アンラボ Cyber Threat Intelligence Report

TLP : AMBER+STRICT

AIをめぐる脅威と攻撃動向

AIを利用した攻撃とAIを標的とした攻撃

V1.0 | 日本語版

アンラボ セキュリティインテリジェンスセンター（ASEC）

2026年11月13日

AhnLab

文書区分に関するご案内

刊行物および提供コンテンツは、以下のとおり、各文書区分で認められた範囲に限り利用できます。

区分	配布対象	注意事項
TLP : RED	特定の顧客または顧客企業の報告書受領者	受領者または受領部門のみ閲覧できます。受領者以外への複製・配布は禁止です。
TLP: AMBER+STRICT	特定の顧客または顧客企業の内部組織	受領組織内で複製・配布できます。組織外で研修などに利用する場合はアンラボの許可が必要です。
TLP : AMBER	特定の顧客または顧客企業、およびその顧客	受領組織内とその顧客に対して複製・配布できます。対象者以外への研修などに利用する場合はアンラボの許可が必要です。
TLP : GREEN	サイバーセキュリティコミュニティ（業務協定を締結した外部機関を含む）	該当業界などで利用できます。出典を明記すれば、社内研修や同業者・セキュリティ担当者向けの研修資料に利用できます。一般向けの発表資料への利用は厳しく制限されます。
TLP : CLEAN	自由に利用可能	出典を明記すれば、営利・非営利のいずれの目的でも利用できます。改変・翻案などの二次的著作物の作成も認められます。

改訂履歴

バージョン｜日付｜内容

1.0｜2026年11月13日｜初版

目次

1. 要約 7
2. 背景 8
3. AI技術とセキュリティの基礎 9
1) AI、ML、LLMの概要 9
2) AIエージェントと連携技術（MCP、Skills、A2A、ハーネス）の概要 10
3) AIサービスの構成（モデル、データ、制御、アクセス、エージェント） 12
4) AIの悪用とアンダーグラウンドAIの登場 14
(1) アンダーグラウンドAIツールの登場 14
(2) 商用AIの安全対策の回避 15
5) AI技術の発展と脅威の変化 16
(1) 対話型チャットボットの登場（2022年） 16
(2) ツール呼び出しと外部連携の始まり（2023年） 17
(3) AIコーディングツールの普及（2024年） 17
(4) 標準プロトコルの登場とエージェントエコシステムの拡大（2024～2025年） 17
(5) AIエージェントの普及とAIによる攻撃の現実化（2026年） 18
4. AIをめぐる主な脅威事例 19
5. AIへの信頼を悪用する攻撃 20
1) AIサービスやプラットフォームへのなりすまし 20
2) AI拡張機能やプラグインを装う攻撃 22
3) 偽のAI APIサービス 23
4) 正規AIサービスの機能の悪用 24
5) AI関連資料を装う攻撃 24
6. AIを活用した攻撃（AI-Augmented Attacks） 25
1) ディープフェイクを用いたなりすましと金融詐欺 25
2) 詐欺コンテンツの作成 26
3) マルウェアの作成と配布 27
(1) マルウェアの作成と開発支援 27
(2) マルウェアの配布支援 29

4) 攻撃の実行と運用の支援 29
5) AIと連携するマルウェア 31
7. AIサービスおよびプラットフォームへの攻撃 33
1) AIサービスおよびプラットフォームへの攻撃 33
(1) AIブラウザー 33
(2) AIアシスタント 34
(3) MCP（Model Context Protocol） 34
(4) 自律型AIエージェント 35
(5) AI開発・運用プラットフォーム 36
(6) ローカルLLMへの攻撃 38
2) AI・LLMを標的とする攻撃手法 39
(1) RAG・検索の改ざんとサービス妨害 39
(2) プロンプトインジェクション 40
(3) AI分析ツールの回避 42
3) AI連携コンポーネントへのサプライチェーン攻撃 43
(1) 悪意あるMCPサーバーとエージェントサーバーのサプライチェーンの悪用 43
(2) 悪意あるAIスキルの配布 44
4) AI開発・ビルド環境へのサプライチェーン攻撃 46
(1) AI開発ツール・CLIへのサプライチェーン攻撃 46
(2) AIフレームワーク・ライブラリへの悪意あるパッケージの混入 47
(3) バックドアを含むAIモデル 48
5) AI関連情報の窃取 49
(1) 会話とプロンプトの収集 49
(2) APIキーと認証情報の窃取 50
8. 対策 51
9. 今後の見通し 53
10. 結論 54
11. 参考文献 54

注意事項

本報告書には、現時点で確認できた情報に基づく分析担当者の見解が多く含まれています。見解は担当者によって異なる場合があり、新たな根拠が確認された場合は、事前の予告なく内容を改訂することがあります。

目次と本文のページ番号は韓国語原文に対応しています。本文の上付き数字は原文の出典注を示します。攻撃事例は経緯、影響、防御上の要点を中心に記述し、実行手順やコードの一部を省略しています。一部の出典注は資料名で記載しています。

1. 要約

生成AIの普及とAIエージェントの登場により、サイバー脅威を取り巻く環境は急速に変化している。従来、AIは主にコンテンツ制作や業務支援に活用されてきた。近年では、メール、文書、コードリポジトリ、クラウドサービス、セキュリティツールなどと連携し、実際の業務を遂行するエージェントへと発展している。この過程で、AIは攻撃者の新たな手段となると同時に、攻撃の標的にもなっている。

AIに関わる脅威は大きく三つに分けられる。一つ目は利用者の信頼を悪用する攻撃である。攻撃者はChatGPT、Claude、DeepSeek、Geminiといった著名なAIサービスを装ったり、AIサービスの共有機能を悪用したりする。二つ目は、攻撃者がAIを活用する攻撃である。フィッシングコンテンツの生成、マルウェアの開発、脆弱性の分析、攻撃の自動化などにAIが使われ、攻撃が効率化している。ペネトレーションテスト用ツールとAIエージェントを組み合わせた攻撃自動化の試みも確認されている。三つ目はAI自体を標的とする攻撃である。プロンプトインジェクション、データの窃取、MCPサーバーの悪用、サプライチェーン攻撃、モデル汚染など、AIサービスやエージェントを直接狙う事例が増加している。

特にMCP（Model Context Protocol）、AIエージェント、スキル、A2A（Agent-to-Agent）などの技術は、AIの活用範囲を広げると同時に、攻撃対象領域を拡大させる。AIが外部システムと連携し、付与された権限で操作できるようになると、攻撃者はその権限や連携先を悪用し、内部システムへのアクセス、情報窃取、サプライチェーンの侵害を試みることができる。これに対応するには、従来のウェブアプリケーションやクラウド環境とは異なるセキュリティポリシーと制御が必要となる。

現時点で確認されている脅威の多くは初期段階にあるが、技術の発展速度を踏まえると、AIを利用した攻撃は今後、自動化と大規模化が進むと見込まれる。自律型エージェント、AIを介したラテラルムーブメント、AI駆動型マルウェア、AIモデルの自己複製に関する研究は、今後想定される脅威の方向性を示している。組織はAIサービスの導入にあたり、プロンプトインジェクション対策、最小権限、サプライチェーンの検証、機密情報の保護、挙動監視など、AI環境に応じた対策を検討する必要がある。本報告書の発行後も、新たな攻撃手法、脆弱性、悪用事例、防御技術が現れると予想される。AIセキュリティには継続的な調査と監視が求められ、今後も変化の速い領域の一つであり続けるだろう。

2. 背景

OpenAIが2022年11月に公開したChatGPTは、5日で利用者100万人、約2カ月で1億人に達し、AIサービスが広く普及する契機となった[1、2]。その後、大規模言語モデル（LLM）を含むさまざまなAIサービスが登場し、生産性の向上、業務の自動化、知識へのアクセス拡大といった効果をもたらした。同じ技術は攻撃者にも新たな手段を提供した。

攻撃者は、精巧な偽コンテンツの作成やソーシャルエンジニアリングの高度化にAIを利用し始めた。AIによるフィッシング文面の作成や、窃取したデータの要約も行われた。AIは標的国の話し方や慣習に合わせた自然な文面を容易に作れるため、従来の翻訳ソフトに特徴的な不自然な表現が減った。攻撃に必要な費用と時間が減り、高度な技能を持たない攻撃者にもさまざまな攻撃を行いやすくなった。マルウェアの分野でも、コードの作成・改善、侵入時の意思決定支援、指令・制御（C2）の開発など、開発と運用の各段階でAIが利用されている[3]。

AIがマルウェアの実行時の判断に組み込まれることもある。端末の構成や権限、所属組織など、感染先の状況に応じて挙動が変化するため、固定的な動作を前提とした分析では攻撃を予測しにくくなる。研究段階ではあるが、AIモデルが自ら別の機器に広がる「AIワーム」の可能性も実験で検証されている[4]。

本報告書では、AIへの信頼を悪用する攻撃、AIを活用した攻撃、AIシステムを標的とする攻撃について、主な事例とともに概観し、対策と今後の見通しを整理する。

1 https://openai.com/ko-KR/index/chatgpt/
2 https://openai.com/ko-KR/global-affairs/new-economic-analysis/
3 Check Point Research ― AIを悪用したC2の分析
4 https://www.anthropic.com/research/self-replicating-ai

3. AI技術とセキュリティの基礎

1) AI、ML、LLMの概要

人工知能（AI、Artificial Intelligence）は、人間のように推論・判断し、新たな結果を生み出す技術全般を指す。機械学習（ML、Machine Learning）はAIを実現する代表的な手法である。人が個々の規則を定義する代わりに、データからパターンを学習してモデルを構築する。深層学習（Deep Learning）は多層のニューラルネットワークを使う機械学習手法で、大量のデータと計算資源により、画像、音声、言語といった複雑なデータの特徴を抽出する。

大規模言語モデル（LLM、Large Language Model）は、深層学習技術の一つであるTransformerを基盤とし、大量の文章を学習して、自然言語を理解・生成するモデルである。入力文をトークンと呼ばれる単語や文字の断片に分け、トークン間の文脈を捉えながら、次に続くトークンを確率的に予測して回答を生成する。学習時点より新しい情報は、検索や外部ツールで補う。ChatGPT、Claude、Geminiなど、広く利用されている対話型AIサービスはLLMを基盤としている[5]。

LLMを用いたサービスは、AIアシスタントからAIエージェントへと発展している。アシスタントは利用者の質問や依頼に応じて回答、文書、コードなどを生成し、それらを使って実際に操作するのは利用者である。エージェントは目標を受け取ると計画を立て、メール、スケジュール管理、ブラウザー、決済などの外部ツールやサービスを使って作業する。2026年初めに関心を集めたオープンソースの個人向けAIエージェントOpenClawが代表例である[6]。Metaも、利用者に代わって日常的なアプリを横断して作業する個人向けAIエージェントMuseを公開した[7]。

AIの活動範囲は物理空間にも広がっている。フィジカルAI（Physical AI）は、ロボット、自動運転車、カメラなどの自律システムが、現実の環境を認識し、判断して行動するための技術である。言語や画像を中心に学習する従来の生成AIには、現実世界の物理法則を理解するうえで限界がある。

5 https://www.sentinelone.com/labs/inside-the-llm-understanding-ai-the-mechanics-of-modern-attacks/
6 https://openclaw.ai/
7 https://about.fb.com/news/2026/09/introducing-muse/

フィジカルAIは、空間的な位置関係や物理法則を考慮し、センサーデータを機械が実行できる動作に変換する[8]。

段階	主な機能	結果・影響範囲
LLMアシスタント	質問への回答を生成	文章、画像、コードなどの成果物
AIエージェント	目標達成のための計画と実行	アカウント、ファイル、サービスなどのデジタル環境
フィジカルAI	物理環境の認識と行動	ロボットや車両などが動作する実空間

表1　AIの段階別に見た機能と影響範囲

この発展に伴って攻撃対象領域も拡大する。アシスタントでは主に入力プロンプトと出力結果が狙われたが、エージェントでは連携ツール、付与された権限、アクセス可能なデータまでが標的となる。フィジカルAIでは、情報漏えいやサービス障害に加え、実空間での物理的被害につながるおそれもある。

2) AIエージェントと連携技術（MCP、Skills、A2A、ハーネス）の概要

AIエージェントは、LLMを中心に、目標設定、推論、ツールの呼び出し、結果の反映を繰り返して自律的に作業する。質問への回答に加え、ファイルシステムへのアクセス、APIの呼び出し、コードの実行などを通じて、外部環境と直接やり取りする。作業を複数のサブエージェント（Sub-Agent）に分担させ、結果を取りまとめるマルチエージェント（Multi-Agent）構成も広く使われている。

エージェントの機能を拡張するため、データソース、API、コマンドラインインターフェース（CLI）、スクリプトなどを組み合わせたツールチェーンで複雑な作業フローを構成する。ブラウザーエージェント（Browser Agent）やコンピューター操作エージェント（Computer Use Agent）は、ページの閲覧、ボタンの操作、フォーム入力を行えるため、専用APIのないサービスも扱える。その一方で、閲覧したページの内容がエージェントの判断に影響する可能性がある。

MCP（Model Context Protocol）は、AIアプリケーションを外部システムと接続するための公開標準である。MCPサーバーがツール、データ、プロンプトなどを提供し、MCP対応アプリケーションは個別の連携開発をせずに利用できる。

8　https://www.nvidia.com/en-us/glossary/generative-physical-ai/

Anthropicが2024年11月にMCPを初めて公開し、2025年12月にLinux Foundation傘下のAgentic AI Foundation（AAIF）に寄贈した。寄贈時点で公開MCPサーバーは1万を超え、ChatGPT、Cursor、Gemini、Microsoft Copilot、Visual Studio CodeなどがMCPに対応している[9]。

エージェントスキル（Agent Skills）は、特定の作業方法を伝える指示、スクリプト、リソースを一つのフォルダーにまとめる形式である。エージェントは、そのときの作業に必要なスキルだけを読み込む。MCPは接続先を、スキルは作業手順を提供する。Anthropicが2025年10月に公開した後、VS Code、GitHub、Cursorなどの開発ツールが対応した。コーディングエージェントにプロジェクト固有の作業指示を伝えるAGENTS.mdも、AAIFの管理する標準となっている[10]。

A2A（Agent2Agent）は、異なる企業やフレームワークで開発されたエージェント同士が通信・協調するための公開標準である。Googleが2025年4月に公開し、同年6月にLinux Foundationに寄贈した。各エージェントは名称、機能、認証方式などを記したエージェントカードを公開し、他のエージェントはそれを参照して作業を依頼する。MCPはエージェントとツールを、A2Aはエージェント同士を接続する[11]。

構成要素	役割	接続先
AIエージェント	目標に沿った計画と実行	利用者、ツール、他のエージェント
ツールチェーン	作業フローの構成	API、CLI、スクリプト、ブラウザー
MCP	外部ツール・データとの連携標準	エージェントとツール・データ
エージェントスキル	作業方法の提供	エージェントと指示・スクリプト
A2A	エージェント間の協調標準	エージェント同士

表2　構成要素と役割

9 https://www.anthropic.com/news/donating-the-model-context-protocol-and-establishing-of-the-agentic-ai-foundation
10 https://agentskills.io/
11 https://a2a-protocol.org/

MCP、Skills、A2Aが接続先や利用機能を規定するのに対し、ハーネス（Harness）は、それらとの連携を実行する。AIモデルを取り囲み、ツールの呼び出し、コマンドの実行、権限の確認、指示の管理を担うプログラムで、Claude Code、Codex、OpenClawなどが代表例である。モデル自体はテキストを生成するため、利用者への承認要求や実行範囲の制限といった安全対策の多くはハーネス側にある。その設定ファイルや指示ファイルは改ざんの標的となり、ハーネス自体が攻撃の自動化に悪用されることもある。

2026年現在、MCPサーバー、スキルのリポジトリ、マーケットプレイスを中心としたエコシステムが急速に形成されている。自動化の効率が高まる一方、エージェントに必要以上の権限があると、誤動作や攻撃を受けた際の被害が拡大する。外部から取得するMCPサーバー、スキル、指示ファイルに悪意あるコードや隠れた指示が含まれると、サプライチェーン攻撃の経路となる。エージェントが参照するウェブページ、文書、他のエージェントの応答が改ざんされた場合も、意図しない動作につながり得る。

3) AIサービスの構成（モデル、データ、制御、アクセス、エージェント）

AIサービスは一般に、モデル層、データ層、制御層、アクセス層、エージェント層の五つで構成される。層ごとに異なる種類の攻撃が想定される[12]。

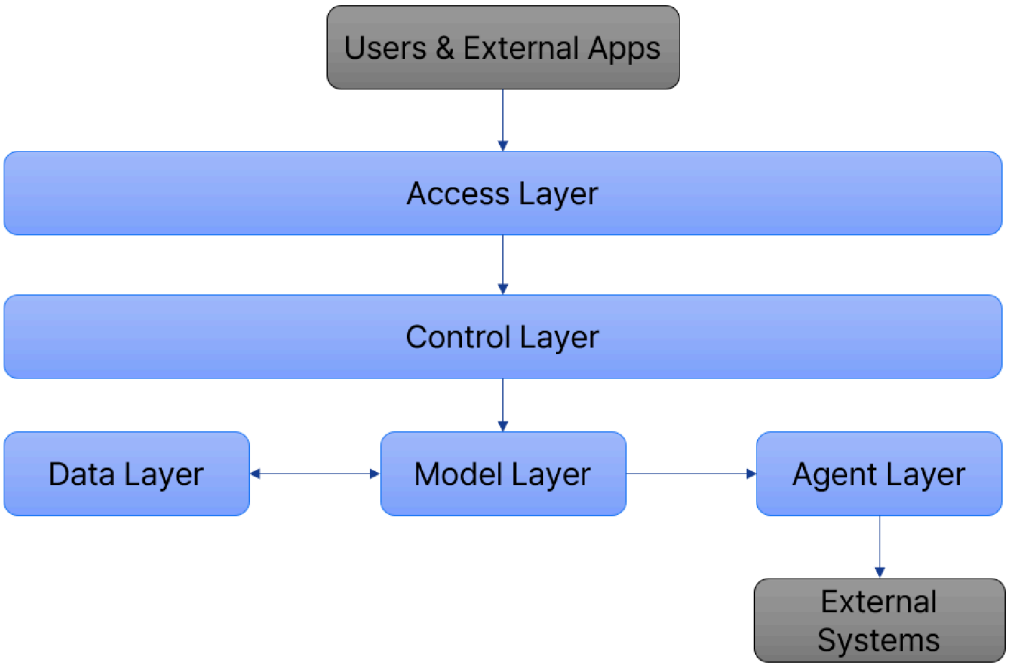


図1　AIサービスの構成

12 https://saif.google/secure-ai-framework/components

サービスの中核は、推論を行うモデルである。企業は事業者が提供するモデルをAPI経由で使うほか、Ollamaなどのツールを使って自社サーバーにモデルを導入したり、Hugging Faceなどの公開リポジトリからオープンウェイトモデルを入手したりする。自社運用の場合、サーバーのセキュリティは運用者が担う。外部から入手したモデルファイルにバックドアが仕込まれている場合もある。

モデルが扱うデータには大きく三種類ある。第一はモデル構築時の学習データである。第二はサービス運用時に参照する外部データで、検索拡張生成（RAG、Retrieval-Augmented Generation）に用いられることが多い。回答を生成する前に、文書リポジトリやデータベースから関連情報を検索し、その内容を回答に反映する。モデルは学習時点以降の情報や社内文書まで参照できる一方、保管元を改ざんされると回答も歪められる可能性がある。第三は会話履歴とメモリである。多くのAIサービスは過去の会話、利用者の好み、業務情報を保存して以降の会話に活用する。この領域には機密情報が集まるため窃取の標的となり、悪意ある内容を記録されると以降の会話に影響が及ぶ。

制御層のガードレールは、モデルに許可されていない挙動を防ぐ。サービス運営者はシステムプロンプトで役割と禁止事項を指定し、入力・出力フィルターや安全性を判定する分類器で有害な要求や応答を選別する。ここはセキュリティポリシーの適用箇所であり、ジェイルブレイクやプロンプトインジェクションの標的にもなる。

アクセス層は、利用者や他のプログラムがモデルを利用する窓口である。チャット画面とAPIに加え、ブラウザー拡張機能、IDEプラグイン、メッセージアプリ連携などの経路が増えている。入力形式も文章から画像、文書、音声に広がっており、ファイルに利用者から見えにくい指示が含まれることがある。APIキーや外部サービス連携用のOAuthトークンが漏えいすると、そのアカウントの権限でサービスや連携データが利用される可能性がある。

最外層のエージェント層では、モデルの判断に基づいてツールを呼び出し、複数段階の作業を行う。MCPサーバー、プラグイン、スキルなどがモデルと連携し、メールの送信、ファイルへのアクセス、コードの実行など、実際のシステムに影響を与える操作ができる。

多くのサービスでは、エージェントがコードを実行したりウェブに接続したりする際に、サンドボックスと呼ばれる隔離環境を使い、被害の範囲を抑える。外部コンポーネントによる機能拡張は便利だが、検証を経ていないコンポーネントはサプライチェーン攻撃の経路となる。エージェントに過剰な権限を与えると、小さな改ざんでも被害が拡大する。

4) AIの悪用とアンダーグラウンドAIの登場

生成AIが広く利用されるようになる一方で、安全対策を取り除いたり回避したりして、サイバー犯罪に悪用する動きも現れた。当初は独自の悪意あるモデルの販売が中心だったが、次第にChatGPTやClaudeなどの商用AIをジェイルブレイクして利用する形に移行した。

(1) アンダーグラウンドAIツールの登場

2023年、Lastと名乗る開発者が、オープンソースモデルGPT-J-6Bを基にしたWormGPTを地下フォーラムやTelegramで販売し始めた。制限のないツールと宣伝され、主にビジネスメール詐欺（BEC）に使われた。セキュリティ記者のBrian Krebsは販売者の過去の投稿を調べ、開発者がポルトガル出身で当時23歳のRafael Moraisだと特定した[13]。本人によれば、自ら書いたコードは全体の30～35%で、残りは他の開発者が寄与した。当時の有料利用者は約200人だった。身元の公表後、運営側は犯罪に関わる一部の回答やBEC関連機能を制限し、サービスは最終的に終了した。

CanadianKingpin12はFraudGPTを公開し、DarkBERT、DarkBARD、DarkGPTも宣伝した。FraudGPTの料金は月額200ドルから年額1,700ドルで、販売件数3,000件以上と宣伝された[14]。その後、販売投稿は姿を消した。

13 https://krebsonsecurity.com/2023/08/meet-the-brains-behind-the-malware-friendly-ai-chat-service-wormgpt/

14 https://outpost24.com/blog/dark-ai-tools/

EvilGPTやWolfGPTなどの後発ツールの中には、ChatGPTが返す拒否応答をそのまま返すだけの詐欺商品もあった[15]。

(2) 商用AIの安全対策の回避

初代WormGPTの終了後もFraudGPT、DarkBARD、Evil-GPTなどの模倣品・派生品が登場し、生成AIの犯罪利用は続いている[16]。2026年時点では、ダークウェブで流通するツールの多くは、GPT-4、Claude、Geminiなど既存の商用モデルの安全対策を回避できるとうたうラッパーである。初代WormGPTとコードを共有しておらず、同じ名称で販売されていても、実態は既存APIを利用したサービスに近い[17]。

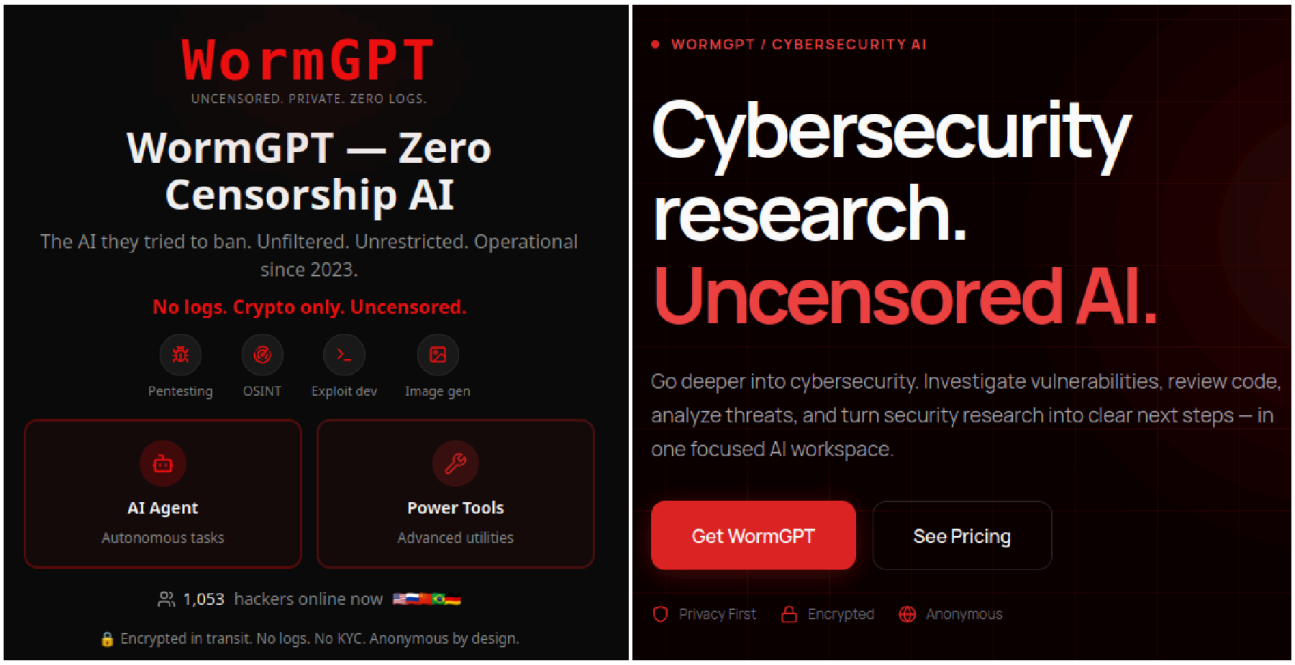


図2　販売されている複数のWormGPT

15 https://abnormal.ai/blog/what-happened-to-wormgpt-cybercriminal-tools
16 https://www.huntress.com/cybersecurity-101/topic/wormgpt
17 TorWiki — ダークウェブ上のAIツール

5) AI技術の発展と脅威の変化

AIは、2022年末に対話型チャットボットが登場してから3～4年で、自律的に行動するエージェントへ進化した。主な技術と脅威の変化を時系列で示す。

時期	技術	脅威
2022年	ChatGPTの登場、GitHub Copilotなどの普及開始	詐欺コンテンツの大量生成、多言語化、AIを使ったマルウェア開発の試み
2023年	外部サービスの操作、RAGの普及、GPT-4のコード生成能力向上	外部データの改ざん、プロンプトインジェクション、マルウェアの品質向上
2024年	Devin、Cursorなどのコーディングエージェントの登場	マルウェア開発の技術的障壁の低下
2024～2025年	MCPの標準化とエージェントCLIの登場	MCPサーバーやスキルを介したサプライチェーン攻撃、エージェント権限を悪用した認証情報の窃取
2026年	AIエージェントの実運用拡大、OpenClawなど個人向けエージェントの普及	エージェントを取り巻くスキルやMCPを狙うサプライチェーン攻撃の増加、ARTEXなどAIペネトレーションテスト用ツールの悪用

表3　AI技術の発展と脅威

(1) 対話型チャットボットの登場（2022年）

2022年11月にOpenAIがChatGPTを公開し、生成AIの普及が始まった。初期のサービスは主にチャット画面内で文章を生成し、外部システムとの連携はほとんどなかった。フィッシングメールや詐欺コンテンツを短時間で自然に作成できるようになり、時間と言語の障壁が下がった。

(2) ツール呼び出しと外部連携の始まり（2023年）

2023年6月、OpenAIが関数呼び出し機能を公開し、モデルが構造化された形式で外部システムの機能を呼び出せるようになった。RAGやベクトルデータベースが普及し、社内文書や最新情報を参照した回答も可能になった。モデルが参照する外部データを改ざんできれば、モデルの判断にも影響する。実行権限を持つシステムでは、プロンプトインジェクションが不適切な回答に加え、意図しない操作の原因にもなり得る。

(3) AIコーディングツールの普及（2024年）

2024年にはCognition AIのDevin、Cursor、Windsurfなど、コードを生成・実行できるツールが開発者に広まった。コード生成の支援により、マルウェア開発の技術的障壁も下がった。

(4) 標準プロトコルの登場とエージェントエコシステムの拡大（2024～2025年）

2024年11月、AnthropicはAIモデルと外部ツール・データソースを接続する標準プロトコルMCPを公開し、サービスごとに個別の連携を実装する負担を軽減した[18]。2025年2月にはClaude Code CLIを公開した。エージェントは開発者のローカル環境でファイルへのアクセスやコマンドの実行など、幅広い操作を自律的に行えるようになった。

[18] https://www.anthropic.com/news/model-context-protocol

MCPの標準化により、第三者のサーバーやスキルを手軽に導入して機能を拡張できるようになった。一方、外部コンポーネントを検証せず導入すると、サプライチェーン攻撃の経路となる。

(5) AIエージェントの普及とAIによる攻撃の現実化（2026年）

2026年に入り、メール、文書、日程管理、コーディング、業務自動化でエージェントの利用が進み、個人向けのOpenClawも短期間で広く普及した。2026年3月の報告では、AI関連の侵害の8件に1件以上が自律型エージェントに関係していた[19]。同年7月のIBMの調査では、悪意ある侵害の4件に1件がAIを使った攻撃で、前年比56%増だった。エージェントの幅広い権限に加え、スキルや拡張機能の検証不足がマルウェアの配布や認証情報の漏えいにつながっている。

攻撃にもエージェントが使われ、従来は個別に行われていた作業の自動化が進んだ。9月末から10月初めにかけて、韓国の金融機関を狙った攻撃でARTEXと複数のLLMが悪用され、データが漏えいした[20]。10月19日には日本企業を狙った攻撃インフラでもARTEXが確認された[21]。繰り返しの作業をエージェントが担うことで、少人数で複数の国や組織を狙うことが可能になり、高度な攻撃への参入障壁も下がった。

19 https://www.hiddenlayer.com/news/hiddenlayer-releases-the-2026-ai-threat-landscape-report-spotlighting-the-rise-of-agentic-ai-and-the-expanding-attack-surface-of-autonomous-systems

20 https://asec.ahnlab.com/ko/96187/

21 https://scan.netsecurity.ne.jp/article/2026/10/21/56522.html

4. AIをめぐる主な脅威事例

AIに関連する主なセキュリティ事例を類型別に整理する。各類型の事例と技術的な背景は、以降の章で取り上げる。

発生時期	攻撃対象	類型	概要
2024年2月	企業	AIの悪用	ディープフェイクでCFOと幹部になりすまし、従業員に2,500万ドルを送金させた。
2025年9月	IT・金融・化学企業、行政機関など約30組織	AIを活用した攻撃	中国の国家支援グループがClaude Codeを悪用し、自律型エージェントによる侵入を試みた。
2025年9月	メールサービスの利用者	AIシステムへの攻撃	npmパッケージpostmark-mcpのバックドアにより、送信メールが攻撃者に転送された。
2025年12月	ChatGPT・DeepSeekの利用者	AIシステムへの攻撃（会話・プロンプトの収集）	悪意のあるChrome拡張機能2種類が、会話を30分ごとに外部へ送信した。
2025年12月	メキシコの行政機関	Claude Codeによる攻撃の自動化	行政機関10カ所と金融機関1カ所への攻撃にClaude Codeが悪用された。
2026年2月	企業の役職員	AIを活用した攻撃（会話・プロンプトの収集）	BEC自動化サービスEvilTokensが窃取したメールをLLMで分析し、取引に合わせた詐欺メールを生成した。
2026年2月	GitHubリポジトリ	AIの悪用（攻撃の自動化）	GitHub Actionsの設定上の弱点を狙うAIボットが確認された。
2026年3月	LiteLLMを利用する企業	AIシステムへの攻撃（開発・ビルドのサプライチェーン）	CI/CDパイプラインの侵害により、バックドアを含むLiteLLMが配布された。

2026年4月	Claudeの利用者	AIサービス・プラットフォームへのなりすまし	偽の公式サイトで、Claude Proのインストーラーを装った不正ファイルが配布された。
2026年10月	韓国の金融機関	ARTEXを用いた攻撃	ウェブサイトの脆弱性を狙う攻撃にARTEXが悪用された。

表4　AIをめぐる主な脅威事例

事例は、AIサービスやツールへのなりすまし、攻撃へのAIの悪用、AIシステム自体への攻撃の三つに大別できる。時間の経過とともに各類型は高度化し、複数の類型が組み合わさる事例も増えている。

5. AIへの信頼を悪用する攻撃

AIへの関心と信頼を利用し、ウェブサイト、モバイルアプリ、ブラウザー拡張機能、APIサービス、モデルを装う事例がある。これらはマルウェアの配布、個人情報の窃取、金銭的被害につながる。

1) AIサービスやプラットフォームへのなりすまし

注目を集めるAIサービスが登場すると、それを模倣したウェブサイトやアプリが短期間で拡散する。公式モバイルアプリの公開前から、App StoreやGoogle PlayにはChatGPTを装ったアプリが多数掲載されていた。正規サービスに似た画面やブランド要素で利用者を信用させ、連絡先、SMS、通話履歴、認証コードを収集したり、バックグラウンドで端末を継続的に監視したりするものがあった。

2026年1月、商標の問題でオープンソースのAIエージェントClawdbotがMoltbot（後のOpenClaw）に改名した。

改名時に空いたGitHubの組織名とXのアカウント名が、数秒で第三者に取得された[22]。正規開発者のPeter Steinbergerを装ったサイトとリポジトリが公開され、Windows向けの偽インストーラーがBingのAI検索結果の最上位に表示され、マルウェアの配布につながった。

2026年3月には、正規の広告主アカウントを悪用したGoogle検索広告を通じて、Claude Codeを検索した開発者が、公式文書の体裁を模したページへ誘導された。この事例では、Windowsに情報窃取型マルウェア、macOSにバックドアが導入された[23]。4月にはClaudeの公式サイトを装い、Claude Proのインストーラーに見せかけた不正ファイルを配布する事例が確認された[24]。AIの回答画面を模したページで、利用者に危険な操作を促す事例もあった。

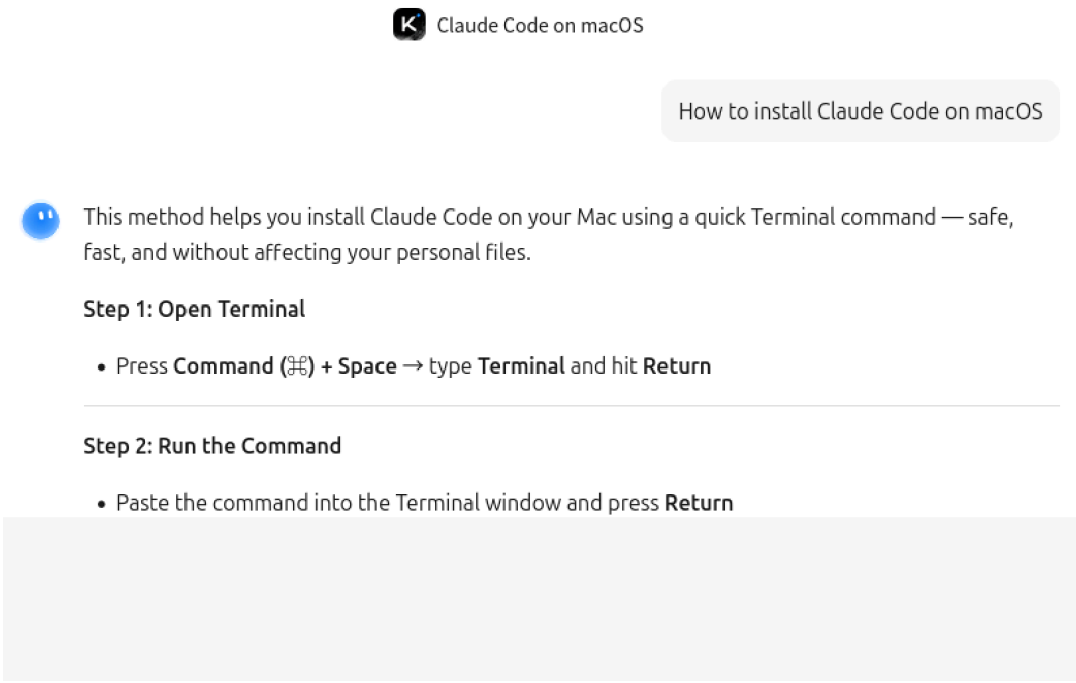


図3　AIの出力を装ったマルウェア感染の誘導

実行用の記述は伏せています。

22 https://atip.ahnlab.com/intelligence/view?id=6fc470a2-a37d-4a50-9166-c20d4c908782
23 https://www.bitdefender.com/en-us/blog/labs/fake-claude-code-google-ads-malware
24 https://www.malwarebytes.com/blog/scams/2026/04/fake-claude-site-installs-malware-that-gives-attackers-access-to-your-computer

攻撃者はAI機能の有無にかかわらず、知名度の高いサービスへの信頼を借りて、検索広告、AI検索結果、よく似たドメインなどからマルウェア配布ページへ誘導する。ダウンロード時には、URLと公式の配布経路を直接確認する必要がある。

2) AI拡張機能やプラグインを装う攻撃

ブラウザー拡張機能は、付与された権限に応じてページ内容、閲覧履歴、セッション情報にアクセスできる。2023年3月、正規の「ChatGPT For Google」を装った拡張機能が検索広告で拡散し、Facebookのセッションクッキーが窃取され、ビジネスアカウントが乗っ取られた[25]。2025年12月には、90万人を超えるChatGPT・DeepSeek利用者が悪意のあるChrome拡張機能2種類をインストールしていた。会話のほか、業務情報や認証情報、セッションデータが30分ごとに外部へ送信された[26]。

2026年2月に確認されたAiFrameでは、Claude、ChatGPT、Gemini、Grokの関連ツールを装う拡張機能が約30種類見つかり、利用者は26万人を超えた。一部はストアの注目バッジ（Featured）を取得していた。機能の多くがリモートのコンテンツに依存していたため、公開時の審査だけではその後の挙動を把握しにくかった。閲覧ページやGmailの本文が収集され、ストアから削除された後、約2週間で再掲載された例もあった[27]。

25 https://guard.io/labs/fakegpt-2-open-source-turned-malicious-in-another-variant-of-the-facebook-account-stealer
26 https://www.ox.security/blog/malicious-chrome-extensions-steal-chatgpt-deepseek-conversations
27 https://layerxsecurity.com/blog/aiframe-fake-ai-assistant-extensions-targeting-260000-chrome-users-via-injected-iframes

ChatGPTの無料プランへの広告導入後、「ChatGPT Ad Blocker」と称する拡張機能がChromeウェブストアに登場し、プロンプトと会話情報が外部へ送信された[28]。注目バッジやインストール数は無害性を保証せず、リモート設定も公開時の審査だけでは捉えにくい。AIに貼り付けられる業務文書やソースコードは、拡張機能の権限を通じて漏えいする可能性がある。

3) 偽のAI APIサービス

無料または安価に正規モデルを使えると称するAPIでは、正常に見える応答が返ってきても、仲介事業者にプロンプト、応答、APIキー、サーバーの接続情報が渡るおそれがある。2024年11月、PyPIのgptplusとclaudeai-engは、それぞれGPT-4 TurboとClaude APIへのアクセスを提供するとうたっていた。実際にはデモ版を利用して応答を返し、その裏でJarkaStealerを配布して、ブラウザーのデータやTelegram・Discordのセッション情報を窃取した[29]。

2026年には、一つの接続先で複数のAIモデルを仲介するAPIルーターが新たなリスクとなった。4月に公開された研究では、調査対象の428件のうち9件で応答に悪意のある内容が見つかり、17件で検証用のAWS認証情報の利用が、1件でテスト用ウォレットからのイーサリアムの移転が確認された。検証対象のエージェントの440セッションのうち401件では、人の承認を経ずにツールが実行されていた[30]。

28 https://dti.domaintools.com/securitysnacks/securitysnack-openai-anti-ads-malware
29 https://www.kaspersky.com/blog/jarkastealer-in-pypi-packages/52640/
30 https://arxiv.org/html/2604.08407v1

API仲介サービスを経由する環境では、AIの回答が正常に見えても機密情報が漏えいする可能性がある。エージェントにファイル変更やコード実行の権限があれば、システムへの影響もあり得る。AI APIへの接続には公式エンドポイントまたは信頼できる事業者を利用し、相場を大幅に下回る料金には注意する。

4) 正規AIサービスの機能の悪用

2026年5月、ChatGPTの会話共有機能と、生成されたHTML・CSSを表示する機能が悪用された。利用者はサービス障害を装う案内から、不正な実行ファイルを配布する偽のダウンロードサイトへ誘導された[31]。7月にはClaudeの公開アーティファクト機能と検索広告が悪用され、Claude Desktopのインストール画面を装うページに利用者が誘導された[32]。

共有ページが正規ドメイン上にあっても、掲載された内容の信頼性は別途確認する必要がある。共有機能の使いやすさ、検索結果や広告の表示、利用者による上位リンクへの信頼が重なるところで同じリスクが生じる。サービス運営者には、公開コンテンツの検証と通報への対応体制が求められる。

31 https://pushsecurity.com/blog/llmshare-malvertising-campaign

32 https://www.huntress.com/blog/fakeagent-claude-desktop-malvertising-ends-in-dotnet-rat

5) AI関連資料を装う攻撃

2026年6月、PostgreSQLやClaude CodeなどのAI関連学習資料を装った7zファイルを通じて、AsyncRATが配布された。圧縮ファイルにはショートカット（LNK）と、隠しファイルとして扱われたPDFが含まれていた。同年4～6月には、架空のAI自動売買ツールを宣伝し、CoinbaseやMetaMaskなどの正規ウォレット用拡張機能を、見た目が同じ改ざん版に置き換える事例もあった[33]。初めて目にするAIツールや資料を入手する際は、提供元が実在し、配布経路を公式情報で確認できるかを確かめる必要がある。

6. AIを活用した攻撃（AI-Augmented Attacks）

生成AIは、ソーシャルエンジニアリング、マルウェア開発・配布、侵入作業の自動化に使われている。マルウェアにAI機能を組み込む事例や、AIモデルにバックドアを仕込む事例もある。

1) ディープフェイクを用いたなりすましと金融詐欺

合成音声や合成映像による企業幹部、家族、機関関係者へのなりすましが、金融詐欺や情報窃取に悪用されている。リアルタイムの映像合成・音声複製を提供するサービス（DaaS）の市場もアンダーグラウンドで拡大している。

33 https://www.hp.com/us-en/newsroom/press-releases/2026/hp-research-cybercriminals-leaning-into-agentic-ai-momentum-to-steal-crypto-wallets.html

複数の調査では、ディープフェイクを使った電話詐欺の被害が拡大し、人が合成を見抜ける割合は低下している[34]。2024年2月の香港の事件では、CFOと幹部の顔や声を合成したビデオ会議で従業員がだまされ、2,500万ドルを送金した[35]。合成コンテンツが実際の金融被害に直結した初期の代表例である。その後も、経営者を装う映像や音声で従業員に送金を指示したり、銀行幹部を装う電話で支店長に資金を移させたりする事例が報告された[36]。

家族や知人の声を合成し、誘拐を装ったり緊急送金を要求したりする詐欺も各地で確認されている。テキストメッセージから音声通話へ誘導するなど、複数の連絡手段を組み合わせた手口も現れている。

2) 詐欺コンテンツの作成

生成AIは、詐欺文面の推敲に加え、行政機関を装うサイトや、実際の取引に合わせた詐欺メールの作成にも使われる。ウェブ開発や外国語の専門知識が乏しくても、それらしいコンテンツを短時間で量産できる。2025年8月に公表されたブラジルの事例では、AIウェブ制作ツールのDeepSite AIとBlackBox AIが、交通当局と教育省を装うサイトの制作に悪用された[37]。

34 https://www.group-ib.com/blog/ai-cybercrime-usecases/
35 https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk
36 https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk
37 https://www.zscaler.com/blogs/security-research/genai-used-phishing-websites-impersonating-brazil-s-government

ブラジルの事例では、無料の運転免許証や求人をうたう偽サイトが、検索結果を操作するSEOポイズニングによって上位に表示された。偽サイトに納税者登録番号（CPF）を入力した利用者には、氏名や生年月日などが表示され、正規の行政サイトであるかのように見えた。その後、登録料としてブラジルの即時決済システムPixでの支払いを求められた。

2026年2月中旬から販売されたフィッシング・アズ・ア・サービス（PhaaS）のEvilTokensでは、Microsoftアカウントの認証情報が窃取され、メールをLLMが分析した。取引の文脈に合わせてビジネスメール詐欺（BEC）の文面が生成されるため、攻撃者がメールを一通ずつ確認する負担が軽減された[38]。同サービスは同年9月に停止された[39]。別の調査では、AIウェブ制作ツールで作られた41万3,524サイトのうち、3万7,447件（9.06%）が悪意のあるサイトに分類された[40]。表示や文章が自然になるにつれ、誤字や不自然な表現といった従来の手掛かりだけで真偽を判断することは難しくなっている。

3) マルウェアの作成と配布

(1) マルウェアの作成と開発支援

生成AIは、専門知識の乏しい個人にもマルウェアの開発を可能にし、熟練した攻撃者の開発も加速させる。2024年5月、日本で、データの破壊・暗号化機能を持つランサムウェア型のマルウェアを生成AIで作成した疑いで、25歳の男性が逮捕された。

38 https://blog.sekoia.io/eviltokens-an-ai-augmented-phishing-as-a-service-for-automating-bec-fraud-part-2/
39 https://blogs.microsoft.com/on-the-issues/2026/09/22/disrupting-eviltokens-the-ai-chatbot-built-for-cybercrime/
40 https://www.zscaler.com/blogs/security-research/one-click-compromise-threatlabz-2026-phishing-and-initial-access-report

捜査で、男性が複数のAIを組み合わせて開発したことが明らかになった。他人になりすましてSIMカードを発行させた過程から、身元が判明した[41]。VoidLinkでは、AIが高度なマルウェアフレームワークのコードの大半を生成し、単独の開発者が約1週間で完成させたと報告された[42]。AIで構造の異なる変種を生成する「プロンプトモーフィズム」は、シグネチャに依存する検知を難しくする[43]。

パキスタンを拠点とするTransparent Tribe（APT36）は「Vibeware」と呼ばれる開発手法で変種を量産した。未完成のサンプルもあったが、正規のコミュニケーションサービスが悪用された形跡が確認された[44]。北朝鮮と関連するFamous ChollimaはChatGPTとCursorをマルウェアの開発に利用し、調査ではバイブコーディングで作られたツールが二つ見つかった。攻撃者がプロンプトを誤って公開したことで、AIの利用が判明した[45]。

41 https://www.yomiuri.co.jp/national/20240528-OYT1T50015
42 https://research.checkpoint.com/2026/voidlink-the-cloud-native-malware-framework/
43 https://www.gendigital.com/blog/insights/research/promptmorphism
44 https://businessinsights.bitdefender.com/apt36-nightmare-vibeware
45 https://expel.com/blog/inside-lazarus-how-north-korea-uses-ai-to-industrialize-attacks-on-developers/

(2) マルウェアの配布支援

生成AIは、マルウェアを正規のファイルやソフトウェアに見せかけるダウンロードページ、案内文、動画の制作に使われる。2025年2月以降、メールを起点とする脅威から、ウェブ制作ツールLovableで作成されたページのURLが毎月数万件検出された[46]。2025年5月に公表された事例では、Windows、Microsoft Office、Spotifyなどを無料で認証できるとうたうTikTok動画を通じて、情報窃取型マルウェアVidarとStealCが広がった。再生数が約50万回に達した動画もあった。複数のアカウントにほぼ同じ動画が投稿されており、分析機関は案内音声などがAIで生成された可能性が高いとみている[47]。

動画で操作を促す場合、プラットフォーム上に直接検査できる実行ファイルが残らないことがある。制作にかかる時間が短くなるほど、攻撃者は誘導用のページや動画を頻繁に差し替え、広く配布できる。

4) 攻撃の実行と運用の支援

AIの悪用は、配布用コンテンツの制作から、攻撃の実行と運用にも拡大している。

46 https://www.proofpoint.com/us/blog/threat-insight/cybercriminals-abuse-ai-website-creation-app-phishing

47 https://www.trendmicro.com/en/research/25/e/tiktok-videos-infostealers.html

AIは侵入に加え、窃取したデータの分析などにも悪用される。少人数の攻撃者が短期間に多くの組織を狙えるようになる。2025年9月中旬、中国の国家支援グループとされるGTG-1002が、技術、金融、化学、行政など約30組織への攻撃にClaude Codeを悪用した。作業の80～90%をAIが担当し、人の介入はキャンペーンごとに4～6回の重要な判断に限られたとされる。Anthropicは2025年11月にこの事例を公表した[48]。2026年2月、600台以上のFortiGate機器が侵害された事例では、AIを使うセキュリティテスト基盤CyberStrikeAIが関係した可能性が報告された[49]。

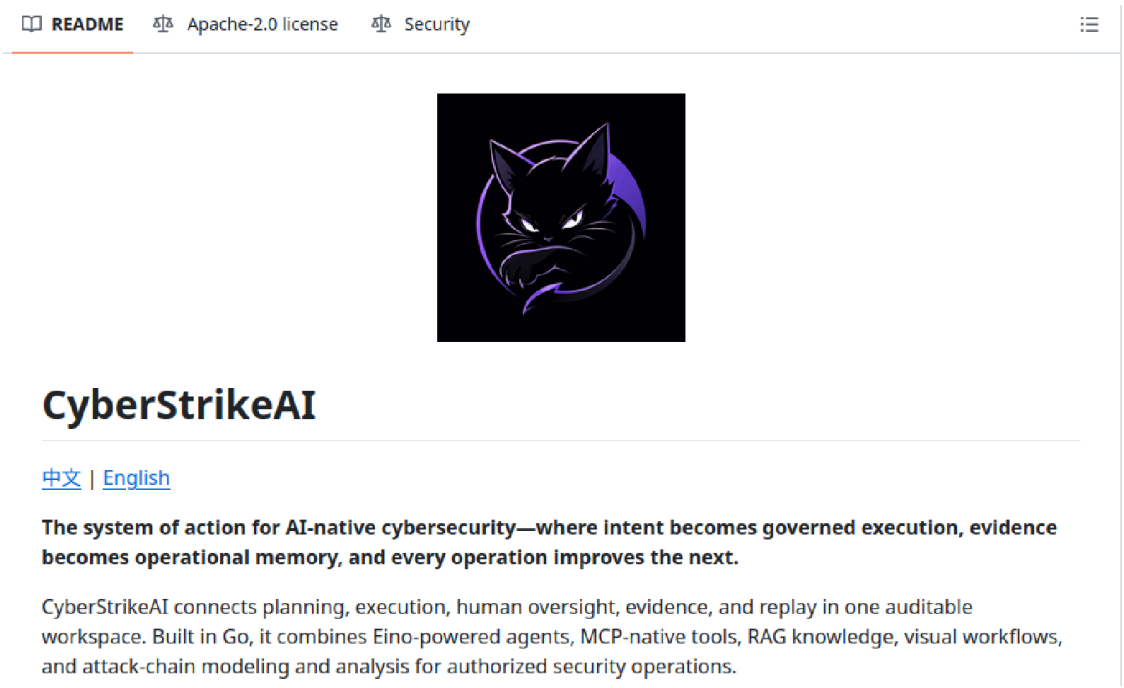


図4 CyberStrikeAIの紹介

48 https://www.anthropic.com/news/disrupting-AI-espionage

49 https://www.team-cymru.com/post/tracking-cyberstrikeai-usage

2026年2月以降、hackerbot-clawと名乗るGitHubアカウントが、Microsoft、Datadog、Aqua Security（Trivy）など7カ所のGitHub Actionsワークフローを狙った。同アカウントはClaude Opus 4.5を基盤とする自律型のセキュリティ研究エージェントを自称していた[50]。

2026年9月末から10月初めに韓国の金融機関が狙われ、銀行、貯蓄銀行、与信専門金融会社の顧客情報が相次いで漏えいした。攻撃側の記録からClaude CodeとARTEXの利用が確認された[51]。2025年12月末のメキシコの行政機関への侵害では、攻撃者がClaude Codeに1,000件を超えるプロンプトを送信し、内部データの分析などを行わせた。一部でGPT-4.1も使用され、住民登録、税務、有権者情報など150GB以上が漏えいした。攻撃の各段階が短時間で進むため、組織の対応が追い付かなくなるおそれがある。

5) AIと連携するマルウェア

AIが開発時に用いられるだけでなく、マルウェアの実行中にモデルが判断や出力の生成を担う事例も確認されている。

50 StepSecurity — GitHub Actionsの侵害事例

51 https://www.crowdstrike.com/en-us/blog/unknown-threat-actor-uses-artex-to-target-south-korean-finance/

2025年8月に確認されたPromptLockは、ニューヨーク大学（NYU）の研究者によるAIランサムウェアの概念実証である[52]。PromptFluxは、Google Gemini APIを使った自己書き換え機能を備えるドロッパーの事例である[53]。PromptStealは、外部モデルの出力を利用するPythonベースの情報窃取型マルウェアである。

2026年1月に確認されたPhantomは、スマートフォンの画面を外部で解析し、広告クリックを自動操作した。広告によって挙動を変え、遠隔接続機能も備えていた[54]。同年2月のPromptSpyは、Androidの画面の状態を把握し、インターフェースの操作を判断するために生成AIを悪用した。報告では、スマートフォンの画面操作に生成AIを利用する初のAndroidマルウェアと位置付けられている[55]。

AIの応答に依存するマルウェアは挙動を固定的に捉えにくい。一方、APIキーやプロンプトが検知・追跡の手掛かりとなり、AI事業者が利用アカウントやキーを停止すれば主要機能を無効化できる。確認済みの多くは実験または初期段階の事例だが、実行時の判断や端末に応じた動作により今後も増える可能性がある。

52 https://www.eset.com/gr-en/about/newsroom/press-releases-1/eset-discovers-promptlock-the-first-ai-powered-ransomware-1/

53 https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools

54 https://news.drweb.com/show/?i=15110

55 https://www.welivesecurity.com/en/eset-research/promptspy-ushers-in-era-android-threats-using-genai/

7. AIサービスおよびプラットフォームへの攻撃

AIサービスは、モデル、コンテキスト、エージェントの実行権限、外部ツールとの連携など、複数の要素から成る。これらは、従来のアプリケーションにはない攻撃対象領域を生む。特に、利用者の権限を引き継いで自動で動作するAIアシスタントや自律型エージェントでは、入力の改ざんや設定ミスにより、一度の要求で情報が漏えいしたり、権限が悪用されたりするおそれがある。本章では、サービスやモデルへの攻撃、入力を操作する攻撃、サプライチェーン攻撃、AI関連情報の窃取を取り上げる。

1) AIサービスおよびプラットフォームへの攻撃

(1) AIブラウザー

AIブラウザーは、ウェブの閲覧に加え、予定の管理、文書の要約、アカウントへのアクセスなどを自律的に行う。エージェントに利用者と同等の権限があると、被害もその権限の範囲に及ぶ。2026年3月3日、PerplexityのCometで、カレンダー招待の処理を起点にローカルファイルが外部へ流出するおそれのある脆弱性が公表された。利用者側には正常な応答が返る一方でデータ送信が起こり得た。1Password拡張機能がロック解除されている場合は、保存された認証情報やアカウントにも影響が及ぶ可能性があった[56]。ウェブコンテンツとエージェントの実行権限の境界を明確にする必要がある。

[56] https://www.zenity.io/company-overview/newsroom/company-news/zenity-labs-discloses-pleasefix-perplexedagent-vulnerability

(2) AIアシスタント

ChatGPT、Claude、Google Gemini、Microsoft CopilotなどのAIアシスタントは、コネクターの導入によりメール、文書、予定、社内システムに接続するようになった。2026年1月、Microsoft 365 Copilot Chatが、データ損失防止（DLP）ポリシーと機密ラベルに従わず、送信済みメールや下書きを要約する問題が確認された[57]。既存の保護ルールがAI機能にも適用されるかを検証する必要がある。

(3) MCP（Model Context Protocol）

MCPの設定には、開発者の端末で行われる処理を指定する項目がある。設定の改ざんや承認手続きの不備は、コード実行や権限の悪用につながり得る。CursorのCurXecute（CVE-2025-54135）では、連携先から取得した外部コンテンツをエージェントが信頼したため、MCP設定が改ざんされ、利用者が変更を拒否しても意図しないコード実行を防げない場合があった[58]。MCPoison（CVE-2025-54136）では、承認済みの設定に後から変更が加わっても、改めて確認を求められなかった[59]。

57 https://www.bleepingcomputer.com/news/microsoft/microsoft-says-bug-causes-copilot-to-summarize-confidential-emails/
58 https://nvd.nist.gov/vuln/detail/CVE-2025-54135
59 Check Point Research — MCPoisonの報告

MCPoisonでは、共有プロジェクトの設定変更が開発者の再承認なしに反映され、開発環境を開くたびに不正な処理が繰り返されるおそれがあった。Anthropicが提供する公式MCP SDKでも、一部の接続設定に信頼できない入力が含まれると、命令として解釈される問題が確認された。関連するSDKのダウンロードは1億5,000万件以上、インターネットからアクセス可能なサーバーは7,000台以上、脆弱なインスタンスは最大20万件と推定された。LiteLLM、LangFlow、Flowise、Windsurfなどで、30件以上の脆弱性が報告された[60]。

2025年12月には、MCPサーバーからクライアント側のLLMに応答生成を依頼できるサンプリング機能について、三つの概念実証が公表された。報告された影響は、トークン利用枠の消費、会話の乗っ取り、意図しないツール呼び出しである[61]。IDEのエージェントは開発者の権限で動くため、設定の再承認と最小権限によって、チームや下流プロジェクトへの波及を抑えることが重要である。

(4) 自律型AIエージェント

OpenClawはメッセージアプリ、カレンダー、開発ツールと連携し、コマンドやメッセージ送信を実行できる[62]。2026年2月の調査では、最初の24時間で4万件以上のインターネット公開インスタンスが見つかった。

60 https://www.ox.security/blog/the-mother-of-all-ai-supply-chains-critical-systemic-vulnerability-at-the-core-of-the-mcp
61 Unit 42 — MCPのセキュリティ調査
62 https://openclaw.ai/

公開インスタンスの35.4%が脆弱と分類された[63]。同月に公開されたCVE-2026-25253とClawJackedでは、悪意のあるウェブサイトに接続すると、OpenClawの管理機能や権限に影響が及ぶ可能性が確認された。エージェントが乗っ取られると、連携状況に応じてメッセージ、APIキー、ファイルが危険にさらされる[64、65]。ローカル接続を前提とするサービスにも、適切な認証と権限制御が必要である。

2026年5月の別の事例では、公開されたGrokの出力を取引エージェントBankrbotが有効な指示として扱い、Baseチェーン上でトークンを送金した。AIの出力が検証を経ずに金融取引につながる設計で、再確認や送金上限が設けられていなかった[66]。管理経路を侵害される場合も、信頼する入力を操作される場合も、被害の大きさはエージェントに与えた権限と確認手続きに左右される。

(5) AI開発・運用プラットフォーム

MLOpsやモデル提供のためのツールはインターネットに公開されることがあり、クラウド認証情報、GPU、モデルファイルに直接アクセスできる。そのため、モデルそのものに加えて、構築・配布のパイプラインも狙われる。

63 https://securityscorecard.com/blog/how-exposed-openclaw-deployments-turn-agentic-ai-into-an-attack-surface/
64 depthfirst — OpenClawの脆弱性報告
65 Oasis Security — ClawJackedの報告
66 https://slowmist.medium.com/behind-the-grok-exploitation-an-analysis-of-ai-agent-permission-chain-abuse-4d832d1bfc73

2025年11月に公表されたShadowRay 2.0では、AIの学習・推論を分散処理するRayの公開環境（CVE-2023-48022）が侵害された。クラスタのGPUが暗号資産の採掘に使われ、他システムへの攻撃やDDoSにも利用された。公開Rayサーバーは23万台以上とされ、ペイロードの作成にLLMが使われた可能性が指摘された[67]。

MLflowでは、モデルレジストリのSSRF（CVE-2026-64849）が2026年8月17日のCVE番号割り当てから数時間で悪用され、クラウドの認証情報やシークレットが窃取された[68]。Langflowでは2026年より前に悪用が確認された脆弱性は1件だったが、2026年に新たに11件が悪用された。同年8月末のCVE-2026-0768では、APIキー、クラウド認証情報、SSH鍵、ソースコードなどが窃取された[69]。

LLMの提供に使われるLMDeployのSSRF（CVE-2026-33626）は、2026年4月21日の公表から12時間以内に悪用の試みが観測された。その短時間にクラウド環境や内部サービスが狙われた[70]。開発・運用プラットフォームの侵害は、クラウドアカウントや接続された他のAIサービスにも波及し得る。

67 https://www.oligo.security/blog/shadowray-2-0-attackers-turn-ai-against-itself-in-global-campaign-that-hijacks-ai-into-self-propagating-botnet

68 https://www.linkedin.com/posts/watchtowr-intel-is-observing-in-the-wild-share-7495481579608236033-RjIj/

69 https://www.vulncheck.com/blog/pwning-the-ai-stack

70 https://www.sysdig.com/blog/cve-2026-33626-how-attackers-exploited-lmdeploy-llm-inference-engines-in-12-hours

(6) ローカルLLMへの攻撃

2026年1月、130カ国で17万5,108件の一意のOllamaホストが観測された。これらはクラウドサービス提供者による監視や悪用防止の対象外で運用されている[71]。認証なしで公開されたモデルは、データ漏えいの原因となるほか、第三者の攻撃にも悪用され得る。2026年10月には、公開モデルを悪用する攻撃エージェントがハニーポットで確認された[72]。

2025年10月から2026年1月までにOllamaのハニーポットで9万1,403件の攻撃セッションが観測された。別の事例では、二つのIPアドレスから11日間で8万469件のセッションが観測され、少なくとも73件のLLMエンドポイントが調べられた[73]。2026年5月に公表されたBleeding Llama（CVE-2026-7482）では、認証なしでプロセスメモリにアクセスでき、利用者やシステムのプロンプト、クラウド認証情報が漏えいするおそれがあった[74]。

インストールや更新にも課題がある。2025年7月には、macOS版Ollamaのインストーラーにコマンド実行に関わる問題が公表された。2024年末に報告された問題で、公表時点では修正されていなかった[75]。2026年4月以降のPoeLLMでは、公開されていたLiteLLMやOllamaなどが悪用され、2,000台以上のサーバーが暗号資産の採掘などに利用された[76]。

71 https://www.sentinelone.com/labs/silent-brothers-ollama-hosts-form-anonymous-ai-network-beyond-platform-guardrails
72 https://www.security.com/threat-intelligence/ai-llmjacking-skynet
73 https://www.greynoise.io/blog/threat-actors-actively-targeting-llms
74 Cyera — Bleeding Llamaの報告
75 Imperva — Ollamaインストーラーの脆弱性報告
76 https://www.lumen.com/blog/en-us/canto-incognito-tracking-the-poellm-malware

ローカル運用では、運用者が認証、ネットワークアクセス、資産管理、パッチ適用を担う。認証のない外部公開は、モデル、会話、クラウド認証情報、GPUを同時に危険にさらす。サーバーの侵害がクラウドアカウントの侵害や第三者への攻撃に波及することもある。インストーラーやアップデートも管理対象となる。

2) AI・LLMを標的とする攻撃手法

(1) RAG・検索の改ざんとサービス妨害

RAGなどで外部情報を参照する場合、参照先が汚染されると回答にも影響する。2025年12月、公開ウェブの内容の改ざんにより、PerplexityとGoogle AI OverviewがEmiratesの問い合わせ先として詐欺電話番号を案内した。一部で正しい番号が返った場合にも、汚染されたページが出典に含まれていた[77]。2026年2月には、ウェブサイトのAI要約機能をきっかけにアシスタントのメモリが汚染される「AI Recommendation Poisoning」が報告された[78]。

[77] https://aurascape.ai/resources/auralabs-research/llm-search-poisoning-fake-support-numbers/

[78] https://www.microsoft.com/en-us/security/blog/2026/02/10/ai-recommendation-poisoning/

この手法では、サービス利用時に意図しない内容がメモリに記録され、その後の会話に影響を及ぼす可能性があった。60日間のメール通信で50件の異なるプロンプトが確認され、31社、金融・医療・法律・SaaSなど14業種が関係していた。報告では企業の販促目的とされており、LLM向けのSEOとしても紹介されていた。

従来型のサービス妨害も発生している。2023年11月のChatGPTとAPIの障害について、OpenAIはDDoSとみられる異常なトラフィックを原因として挙げ、Anonymous Sudanが犯行を主張した[79]。2026年3月には、ウェブコンテンツを介してエージェントの処理資源を消費させたり、サービス妨害を引き起こしたりする試みが報告された[80]。検索、メモリ、知識ベースを汚染する攻撃では、AIが複数の情報源を一つの断定的な回答にまとめるため、利用者からは出典ごとの問題が見えにくい。

(2) プロンプトインジェクション

プロンプトインジェクションは、AIが処理する入力に悪意ある指示を混ぜ、意図しない動作を誘発する手法である。自然言語を使うため検出が難しく、ログがなければ機密情報へのアクセスの試みが把握できない場合もある。直接型では、利用者からの入力がシステム指示や安全ポリシーに干渉する[81]。

79 https://status.openai.com/incidents/21vl32gvx3hb
80 https://unit42.paloaltonetworks.com/ai-agent-prompt-injection/
81 https://www.microsoft.com/en-us/security/blog/2026/03/12/detecting-analyzing-prompt-abuse-in-ai-tools/

間接プロンプトインジェクション（IDPI）では、AIが読み取るウェブページ、文書、メールなどに含まれた指示が、意図しない動作の原因となる。2025年8月に公表されたAgentFlayerでは、ChatGPTのコネクターを通じてGoogle Drive上の機密情報が、利用者の追加操作なしに外部に送信される可能性が実証された[82]。HashJackでは、通常のURLに由来する入力により、Comet、Copilot for Edge、Gemini for Chromeで偽の問い合わせ先の案内や虚偽の情報が表示されるおそれがあった。Cometでは利用者情報の外部送信も確認された[83]。

GitHub Actionsを狙った事例では、7件の標的のうち少なくとも4件でリモートコード実行に至った[84、85]。2026年3月には、Microsoft Copilotのメール要約に偽のセキュリティ警告とリンクが表示される事例が公表された。利用者が要約をシステムからの案内と見なすことを利用した、モデルを介したフィッシングである。一部の環境では、Teamsなどの社内情報にも影響が及ぶ可能性が確認された[86]。

実際のウェブ上でも、広告審査の回避、SEOの操作、情報漏えい、データ破壊、サービス妨害、無断取引などを目的とする不審な指示が検出された[87]。間接型の攻撃ツールは、メール、PDF、カレンダー招待、ウェブページ向けとして地下フォーラムで宣伝されていた[88]。

82 Zenity Labs — AgentFlayerの報告
83 https://www.catonetworks.com/blog/cato-ctrl-hashjack-first-known-indirect-prompt-injection/
84 StepSecurity — GitHub Actionsの侵害事例
85 https://www.datadoghq.com/blog/engineering/stopping-hackerbot-claw-with-bewaire/
86 Permiso — Copilotを介したフィッシングの調査
87 https://unit42.paloaltonetworks.com/ai-agent-prompt-injection/
88 https://www.proofpoint.com/us/blog/threat-insight/notes-underground-adversarial-prompt-injection

カレンダー招待などを通じて、エージェントに意図しない外部送信をさせようとする試みも報告された。報告の多くは研究や推定に基づき、この時点で広範に普及した手法とまでは確認されていない。

(3) AI分析ツールの回避

マルウェアやパッケージの分類にAIが使われるようになり、サンプル内の文字列で分析結果を歪めようとする例が現れた。分析依頼とファイルの内容を区別できないモデルは、検査対象の文字列を指示として扱うおそれがある。2025年6月に発見された試作段階のSkynetは初期の事例である。2025年1月から2026年7月までに、4系統84サンプルで類似の手口が確認された[89、90]。観測された指示は、主にファイルを無害と判定させるものと、判定自体を阻むものに分かれる。

検体内の分析妨害プロンプトとコードの例

図5　AI分析の回避を狙うマルウェア内のプロンプト

実行用の記述は伏せています。

FRUITSHELLに含まれる文面は、15カ月間に少なくとも四つの攻撃主体のスクリプトで再利用された。PLOTSAFEでも、同じ文面を再利用した形跡が確認された。

89 Check Point Research — AI分析へのプロンプトインジェクション

90 https://blog.talosintelligence.com/ignore-all-instructions-and-read-this-blog-the-state-of-ai-analysis-evasion-in-malware/

著作権や知的財産権を理由に分析の拒否を求めたり、モデルの応答を途中で止めてサンプルの後半を判定させなかったりする試みもあった。実験では、応答を続けられるかどうかはモデルと利用形態によって異なった。2026年8月末には、ロシアとの関連が指摘されるUAC-0099がウクライナを狙ったVBSスクリプトに分析を妨げる記述を加えた事例や、npmパッケージのファイルを9.28MBに膨らませ、モデルの入力上限を消費させようとした事例が報告された[91~93]。

5種類のローカルLLMを比較した検証では、最も効果が高い手法でも、攻撃側に有利な判定に変わったのは実行の約35%だった。モデルが読み取れる文字列は検知の手掛かりにもなり、従来の静的・動的解析は、こうした自然言語の指示の影響を受けない。

3) AI連携コンポーネントへのサプライチェーン攻撃

MCPサーバーやエージェントスキルは、エージェントが信頼して使うツールや指示を提供するため、サプライチェーン攻撃の標的となる。

(1) 悪意あるMCPサーバーとエージェントサーバーのサプライチェーンの悪用

利用者の端末で動くMCPサーバーには、環境変数、APIキー、ローカルファイルへのアクセス権を持つものがある。悪意あるサーバーを導入すると、その権限の範囲で情報が窃取され得る。

91 JFrog — AIスキャナーへのプロンプトインジェクション
92 https://x.com/ESETresearch/status/2092885117286879707
93 https://socket.dev/blog/npm-package-uses-prompt-injection-and-token-flooding-to-disrupt-ai-malware-scanners

2025年9月、悪意のあるMCPサーバーの初期事例として、npmのpostmark-mcp 1.0.16にバックドアが見つかった。パスワード再設定、金融関連の通知、社内メモを含む送信メールが外部へ転送された[94]。

2026年2月、開発用ユーティリティなどを装う19件のnpmパッケージを通じて広がったワームSANDWORM_MODEは、Claude Desktop、Cursor、VS Code、Windsurfに悪意のあるMCPサーバーを登録し、SSH鍵、AWS認証情報、npmトークンを窃取した[95]。同年8月には公式MCPレジストリの登録項目V.A.P.Eを介したShai-Huludの変種が確認された。PyPI上のパッケージ自体は正常だったものの、関連リポジトリを開いた開発環境で、開発者トークンやクラウド認証情報が窃取された[96]。

公式レジストリに登録されていても、コンポーネントの安全性が保証されるわけではない。パッケージ自体に加え、MCPやプロジェクトの設定、サーバーがアクセスできるファイルや環境変数の範囲も確認する必要がある。

(2) 悪意あるAIスキルの配布

外部のマーケットプレイスを通じて配布されるスキルは、エージェントの指示やツールをまとめたコンポーネントである。文書と実行内容の両方を確認する必要がある。

94 https://snyk.io/blog/malicious-mcp-server-on-npm-postmark-mcp-harvests-emails
95 https://socket.dev/blog/sandworm-mode-npm-worm-ai-toolchain-poisoning
96 https://www.ox.security/blog/shai-hulud-outbreak-debrief-the-worm-evolves-into-mcp

2026年2月、OpenClawのスキルマーケットプレイスClawHubで2,857件を調査した結果、341件が悪意のあるスキルと判定され、335件が一つのキャンペーンClawHavocに関連していた[97]。VirusTotalは3,016件を超えるスキルを分析し、特定の利用者が投稿した314件はすべて悪意のあるものだったと報告した。Windows用のトロイの木馬やAtomic macOS Stealer（AMOS）が確認され、ファイル自体にマルウェアがなくても、記載された作業手順を通じて配布される場合があった[98]。

不正な外部接続を確立したり、APIキーを窃取したりするスキルも確認された[99]。AMOSがOpenClawのスキルを装って配布された事例では、キーチェーン、19種類のブラウザー、150種類の暗号資産ウォレット、Telegramのメッセージなどが窃取の対象となった[100]。

OpenClawは2026年2月からVirusTotalと連携し、悪意のあるスキルと判定されたもののダウンロードをブロックした。自然言語のプロンプトインジェクションは検知をすり抜ける可能性があり、ハッシュやシグネチャによる検査に加え、スキルの指示や挙動を検証する必要がある[101]。

97 https://www.koi.ai/blog/clawhavoc-341-malicious-clawedbot-skills-found-by-the-bot-they-were-targeting
98 https://blog.virustotal.com/2026/02/from-automation-to-infection-how.html
99 https://www.bitdefender.com/en-us/blog/businessinsights/technical-advisory-openclaw-exploitation-enterprise-networks
100 https://www.trendmicro.com/en_us/research/26/b/openclaw-skills-used-to-distribute-atomic-macos-stealer.html
101 https://openclaw.ai/blog/virustotal-partnership

4) AI開発・ビルド環境へのサプライチェーン攻撃

AI関連の開発ツール、配布パイプライン、フレームワーク、モデルファイルがサプライチェーン攻撃の標的となっている。

(1) AI開発ツール・CLIへのサプライチェーン攻撃

AIコーディングエージェントやCLIツールは開発者のコードや認証情報にアクセスし、自律的に処理できる。そのため、配布元の侵害は幅広い利用者に影響する。2025年8月、Nxの悪意のあるnpmバージョンが、Claude Code、Gemini CLI、Amazon Qなど開発環境内のAIツールを悪用し、2,349件のシークレットが漏えいした[102]。

2026年2月、窃取されたnpm公開用トークンを使ってCline CLIの不正バージョン2.3.0が配布された。このバージョンは利用者の同意なくOpenClawをインストールし、約8時間で約4,000回ダウンロードされた。Clineによれば、それ以外の不正な挙動はなく、VS Code拡張とJetBrainsプラグインにも影響はなかった[103]。

同年3月のgemini-ai-checkerはGeminiの認証情報を確認するツールを装っていた。研究者は、北朝鮮と関連するContagious InterviewキャンペーンのOtterCookieの変種と、中～高程度の確度で評価した。AIコーディングツールのトークン、キー、会話履歴が標的となった[104]。同月に報告されたTroyDen's Lure Factoryでは、OpenClawのDocker配布ツールやゲーム関連ツールなどを装う300以上のパッケージが確認された。

[102] https://github.com/nrwl/nx/security/advisories/GHSA-cxm3-wv7p-598c

[103] https://www.endorlabs.com/learn/supply-chain-attack-targeting-cline-installs-openclaw

[104] https://cyberandramen.net/2026/04/04/ottercookie-expands-targeting-to-ai-coding-tools-analysis-of-a-trojanized-npm-campaign/

TroyDen's Lure Factoryのパッケージは、画面の取得、位置情報の確認、データ送信などに使われるトロイの木馬を配布した[105]。

(2) AIフレームワーク・ライブラリへの悪意あるパッケージの混入

インストール時にコードを実行するパッケージは、認証情報が集まる開発端末やCI環境に大きな影響を及ぼす。2026年3月、累計ダウンロード4億8,000万回以上のLiteLLMのCIが、侵害された外部スキャナーTrivyをバージョンの固定やチェックサムの検証なしに取り込んだ。PyPI公開用トークンが窃取され、SSH鍵、クラウドやKubernetesの認証情報を狙うバックドア入りのバージョンが配布された[106]。LiteLLMは非公開リポジトリとTrusted Publisherを使う配布方式に移行したが、同年10月21日にリリース工程が再び侵害された[107]。

2026年4月、無料のOpenAI互換プロキシを装うhermes-pxが見つかった。実際にはチュニジアの私立大学Université Centraleの内部チャットボットを無断利用していた[108]。5月のTeamPCP Shai-Huludでは、TanStackとMistral AIから、Guardrails AI、UiPath、OpenSearchなど数百のパッケージに被害が及んだ。TanStackの42パッケージに含まれる不正な84バージョンには、有効な署名とビルド来歴証明（provenance）が付いていた[109]。6月のHadesはPyPIの19パッケージに影響し、AnthropicのAPIキー、クラウドのシークレット、ClaudeやMCPの設定が窃取された[110]。同月、Mastraの保守担当者のアカウントも侵害された。

[105] https://threatlabs.netskope.com/blog/2026/03/openclaw-trap-ai-assisted-lure-factory-targets-developers-gamers
[106] https://research.jfrog.com/post/litellm-compromised-teampcp/
[107] https://docs.litellm.ai/blog/security-incident-oct-21-2026
[108] https://research.jfrog.com/post/hermes-px-pypi/
[109] https://www.endorlabs.com/learn/shai-hulud-compromises-the-tanstack-ecosystem-80-packages-compromised
[110] https://socket.dev/blog/shai-hulud-descends-to-hades-miasma-pypi-wave

Mastraでは約30分で対象パッケージが再配布され、改ざんされた依存関係を通じて外部のコードが実行された。月間ダウンロードは合計2,800万回以上で、通常付いていた来歴証明が欠けていた[111]。2026年10月8日に公開されたTensorlakeのnpm SDK（0.5.144、週約1万2,000ダウンロード）には、認証情報を窃取するワームが含まれていた。正規リリースパイプラインで処理されたため、有効な来歴証明が付いていた[112]。有効な署名や来歴証明があっても、ソースやビルドが既に侵害されている場合がある。

(3) バックドアを含むAIモデル

公開ハブにあるモデルでも、ファイル形式によっては読み込み時にコードが実行される。2024年2月、Hugging FaceでPyTorchモデルを装う悪意あるファイルが発見された[113]。2025年2月にReversingLabsが報告した別の2ファイルは、Hugging Faceの検査を通過していた。同社は24時間以内にモデルを削除し、スキャナーを修正した[114]。同年12月にはPicklescanについて三つの脆弱性が公表された。

111 https://www.endorlabs.com/learn/mastra-npm-org-compromised-multiple-packages-trojanized-to-drop-a-remote-payload-via-easy-day-js

112 https://www.endorlabs.com/learn/tensorlake-npm-package-compromised-by-shai-hulud-in-latest-software-supply-chain-attack

113 https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/

114 https://www.reversinglabs.com/blog/rl-identifies-malware-ml-model-hosted-on-hugging-face

Picklescanの脆弱性はCVE-2025-10155、CVE-2025-10156、CVE-2025-10157として識別され、検査時と実際の読み込み時の挙動の差が問題となった[115]。MLflowでも、モデル成果物の読み込み時にコードが実行される問題が2026年9月8日にCVE-2026-79721として登録された。関連の識別子はCVE-2026-96775とCVE-2026-96804である[116]。公開ハブから入手したモデルの取り扱いもサプライチェーン管理に含める必要がある。

5) AI関連情報の窃取

窃取の対象には、企業秘密や個人情報を含む会話・プロンプトと、無断利用や連携先へのアクセスを可能にするAPIキー・認証情報がある。

(1) 会話とプロンプトの収集

無料VPN拡張機能Urban VPN Proxyは、2025年7月9日配布のバージョン5.5.0から、ChatGPT、Claude、Gemini、Copilotなどとの会話を収集していたと報告された。インストールは600万件に達し、同じ公開者の拡張機能を合わせると、ChromeおよびEdgeの利用者800万人以上に影響が及んだ[117]。別の二つの拡張機能「Chat GPT for Chrome with GPT-5, Claude Sonnet & DeepSeek AI」と「AI Sidebar with Deepseek, ChatGPT, Claude and more」は、匿名の解析データを収集すると説明していたが、実際にはChatGPTとDeepSeekの会話、閲覧URL、検索語を収集し、30分ごとに外部へ送信していた[118]。

[115] JFrog — Picklescanの脆弱性報告
[116] https://www.sentinelone.com/vulnerability-database/cve-2026-79721/
[117] https://thehackernews.com/2025/12/featured-chrome-browser-extension.html
[118] https://www.ox.security/blog/malicious-chrome-extensions-steal-chatgpt-deepseek-conversations/

これらの拡張機能は2万社以上で見つかった。利用者が収集を無効にしても更新後に再び有効になることがあり、一部のエージェント型ブラウザーでは利用者の承認なしに導入された[119]。

2026年1月に確認されたMicrosoft Copilot PersonalのRepromptでは、正規ドメインのリンクを開くと、ログイン中の利用者の名前、所在地、直近の会話の要約などが漏えいするおそれがあった。後続の処理が外部から制御されるため、最初の入力だけを見て被害を把握するのは難しかった[120]。同年6月にはMicrosoft 365 Copilot Enterprise SearchのSearchLeak（CVE-2026-42824）が公表され、メール、予定、SharePoint、OneDriveの情報が外部に送信される可能性が示された。正規ドメイン上のURLであったため、URLだけを判定する対策では見つけにくかった[121]。業務用アシスタントは利用者の権限で広い範囲にアクセスするため、同種の問題でも個人用より影響が大きくなり得る。

(2) APIキーと認証情報の窃取

AIサービスのAPIキーには、推論、利用枠、課金に関わる権限がある。漏えいしたキーは、大量利用や不正請求などの原因となる。2026年2月、公開GitHubリポジトリ5,000件以上、稼働中のウェブサイト約3,000件で、ソースコード、設定ファイル、配信されるフロントエンドファイルなどにChatGPTのAPIキーが埋め込まれていることが確認された[122]。

119 https://www.microsoft.com/en-us/security/blog/2026/03/05/malicious-ai-assistant-extensions-harvest-llm-chat-histories/
120 Varonis — Repromptの報告
121 Varonis — SearchLeakの報告
122 https://cyble.com/blog/when-ai-secrets-go-public-chatgpt/

公開リポジトリなどを自動で監視するスキャナーがあるため、キーの露出から悪用までが数時間、場合によっては数分に縮まる。漏えいしたキーは、大量の推論リクエスト、フィッシングや詐欺の文面作成、課金アカウントの利用枠の消費に使われた。AI APIの利用が集中管理ログやSIEMに記録されず、料金の急増で初めて発覚することもある。

MCP設定ファイルからは2万4,008件の秘密情報が見つかった。Claude Codeが共同作成したコミットで秘密情報が漏えいする率は、全体平均の約2倍だった[123]。GitHub以外でも同様の問題が確認された。2026年10月、中国の大手コードホスティングサービスGiteeの公開リポジトリ725万件を調査したところ、有効なクラウド・AI認証情報が2万6,400件見つかった。AIサービスのキーはDeepSeekが3,170件、Zhipu BigModelが1,489件、OpenAIが1,348件だった。大手企業向けAIプラットフォームの組織管理者の認証情報も含まれていた。

2026年10月、AIコーディングエージェントOpenCodeのホスティングサービスConsoleが侵害され、外部のAIサービスを連携させていた利用者に影響が及んだ[124]。Consoleに保存された事業者のAPIキーが漏えいした可能性があり、身に覚えのない請求が発生した利用者もいた。この事案には、別のモデルの応答を大量に収集し、それを使って独自のモデルを学習させる大規模な蒸留攻撃が関係していたとみられる。

セキュリティ侵害がなくても認証情報が公開されることはある。ChatGPTの会話共有で検索を許可すると、Googleなどの検索エンジンに登録され、ソースコード、個人情報、コードに記載されたパスワードなどを第三者が検索できる状態になった[125]。OpenAIは2025年7月末から8月初めにかけて、このオプションを削除し、登録済みのリンクが検索結果から削除されるよう対応した。

8. 対策

AI関連の脅威には、AIへの信頼を利用する攻撃、AIを活用した攻撃、AIシステムを狙う攻撃がある。影響を受けるのは、個人利用者、AIを導入する組織、AIサービスやモデルの開発企業である。

123 https://gitguardian.com/state-of-secrets-sprawl-report-2026

124 https://x.com/opencode/status/2111519637049319818

125 https://www.bitdefender.com/en-us/blog/hotforsecurity/your-shared-chatgpt-chats-may-be-publicly-searchable-heres-how-to-delete-them

共通する原則は、攻撃が成立した場合でも被害を抑えられるように備えることである。例えば、LLMはデータと指示を確実に区別できないため、プロンプトインジェクションを完全に防ぐことは難しい。モデルが誤った判断をした場合でも、実行できる操作が限られるように設計する必要がある。

AIの利用者は、サービス、拡張機能、APIを公式の配布・提供経路から利用する。AIの回答や共有された会話に、コマンドを実行画面やターミナルに貼り付けるよう書かれていても従わない。音声や映像で送金を求められた場合は、顔や声だけで判断せず、依頼を別途確認する。家族間で合言葉を決めておく方法もある。エージェントに任せる作業は範囲を絞り、認証が不要なものはログアウトした状態で実行させる。業務用アカウントへのログインにはAIブラウザーを利用せず、実行前の確認画面だけに頼らない。

AIを導入する組織は、製品ごとに個々のAI機能を確認して利用の可否を決める。例えば、ウェブページの要約を許可しながら、エージェントによるメール作成を禁止し、AIブラウザーのデータ保存を無効にできる。既存のデータ損失防止（DLP）ポリシーをブラウザー内のAI機能にも適用できる。合成音声・映像や偽のAIサービスのページに備え、業務手順も見直す必要がある。

AIサービスの開発・運用者は、エージェントの権限を最小限に抑える。2026年8月発表の「OWASP Top 10 for LLM Applications 2026」では、過剰なエージェンシー（Excessive Agency）が6位から3位に上がった[126]。エージェントが越えられない制限を設け、既定で有効にする。利用者からの入力、ツール呼び出し、ツール応答、最終出力の四つをすべて検査する。ツール応答を検査しなければ、文書やウェブページの悪意ある指示が信頼できる内容として扱われ得る。ウェブコンテンツを読むモデルと、操作の妥当性を判定するモデルを分け、決済やメッセージ送信前に利用者の確認を求める設計も参考になる[127]。MCPサーバー、スキル、ライブラリ、モデルファイルの出所を確認する。新バージョンの導入まで待機期間を設ける。npmでは `npm config set min-release-age 7` で7日間の待機を適用できる[128]。

[126] https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/
[127] https://blog.google/security/architecting-security-for-agentic/
[128] https://www.elastic.co/security-labs/threat-command/how-we-caught-the-axios-supply-chain-attack

AIモデルの開発者は、攻撃への耐性を高めるため、モデルの学習を継続する。自動化された検証で新たな攻撃を見つけ、得られた事例を使ってモデルを再学習させる方法が代表的である[129]。再学習によって攻撃の成功率を下げられるが、サービス側でも権限を制限する必要がある。悪用事例の検知と公表もモデルの開発者が担う。2025年12月から2026年8月までの間には、窃取したAPIキーで他者のアカウントが利用された事例が確認された[130]。アカウントの停止に加え、キーの異常な利用を検知し、所有者に通知する必要がある。

9. 今後の見通し

AIエージェントはメール、予定、金融情報にアクセスし、実際に操作を行う。一つの設定ミスやプロンプトインジェクションが、情報漏えい、権限昇格、内部ネットワークへのアクセスにつながり得る。Clineをめぐる事例では、GitHubの課題管理機能を介したプロンプトインジェクションにより、数千台のシステムにOpenClawが同意なくインストールされたと報告されている[131]。原文では、この節の事例に2026年1月、開発ツールの節の不正バージョンの配布に同年2月が記されている。

侵入後の活動にもAIが悪用されている。2026年2月、ある攻撃者が商用生成AIを悪用し、600台以上のFortiGate機器を侵害した。侵入後にもAIを利用したと報告された[132]。組織内のエージェントはラテラルムーブメントの経路となり得る[133]。今後、AIの応答に応じて動作するマルウェアや[134]、IoT機器で小型モデルが自己複製するAIワームが現れる可能性がある[135]。いずれも現時点では概念実証または管理された実験の段階にある。

セキュリティでは、何を防ぐかに加え、エージェントにどこまで任せるかが重要になっている。

129 https://openai.com/index/hardening-atlas-against-prompt-injection
130 https://www-cdn.anthropic.com/e50be2e51e7695dc4b1366a37a245a597377d3b5/Anthropic-Detecting-and-countering-091026.pdf
131 https://krebsonsecurity.com/2026/03/how-ai-assistants-are-moving-the-security-goalposts/
132 https://aws.amazon.com/blogs/security/ai-augmented-threat-actor-accesses-fortigate-devices-at-scale
133 https://orca.security/resources/blog/ai-induced-lateral-movement-ailm/
134 Check Point Research ― AIを悪用したC2の分析
135 https://www.anthropic.com/research/self-replicating-ai

プロンプトで指示を与えるだけでは、エージェントの動作を完全に制御できない。アクセス可能なデータと自動実行できる操作の範囲を厳格に制限し、実行段階で検証を強制する必要がある。MCP、Skills、A2Aなどの標準が急速に普及するにつれ、攻撃対象領域も広がる。AIセキュリティは、変化するエコシステムを継続的に追跡する長期的な課題である。

10. 結論

生成AIは外部ツールと連携して自ら動くエージェントへと発展し、攻撃者はAIに攻撃の実行も担わせるようになった。同時に、AIサービス、エージェント、MCPサーバー、スキル、モデルのサプライチェーンも新たな攻撃対象となった。

システムやネットワークを保護するとともに、AIがどのデータを信頼し、どの権限で何を実行できるかを制御する必要がある。AIを導入する組織は、エージェントへの最小権限の付与、サプライチェーンの検証、プロンプトインジェクション対策、AI関連情報の保護を基本要件とすべきである。本報告書は、作成時点で確認した事例を整理したものである。発行後も新たな手法や脆弱性が現れるため、継続して動向を追う必要がある。

11. 参考文献

［1］Inside the LLM | Understanding AI & the Mechanics of Modern Attacks
https://www.sentinelone.com/labs/inside-the-llm-understanding-ai-the-mechanics-of-modern-attacks/

［2］The Growing Threat of ChatGPT-Based Phishing Attacks
https://cyble.com/blog/the-growing-threat-of-chatgpt-based-phishing-attacks/

［3］Criminal AI-as-a-Service in 2026: How the Underground Market Is Operationalizing Cybercrime
https://www.rapid7.com/blog/post/tr-criminal-ai-underground-market-operationalizing-cybercrime-2026/

［4］Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign
https://www-cdn.anthropic.com/d7dd50dd1185f59be051b307150d877f2b82bd2c.pdf

［5］Ignore All Instructions and Read This Blog: The State of AI-Analysis Evasion in Malware
https://blog.talosintelligence.com/ignore-all-instructions-and-read-this-blog-the-state-of-ai-analysis-evasion-in-malware/

［6］プロンプトインジェクションを理解する：最前線のセキュリティ課題

https://openai.com/ko-KR/index/prompt-injections/

［7］How AI Assistants Are Moving the Security Goalposts

https://krebsonsecurity.com/2026/03/how-ai-assistants-are-moving-the-security-goalposts/

［8］Self-Replicating AI

https://www.anthropic.com/research/self-replicating-ai

本報告書は著作権法によって保護されています。営利目的での無断転載・無断複製を禁じます。
本報告書の全部または一部を引用・加工する場合は、アンラボが発行した報告書であることを明記してください。
掲載内容または配布に関するお問い合わせは、アンラボ（+81 3-6453-8315）までお願いします。
本報告書は https://atip.ahnlab.com からご利用いただけます。

(株)アンラボ
〒108-0014 東京都港区芝4丁目13-2 田町フロントビル3階ホームページ：
www.ahnlab.com/jp
Tel： +81 3-6453-8315 ｜ Fax： +81 3-6453-8316
© 2026 AhnLab, Inc. All rights reserved.

AhnLab