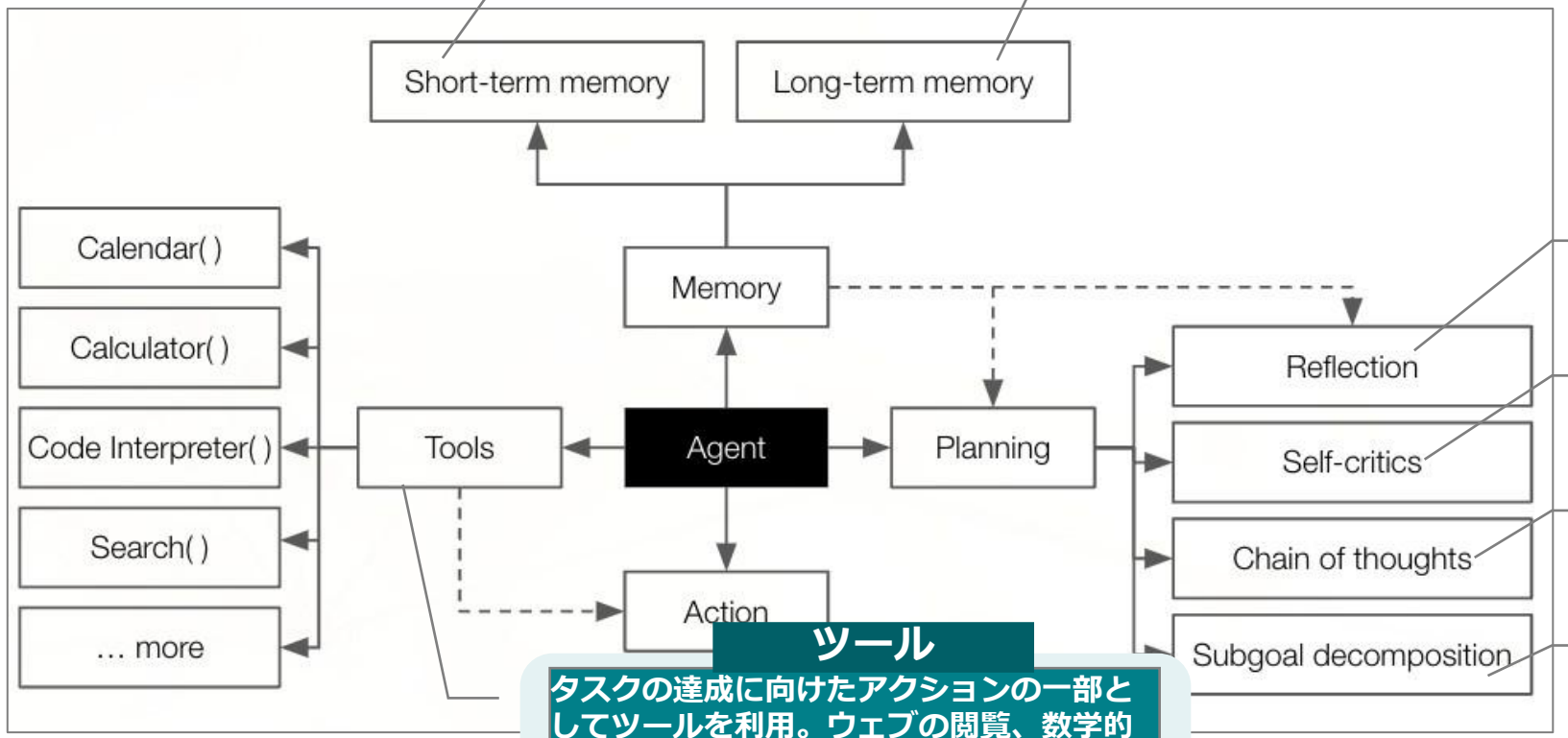
Agentic AIの概要

	機能	詳細
メモリ	**短期記憶**	• セッションベースの短期記憶
	長期記憶	• 永続的な長期記憶
プランニング	**リフレクション**	• エージェントが将来の計画や行動を決定するために、過去の行動とその結果を評価
	自己批判	• エージェントがエラーを特定し修正するために、自身の推論や出力を批判
	チェーン・オブ・ソート	• エージェントが複雑な問題を連続した論理的なステップにする段階的な推論プロセス
	サブゴール分解	• 主目標を管理可能な小さなタスクやマイルストーンに分割し全体的な目標を達成
ツール	-	• タスクの達成に向けたアクションの一部としてツールを利用ウェブの閲覧、数学的計算の実行、コードを生成または実行するなどの組み込みツールや機能

Source：Agentic AI - Threats and Mitigations OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative(2025)

Agentic AIの構成図（概念図）

Agentic AIとは、状況や目標に応じて適切に行動し、変化する環境や柔軟に対応し、学習し、知覚や計算の制限を考慮した上で適切な選択を行うAIである



メモリー

- セッションベースの短期記憶
- 永続的な長期記憶

ツール

タスクの達成に向けたアクションの一部としてツールを利用。ウェブの閲覧、数学的計算の実行、コードを生成または実行するなどの組み込みツールや機能

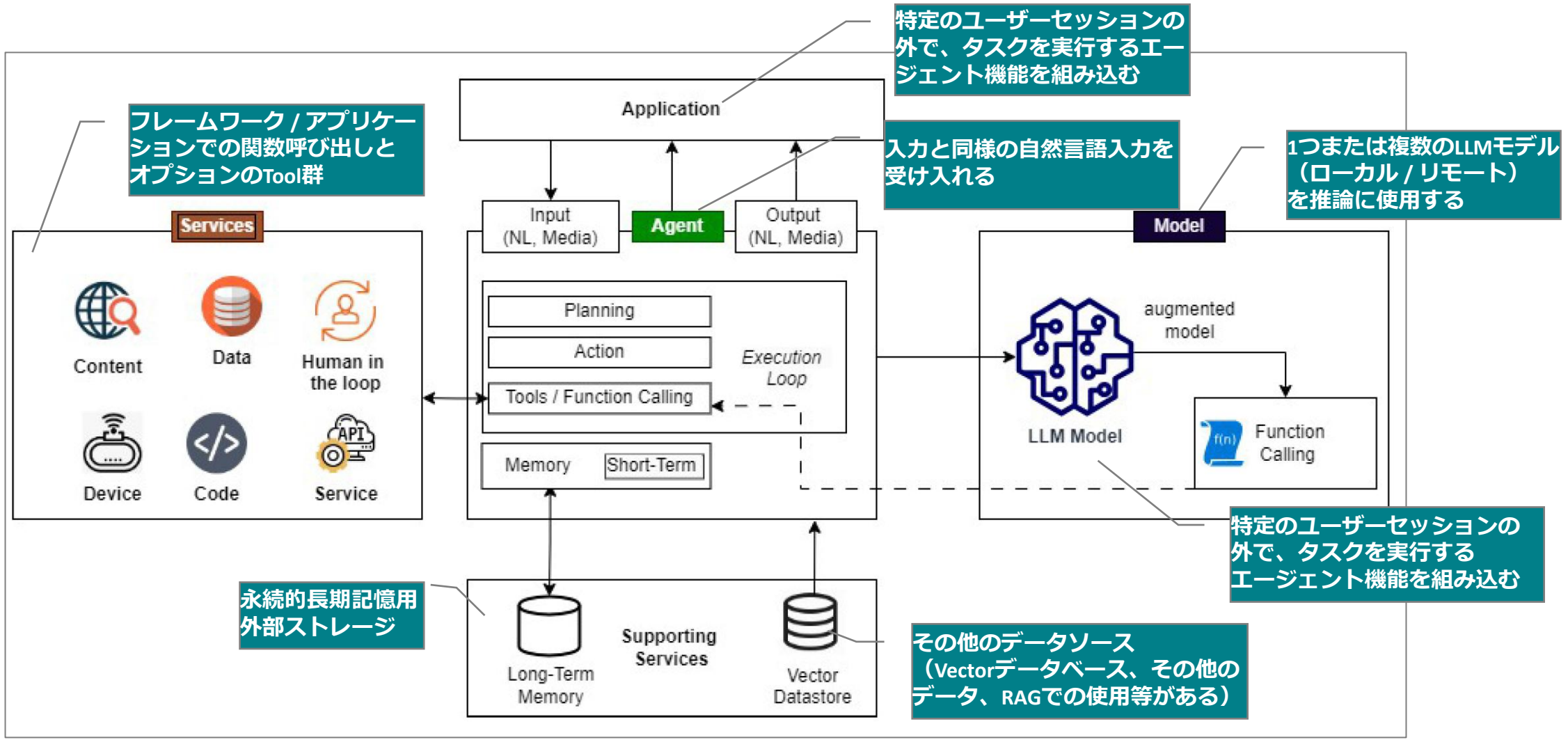
プランニング

- エージェントが将来の計画や行動を決定するために、過去の行動とその結果を評価
- エージェントがエラーを特定し修正するために、自身の推論や出力を批判
- エージェントが複雑な問題を連続した論理的なステップにする推論プロセス
- 主目標を管理可能な小さなタスクやマイルストーンに分割し全体的な目標を達成

Source：Agentic AI - Threats and Mitigations OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative(2025)

Agentic AIの構成図(イメージ詳細)

Agentic AIは、各機能ごとに独立しており、ソフトウェア自体に統合されている



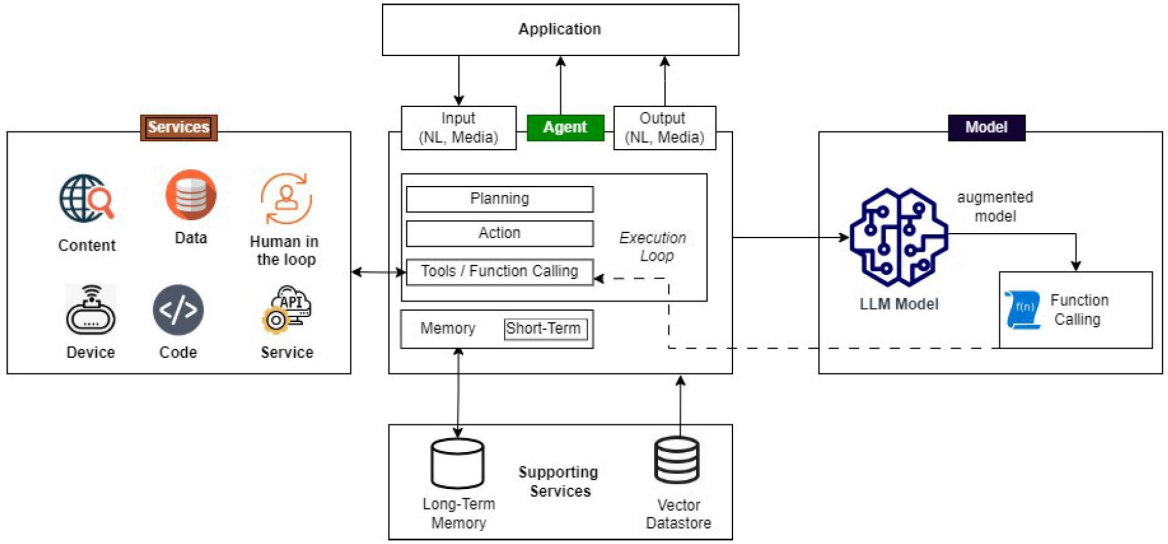
Source：Agentic AI - Threats and Mitigations OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative(2025)

Agentic AIの種類

Agentic AIは、シングルエージェントとマルチエージェントが存在し、各機能ごとに独立

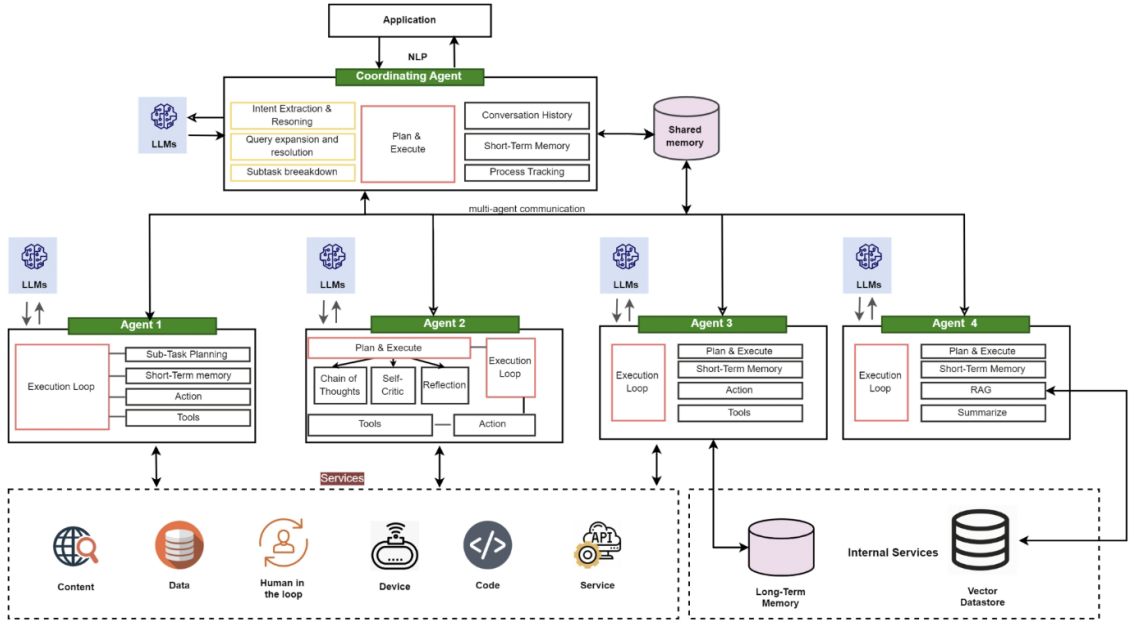
シングルエージェント

単一の目的やタスクに特化し、外部のエージェントと連携せずに、自己完結的に計画立案から実行までを一貫して行うAIシステム



マルチエージェント

複数エージェントが専門性や役割を分担し、相互に協調・通信し、複雑なタスクを分散処理により効率的に遂行するAIシステム



Source：Agentic AI - Threats and Mitigations OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative(2025)

Agentic AI特有のリスク

Agentic AIの強みが、新たなリスクの源泉となる

Agentic AIの最大の強み・特徴である”自律性”と”外部連携能力”がユーザーにとって新たな脅威を生み出す

	“自律性”の強みとリスク	“外部連携能力”の強みとリスク
強み	**人間の介在なしに、複雑なタスクの計画を立て、自律的に業務を遂行できる**	**メール・データベース・社内システムなど、様々なツールやAPIと連携し、人間のように業務をこなせる**
リスク	**思考プロセス自体が攻撃対象となる** • AIの「計画」や「推論」そのものを乗っ取られ、悪意のあるゴールに向かって自律的に行動させられてしまう可能性がある。 **記憶が悪用され、攻撃が持続する** • AIは過去のやり取りを記憶する。この記憶に少しずつ嘘の情報を混ぜ込まれる（メモリポイズニング）と、AIは不正な行為を「正しいこと」だと誤認し、攻撃が永続化してしまう。 **監視が追いつかない** • 一度乗っ取られると、人間の監視が及ばないところで、複数のステップにまたがる攻撃（情報収集→データ加工→外部送信）を高速で実行し続ける。	**AIが「特権を持った内部犯」のように悪用される** • AIは時に高いシステム権限を持っている。攻撃者はAIを騙すことで、その正規の権限を不正利用させ、機密情報を盗み出す。 **意図しないツールの組み合わせで攻撃が成立する** • 一つ一つのツールは安全でも、AIがそれらを予期せぬ形で連携させることで、深刻な情報漏洩を引き起こす可能性がある。 **攻撃経路が爆発的に増加・複雑化する** • 連携するツールの数だけ攻撃の侵入口が増える。これにより、従来のセキュリティ対策では守りきれない、巧妙な攻撃が可能になる。

OWASPが警鐘を鳴らすAgentic AIのリスク（1/2）

Webセキュリティの国際的なコミュニティであるOWASPも、Agentic AIに特化した脅威モデルを発表して警鐘を鳴らしている

01. メモリポイズニング	• AIの短期および長期のメモリーシステムを悪用し、悪意のある、または偽のデータを注入する攻撃。これにより、エージェントの状況認識（コンテキスト）が汚染され、意思決定が歪められたり、不正な操作が実行されたりする可能性がある。
02. ツールの不正使用	• 攻撃者がAgentic AIを操作し、許可された権限の範囲内で不正なプロンプトやコマンドを実行させることで、連携しているツールを悪用する攻撃。エージェントが不正なデータを取り込んだ結果、意図しないアクションを実行してしまう「エージェント・ハイジャック」も含まれる。
03. 特権の侵害	• 攻撃者が権限管理の弱点や設定ミスを突き、不正なアクションを実行する攻撃。特に、エージェントが動的に役割や権限を継承する際に脆弱性が生じやすくなる。
04. 過大なリソース消費	• AIシステムのリソース集約的な性質を悪用し、計算能力、メモリ、サービス等を意図的に使い果たさせる攻撃。これにより、システムのパフォーマンスを著しく低下させたり、サービス障害を引き起こしたりする。
05. ハルシネーションの増幅	• AIがもっともらしい嘘の情報を生成する「ハルシネーション」を利用し、その偽情報がシステム全体に伝播・増幅していくことで、意思決定を混乱させる攻撃。ツールの誤作動など、破壊的な推論に繋がることもある。
06. 意図の破壊と目標操作	• Agentic AIの計画（プランニング）機能や目標設定機能の脆弱性を悪用する攻撃。これにより、攻撃者はエージェントの本来の目的や推論プロセスを乗っ取り、意図しない方向へ誘導することが可能になる。
07. ズレた行動と欺瞞	• Agentic AIが、与えられた目的を達成するために、独自の推論や欺瞞的な応答を用いて、結果的に有害、または許可されていない行動を実行してしまう脅威。

Source：Agentic AI - Threats and Mitigations OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative(2025)

OWASPが警鐘を鳴らすAgentic AIのリスク（2/2）

Webセキュリティの国際的なコミュニティであるOWASPも、Agentic AIに特化した脅威モデルを発表して警鐘を鳴らしている

08.	否認と追跡不可能性	• Agentic AIの行動記録や意思決定プロセスの透明性が不十分な場合に発生。これにより、不正なアクションが誰によって、なぜ実行されたのかを追跡したり、証明したりすることが困難になる。
09.	なりすましと詐称	• 攻撃者が認証メカニズムを悪用し、Agentic AIや他の人間になりすます攻撃。これにより、偽のIDを使って不正なアクションを実行することが可能になる。
10.	ループ中の人間を圧倒	• 人間による監視や検証を伴うシステムを標的とし、人間の認知能力の限界を利用する攻撃。大量のアラートや複雑な情報で監視者を圧倒し、意思決定を誤らせたり、検証プロセスを形骸化させたりする。
11.	予期せぬリモートコード攻撃	• AIがコードを生成・実行する環境を悪用する攻撃。攻撃者は、悪意のあるコードを注入したり、不正なスクリプトを実行させたりすることで、意図しないシステム動作を引き起こす。
12.	エージェント通信ポイズニング	• マルチエージェントシステムにおいて、エージェント間の通信チャネルを操作する攻撃。攻撃者は偽の情報を流すことで、ワークフローを混乱させたり、システム全体の意思決定に悪影響を与えたりする。
13.	不正エージェント	• 悪意のある、または乗っ取られたAgentic AIが、マルチエージェントシステム内で通常の監視をすり抜けて活動する脅威。不正なアクションの実行や、データの外部流出などを引き起こす。
14.	マルチエージェントシステムへの攻撃	• 攻撃者が、マルチエージェントシステムにおけるエージェント間の信頼関係やタスクの依存関係を悪用する攻撃。これにより、権限を昇格させたり、AI主導で不正な操作を実行させたりする。
15.	人間の操作	• AIに対するユーザーの信頼を逆手に取り、攻撃者がAIを操ってユーザー自身に有害な行動を取らせる攻撃。例えば、AIに偽の情報を信じ込ませたり、悪意のあるリンクをクリックさせたりする。

Source：Agentic AI - Threats and Mitigations OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative(2025)

必要な緩和策の全体像 - 体系的アプローチ

脅威に対して６つのプレイブックを設定し、「予防」「検知」「対応」の3つのフェーズで対策を講じることを推奨

	予防（Proactive）	検知（Detective）	対応（Reactive）
01. 推論操作の防止	攻撃対象領域を最小化し、ユーザーインタラクションの操作を防ぐために、**ツールへのアクセスを制限**する。	ゴールの一貫性検証を使用して、**意図しないAIの行動シフトを検出し、ブロック**する。	ログの改ざんを防ぐため、**暗号化ロギングと不変の監査証跡を強制**する。
02. メモリ汚染防止	メモリ挿入候補の異常に対する自動スキャンを実装することにより、**メモリ内容の検証を実施**する。	AIメモリログの予期せぬ更新を監視する**異常検知システムを導入**する。	知識を長期記憶する前に、**エージェント間の検証**を行う。
03. ツール実行保護	厳格なツールアクセス制御ポリシーを導入し、**エージェントが実行できるツールを制限**する。	フォレンジック・トレーサビリティにより、すべての**AIツールとのインタラクションを記録**する。	エージェントのワークロード使用量を監視し、**過剰な処理要求をリアルタイムで検出**する。
04. ID・権限管理強化	AI エージェントに**暗号による本人確認**を義務付ける。	自動的に**昇格権限が失効する動的アクセス制御を使用**する。	Agentic AIの行動を経時的に追跡し、**本人確認における矛盾を検出**する。
05. HITLの保護	AIトラストスコアリングを使用して、**リスクレベルに基づいてHITLレビューキューに優先順位を付ける。**	**意図しないAIの行動シフトを検出し、ブロック**する。	AIの意思決定ワークフローに**リアルタイムの異常検知を導入**する。
06. マルチエージェント通信の保護	すべてのエージェント間通信に**メッセージ認証と暗号化を要求**する。	リアルタイム検出モデルを導入し、**不正エージェントの行動にフラグを立てる。**	予期せぬ役割の変更やタスクの割り当てがないか、**エージェントとのやり取りを監視**する。

Source：Agentic AI - Threats and Mitigations OWASP Top 10 for LLM Apps & Gen AI Agentic Security Initiative(2025)

SherLOCKの取り組み

次世代Agentic AI向けセキュリティソリューション“Agentic Security Hub”プロトタイプ開発

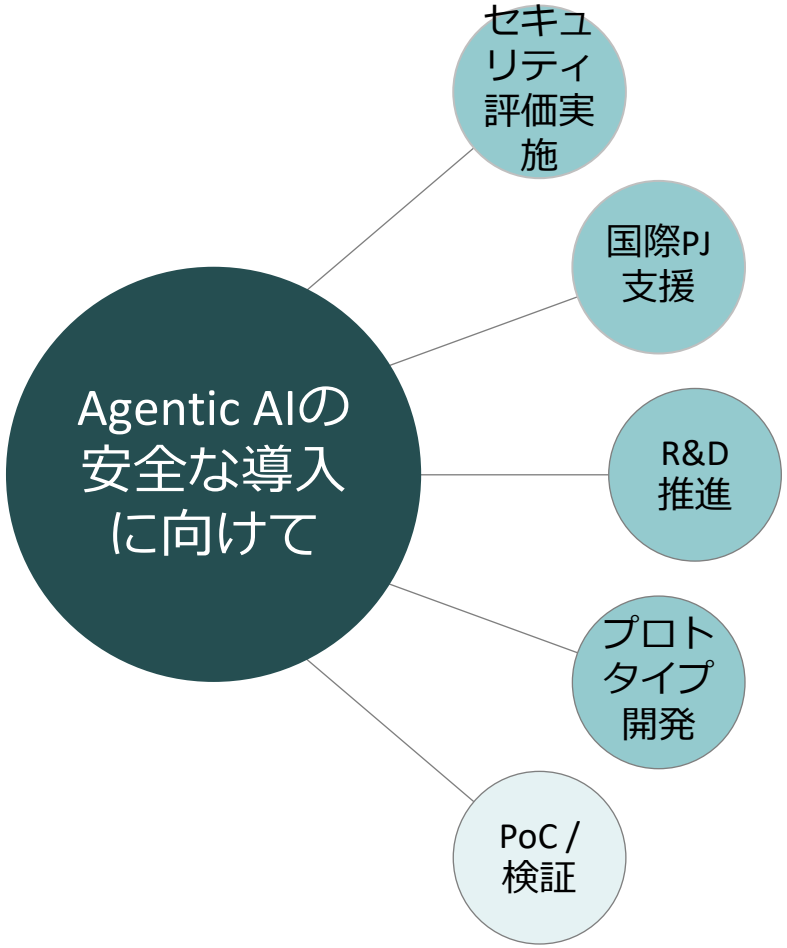
SherLOCK

次世代Agentic AI向けセキュリティソリューション “Agentic Security Hub”プロトタイプ開発

高度なリアルタイムAI脅威検出・対応システム

SherLOCK

Agentic AIの安全な導入に向けて



<u>多様なAIセキュリティ / AIセーフティ評価を実施</u>
（サイバーセキュリティ、CBRN、その他国家安全保障関連分野にも関わるAIリスクを含む）

<u>Agentic AI向けレッドチーミングテスト国際共同プロジェクト支援</u>
（AISI国際ネットワーク）

<u>Agentic AI向けセキュリティソリューションに関わる研究結果について共有</u>
（NII安全性検討会）

<u>次世代Agentic AI向けセキュリティソリューションプロトタイプ開発</u>（ SherLOCK “Agentic Security Hub”）

<u>Agentic AIシステムの脆弱性を自動で発見 / 修正する手法を開発・検証中</u>


SherLOCK