レポート

AI セキュリティ 2026 年版

世界中の 62 万 5,000 社を超える組織の顧客基盤全体における観測結果

SOPHOS

SOPHOS | X-OPS

John Peterson からのメッセージ

2026 年 5 月、ソフォスは、顧客のネットワーク内で、実質的にソフトウェア開発業務と呼べる活動を行っている攻撃者を発見しました。約 12 個の AI エージェントが、商用のコーディングアシスタントを介して連携し、それぞれ専用の仮想マシン上で、ソフォス、CrowdStrike、Windows Defender のエンドポイント保護製品を回避するマルウェアの作成とテストを行っていました。この活動によって、約 80 個のモジュールと 70 件を超える回避手法が生み出され、すべての成果物がバージョン管理システムにコミットされたうえで、次の試行で改良されていました。同じオペレーターはその後、こうして生成した成果物を利用してランサムウェアを展開し、データを窃取しました。

攻撃手法そのものは新しいものではありませんでしたが、AI エージェントによって、数週間にわたって手作業で行っていた反復作業が数日間の自動化された反復作業へと短縮されていました。本レポートでは、このような変化について取り上げます。AI は、侵入の基本的な仕組みを変える以上に、セキュリティ運用のスピードと形態を変えています。攻撃者には依然として初期アクセスが必要であり、ラテラルムーブメントを行い、データを外部へ持ち出していますが、その経路は観測可能です。変わったのは、攻撃にかかる時間です。

本レポートでは、データを示しながらこの点を明らかにします。

ソフォスでは、MDR (Managed Detection and Response) およびインシデントレスポンス (IR) のアナリストが、年間数千件のインシデントを調査しています。また、ソフォスには、複数の言語でアンダーグラウンドフォーラムを監視する Counter Threat Unit (CTU) があります。膨大なマルウェアサンプルを分析する SophosLabs のリサーチャーもいます。また、革新的な防御技術を開発する AI チームもあります。さらに、世界中の 62 万 5,000 社を超える顧客基盤から、エンドポイント、ファイアウォール、メール、クラウド、ネットワーク、アイデンティティに関するテレメトリデータを収集しています。

この規模、対象範囲、専門知識によって、急速に進化する AI とサイバーセキュリティが交差する領域を、他に類を見ない視点から捉えています。ソフォスは、この視点を生かして、業界が攻撃のスピードに後れを取らず、顧客が確実に保護され続けるよう支援しています。

J. P. Peterson, CTO, Sophos

エグゼクティブサマリーと主な調査結果

本レポートは、Sophos MDR およびインシデント対応の事例研究、SophosLabs の解析、Sophos CTU のインテリジェンス、そして世界中の 62 万 5,000 社を超える組織の顧客基盤におけるエンドポイントおよびネットワークの観測結果に基づいています。これらすべてに共通するパターンが 1 つあります。AI によって、攻撃者と防御者が既存のオペレーションをより高速に実行できるようになりました。ただし、まったく新しい種類の攻撃の確認例は限定的であり、存在しないわけではありません。そして、AI の能力が拡大するにつれて、その数は増加する可能性があります。

2026 年は転換点となっています。フロンティアモデルとオープンウェイトモデルは、エンジニアリング作業を委任できるほど実用的になり、一方で、推論コストの低下によって攻撃側と防御側の双方のコスト構造も変化しています。実際に問われているのは、もはや AI を利用するかどうかではありません。高コストなフロンティアシステムをどこで利用するのが適切なのか、より低コストのオープンウェイトモデルで十分なのはどこなのか、そしてリスクを拡大させることなく、どのように処理を振り分けるのかが重要となっています。

レポートの要点となる調査結果は以下のとおりです。

1. **AI は攻撃の進行を高速化していますが、新しい攻撃の種類を発明しているわけではありません。**ソフォスが STAC6994 として追跡しているサイバー攻撃者は、AI エージェントを利用し、人間の開発者では実現できないほどの速さで EDR 回避手法の改良を繰り返していました。その後、Sophos CTU のアナリストは、STAC6994 の背後にいる攻撃者がランサムウェアの展開やデータ窃取を行っていたことを確認しました。
2. **エンタープライズ AI サービスを取り巻く認証情報およびアイデンティティ層は現在、収集の標的となっています。**Salesloft/Drift のインシデントは、チャットボットの OAuth トークンが Salesforce 環境の侵害に直結することを示しました。このインシデントとは別に、Sophos CTU のアナリストは、サイバー攻撃の過程で AI へのプロンプトや生成データが副次的に取得される可能性があると、サイバー攻撃者がアンダーグラウンドフォーラムで主張していることを観測しました。
3. **アンダーグラウンドでは AI をインフラとして取り込みつつあります。**ソフォスは、サイバー攻撃者が Claude、ChatGPT、Grok の API キーを仲介販売している事例を特定して文書化しました。フォーラムには、AI プロンプトエンジニアリング、ジェイルブレイクの手法、マルウェア開発のワークフローに特化したチャンネルが登場しています。一部の攻撃者は、AI プロンプトエンジニアを雇用しています。一方で、AI によって自らの活動が変わることに、依然として懐疑的な攻撃者もいます。AI の導入曲線は、突然の変革というよりも、新しいツールが既存の犯罪活動のワークフローに徐々に組み込まれていく過程を示しています。
4. **サプライチェーン攻撃は、特に AI 開発インフラストラクチャを標的としています。**Nx Console VS Code 拡張機能の侵害では、認証情報を窃取するマルウェアが配信されました。SophosLabs のアナリストは、偽の Claude ダウンロードページから新種の RAT を配信する事例と、正規の ChatGPT 共有チャット URL を悪用して MacSync インフォスティーラーを配信する ClickFix キャンペーンを調査しました。
5. **AI を活用したソーシャルエンジニアリングは、実験段階から運用段階へと移行しています。**Sophos CTU は、音声ボット、ロマンス詐欺向けの AI 生成ペルソナ、合成プロフィール画像サービスを宣伝するアンダーグラウンド市場の広告を特定しています。AI がもたらした変化は、生成 AI が登場する前から有効だった同じ欺瞞の手法を、より大規模に、そして多言語で高品質に適用できるようになったことです。

6. **脆弱性の悪用までの時間は、パッチ適用までの時間よりも速くなっています。**CISA が 2026 年 6 月 10 日に発行した「Binding Operational Directive 26-04」では、リスクが最も高い脆弱性について、パッチ適用期限が 3 日間に短縮され義務付けられました。同指令では、脆弱性の開示から悪用までの期間を短縮させる要因の 1 つとして、AI が名指しされています。この指令が発行されたほぼ直後に、実際の環境で初めて試される事例が発生しました。CVE-2026-10520 (Ivanti Sentry) は、概念実証 (PoC) の公開から 24 時間以内に悪用 され、6 月 12 日には KEV カタログに追加されました。CVE-2026-42208 (LiteLLM) は 36 時間以内に悪用され、攻撃者は上流の LLM プロバイダーのキーを含むデータベースを標的にしました。

7. **攻撃に AI が利用されたことを証明するのは、依然として実務上困難です。**STAC6994 の事例が異例だったのは、ソフォスが調査できるデバイス上に開発フレームワークが残されていたことです。多くのインシデントでは、AI による支援はテレメトリデータでは確認できません。AI を利用した攻撃の実際の広がりは、ベンダーが直接的な証拠によって確認できる範囲を上回っている可能性があります。

8. **AI 推論モデルは、調査を加速させます。**AI 推論モデルは、既知の攻撃パターンに対するアナリストの調査を、数時間から数分に短縮します。これらは、数十億件のイベントに埋もれた単一の脅威を検知する振る舞い検知システムに取って代わるものではありません。極端なクラス不均衡による統計的制約、環境固有のベースラインの確立、攻撃者の適応の速さへの対応には、いずれも専門的なエンジニアリングが必要です。

第 1 部：
攻撃チェーンにおける AI

Sophos X-Ops による AI 脅威の分類

Sophos X-Ops は、AI 脅威を大きく 2 つのカテゴリに分類しています。AI の悪意ある利用 (防御する対象となる脅威) と、AI を標的とした悪意ある攻撃 (保護する対象となる AI) です。前者には、AI が生成したアーティファクト、AI によって強化されたランタイム機能、AI によってオーケストレーションされた侵入が含まれます。後者には、エージェントによって開始される侵害、AI ソフトウェアのなりすまし、LLM のポイズニング (汚染)、そしてモデルそのものへの攻撃が含まれます。この 2 つには、それぞれ異なる対応が求められます。悪意ある利用は生産性と規模の問題であり、AI を標的とした悪意ある攻撃は信頼、サプライチェーン、ガバナンスの問題です。この分類体系は、MITRE ATLAS、Microsoft のエージェントフェイルモード分類、MIT の AI リスクリポジトリ、NIST AI 100-2 など、既存のフレームワークを補完するものです。

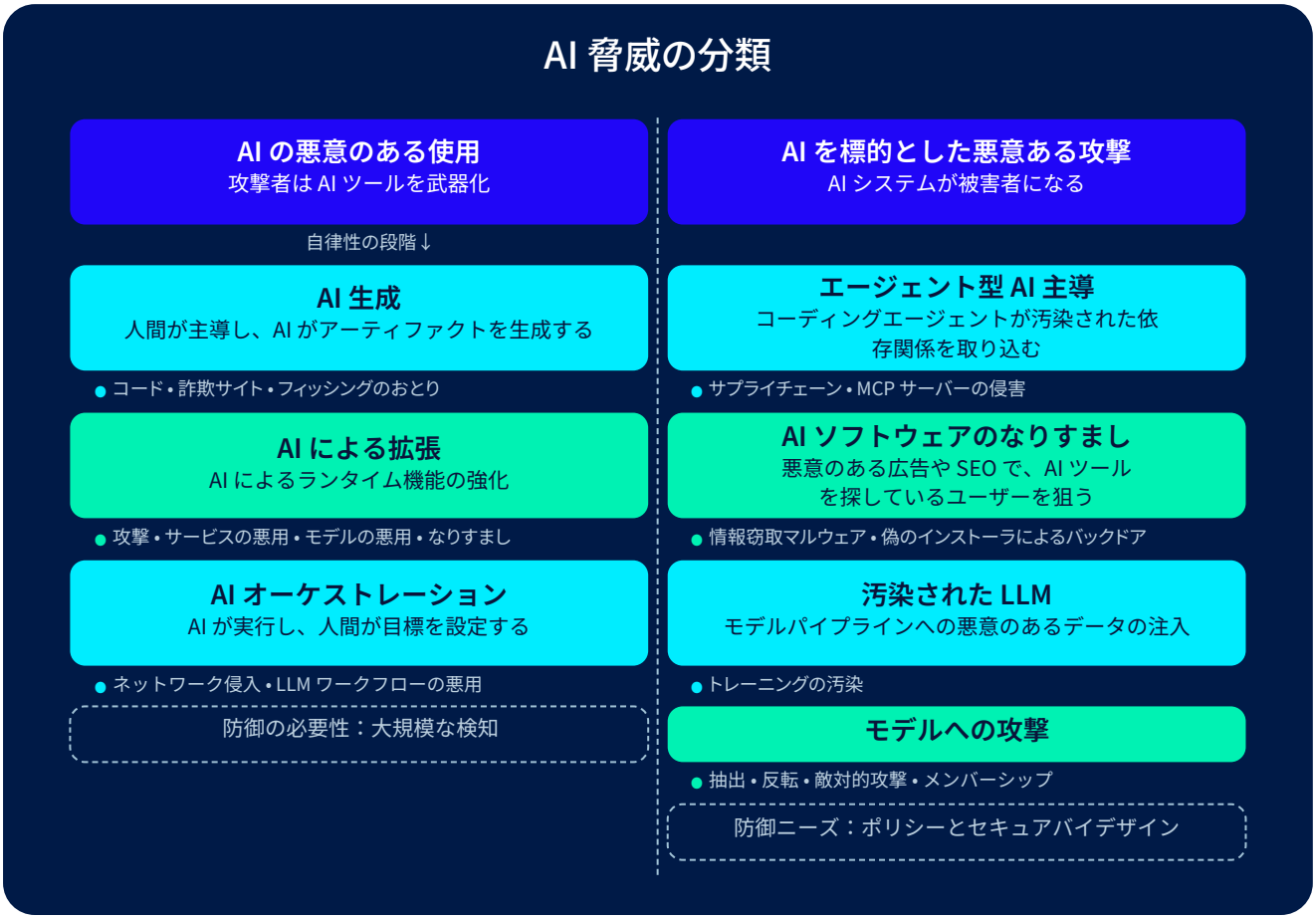


図 1：AI 脅威の分類の概要

自律性の段階

最も AI への依存度が低いケースでは、生成 AI を使って人間が攻撃用のアーティファクトを作成し、それを手動で展開します。メキシコ政府機関を標的とするサイバー攻撃者は、Claude Code と GPT を使用してスクリプトを生成しました。The Gentlemen ランサムウェアグループからチャットが流出したことで、同様のアプリケーションが確認されました。また、ソフォスのインシデント対応チームが調査した最近の DragonForce ランサムウェア事例では、サイバー攻撃者が標的組織との交渉中に、AI によって生成されたことが強く疑われるレポートを作成していました。このレポートには、窃取したデータについて、PII や法的リスクを含む詳細な分析が記載されていました。AI によって生成された分析は、標的組織に身代金を支払うよう圧力をかけ、説得するための手段として利用されました。

自律性の段階がさらに進むと、AI によって強化された攻撃では、モデルが実行時に攻撃能力を強化します。LameHug は、CERT-UA が中程度の確度で IRON TWILIGHT (APT28) によるものと評価している Python ベースのマルウェアで、ホスト型のオープンウェイトモデルに問い合わせて、環境ごとに Windows の偵察コマンドを動的に生成していました。その結果、静的解析の有効性が大幅に低下し、AI/ML API エンドポイントへの外部トラフィックが、数少ない信頼性の高い検知ポイントの 1 つとなりました。AI によって強化されたなりすましも、現在進行中の攻撃手法の 1 つです。音声クローンや顔交換ツールは、リアルタイム詐欺、経営幹部へのなりすまし、KYC (本人確認) の回避に利用されています。

自律性の最も高い段階では、Anthropic が 2025 年 11 月に公表した事例で、中国政府系の攻撃者が Claude Code を利用し、約 30 の標的への侵入を試みたことが明らかになりました。こうした攻撃は依然としてまれですが、偵察から実行に移るまでの時間を大幅に短縮し、従来の防御における前提を揺るがしています。

ケーススタディ：STAC6994

ソフォスのアナリストは、サイバー攻撃者が「レッドチーム」型の侵害後フレームワークで AI を利用して EDR を回避する手法をテストしていることを確認しました。ソフォスは、顧客の環境でペイロードを継続的に生成している不正なデバイスを発見しています。攻撃者は Ludus から仮想マシンをプロビジョニングし、AI エージェントを搭載した Cursor IDE を使用して、侵害後のツールを開発およびテストしていました。

このフレームワークは、3 つの環境で並行テストを実行していました。1 台はソフォスのエンドポイントプロテクションを導入した VM、1 台は CrowdStrike を導入した VM、そして 1 台は比較対象として EDR を導入していない VM でした。4 台目の VM は Sliver C2 サーバーとして動作していました。約 12 個の AI エージェントが、それぞれ明確な役割を担って動作していました。

Claude Opus 4.5 を実行する 1 つのエージェントが、コアの操作とルール設定を担当していました。その他のエージェントは、運用上のセキュリティ (OPSEC) の強化、ドキュメント作成、プロキシのストレステスト、VM の展開、EDR エージェントに対するツールのテストを担当していました。コードのコミットは、Model Context Protocol (MCP) を介して Git に送られていました。

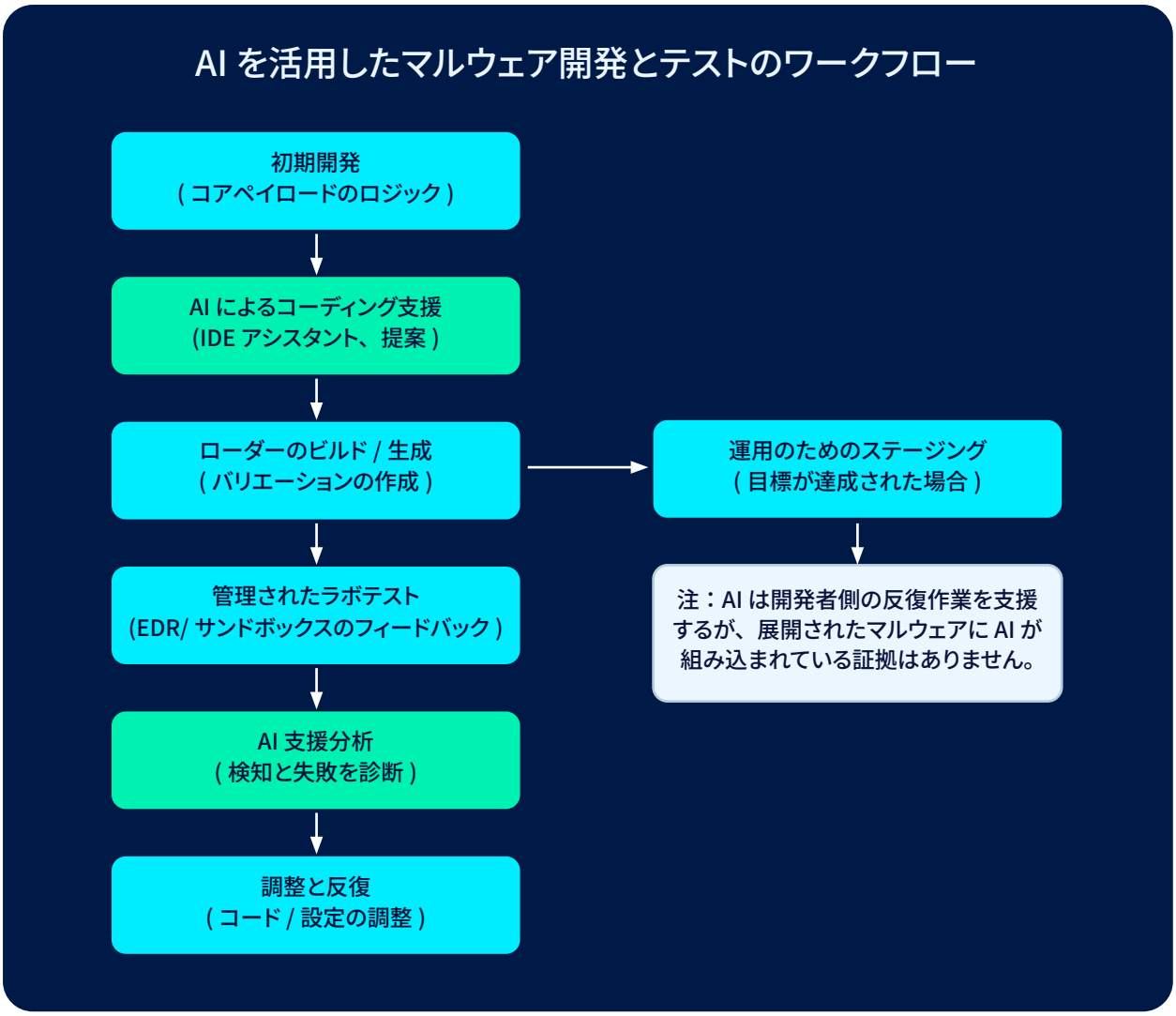


図 2：マルウェア開発ワークフローにおける AI の役割を示した図

これらの手順は体系化されていました。復元されたアーティファクトによると、エージェントは SpecterOps のブログからスクレイピングされた調査記事を読み込み、攻撃手法を抽出し、それらを MITRE ATT&CK フレームワークにマッピングしたうえで、各手法を再現するために必要な手順とツールを特定し、ラボ環境を準備して実行し、結果を報告していました。

中核には、Python ベースのモジュール型ペイロード生成ツールがあり、暗号化と回避手法を何層にもわたって元のペイロードに適用し、Rust および Go でカスタム実行ファイルを生成していました。70 種類を超えるさまざまな手法をテストする、約 80 種類のモジュールが開発されました。実際の EDR 回避プロセスは、人間によるレビューと反復を組み込んだ、体系的なエンジニアリングテストサイクルでした。AI はワークフローを調整し、実験を支援していました。一方、展開されたマルウェアに AI が組み込まれていたことを示す証拠は確認されませんでした。

さまざまな反復を経て、エージェントのドキュメントでは、EDR エージェントに対してほぼ全面的に成功したと主張されていました。しかし、ソフォスのアナリストが調査した結果、その結論を十分に裏付ける証拠はありませんでした。運用上の重要なシグナルは、そのスピードです。完全なサイクルに要していた期間は、数週間から数日へと短縮されました。

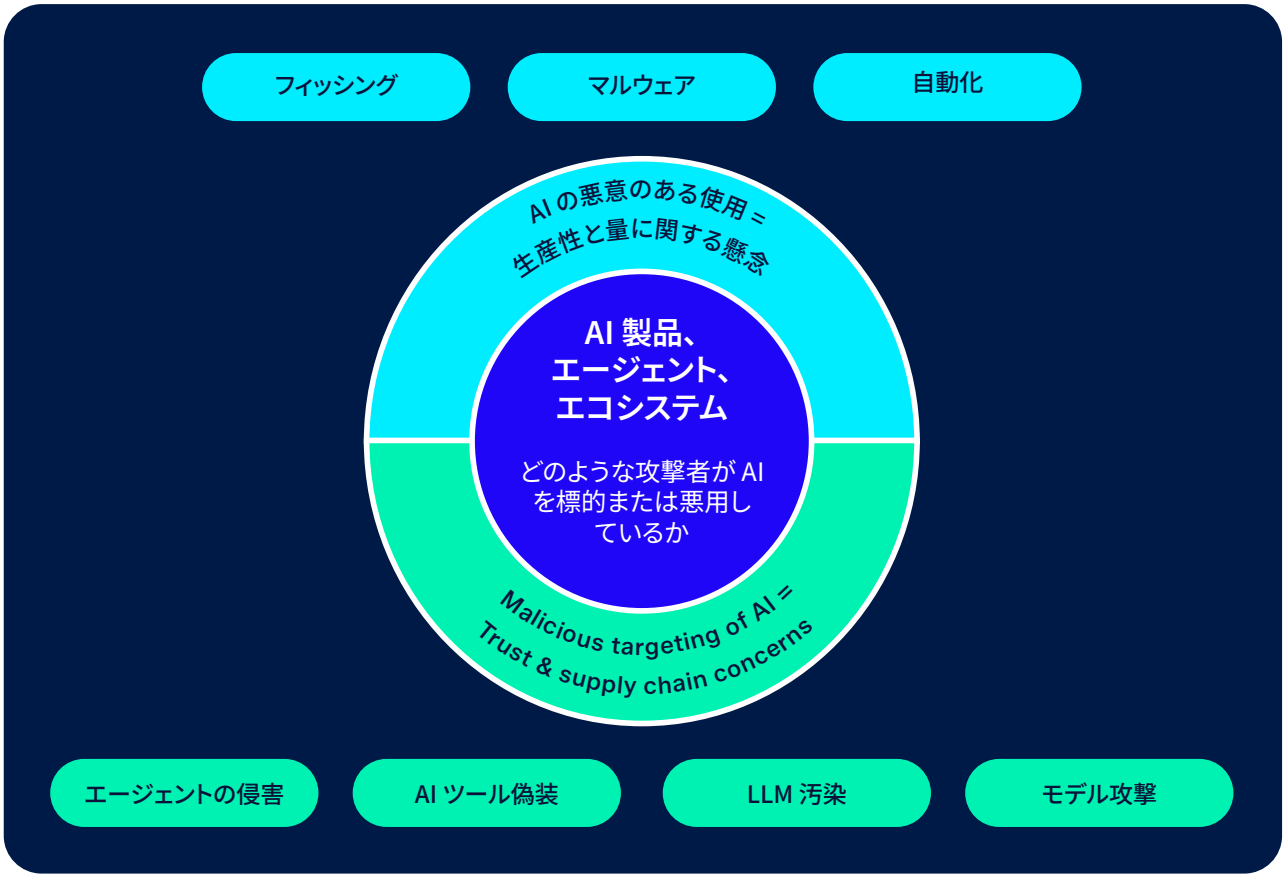
CTU はその後、STAC6994 の攻撃者がランサムウェアの展開とデータ窃取に関与したことを確認しました。このフレームワークは、実際の侵入に使用されるツールだったのです。

AI を標的とした悪意ある攻撃

この分類体系の第 2 のカテゴリでは、AI を攻撃のツールとして利用するのではなく、AI 製品、AI エージェント、AI エコシステムそのものを標的または攻撃対象とする攻撃を扱います。これについては、第 2 部で詳しく取り上げます。エージェントによって開始される侵害は、最も差し迫った懸念があります。たとえば、コーディングエージェントがタスクを実行する過程で、侵害された NPM パッケージをダウンロードしたり、汚染された MCP サーバーとやり取りしたりするケースです。ここで懸念されるのは、公開されてから実行されるまでの時間が短縮されることです。しかも、AI が開発と改良を担うことで、その過程を人間がレビューするという従来のブレーキも効かなくなっています。

この分類体系には、AI ソフトウェアのなりすまし (サイバー攻撃者が、正規の AI ツールを探しているユーザーを狙う攻撃)、LLM の汚染 (カスタムモデルを構成したり、悪意のある、または誤解を招くコンテンツを意図的に注入して、モデルの動作を変化させたりする攻撃)、さらに、モデル抽出、モデルから学習データを推測して復元する攻撃、敵対的サンプル、メンバーシップ推論など、モデルそのものを対象とするさまざまな攻撃が含まれます。

この 2 つの大分類を分けて捉えることで、脅威の全体像がより明確になります。AI の悪意ある利用は、本質的には生産性と量の問題です。既存の検知基盤で多くの攻撃を検知できる可能性はありますが、問題となるのは規模と反復のスピードです。AI を標的とした悪意ある攻撃は、信頼性とサプライチェーンに関わる問題であり、ポリシーやセキュアバイデザイン・フレームワークなどの取り組みによって対処するのがより適切です。



アンダーグラウンドにおける AI の導入と懐疑論

ソフォスは 2023 年以降、犯罪者フォーラムにおける生成 AI に対する見方を追跡してきました。2025 年の追跡調査を経て、現在も継続的に監視しています。状況は大きく変化しています。サイバー攻撃者は ChatGPT、Claude、Grok の仲介 API キーを購入しており、フォーラムには AI 専用のチャンネルも設けられ、ペルソナ同士がプロンプトのテンプレートやジェイルブレイクの手法を共有しています。Sophos CTU のリサーチャーは、以前からブロックチェーン開発者やフィッシング用のコールセンタースタッフを募集していた、ある既知のアンダーグラウンドのリクルーターが、今回は AI プロンプトエンジニアを専門に募集していたことを確認しました。これは、サイバー攻撃者がマルウェア開発や初期アクセスを外部委託するのと同じように、AI に関する専門知識も外部に委託していることを示しています。

CTU はまた、ビッシングや電話を利用した詐欺向けに設計された AI 音声ボットの広告も確認しました。ユーザー (サイバー攻撃者) からの肯定的なレビューも寄せられており、これらのボットが声のトーンや話すテンポ、会話パターンを説得力のある形で模倣できることが示されています。「HackingRealm」という人物は、ロマンス詐欺やソーシャルエンジニアリング向けの「AI OnlyFans Models」サービスを宣伝していました。このサービスは、合成されたプロフィール画像や会話コンテンツを生成し、出会い系サイトやソーシャルメディア上で、大規模に運用可能なペルソナを作成するものでした。さらに、専用の Web サイトや、肯定的なレビューを並べたフィードバックページも用意されていました。

「AI 主導」をうたうツールも登場しています。Leak Bazaar (機械学習を活用した窃取データのトリアージ) や、REST API および MCP サーバー統合に対応した Cobalt Strike の更新版などです。攻撃者は、従来からあるワークフローに LLM の統合機能を組み込んでいます。Cobalt Strike の広告で Claude に言及しているのは、最新技術を取り入れていることを示すアピールの一面がありますが、同時に、LLM の統合が犯罪市場で差別化要因になりつつあることも示しています。たとえそれが攻撃手法を洗練させるものではなく、単に利便性を高めるだけであってもです。

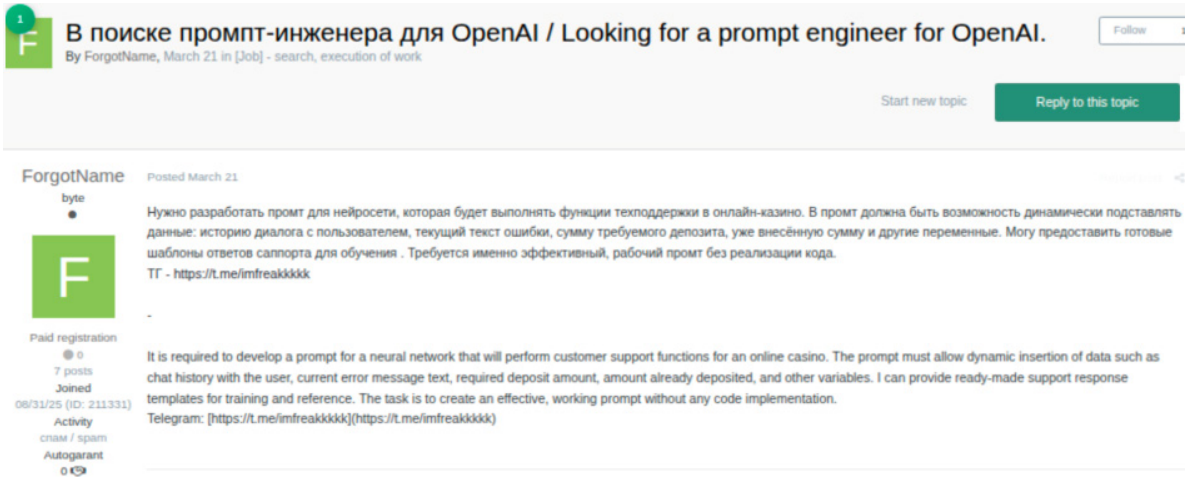


図 3：OpenAI プロンプトエンジニアの募集

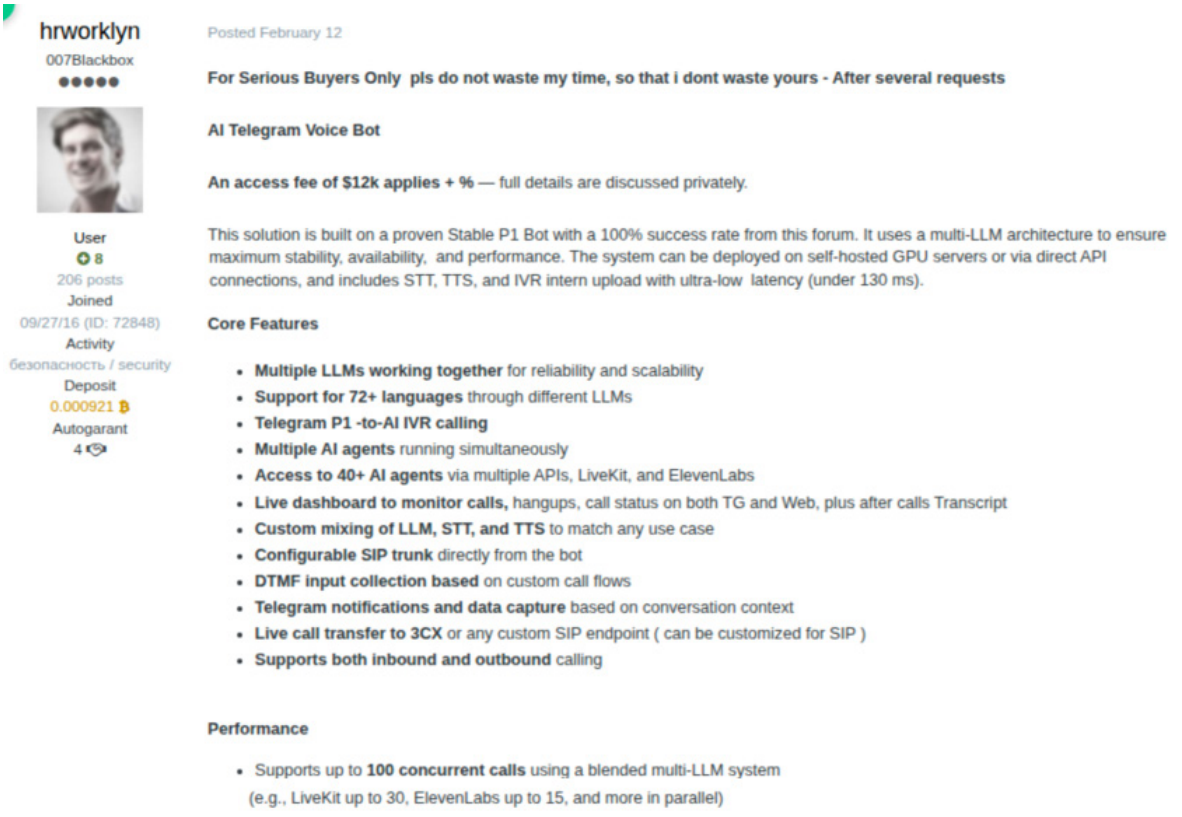


図 4：AI Telegram 音声ボットの広告

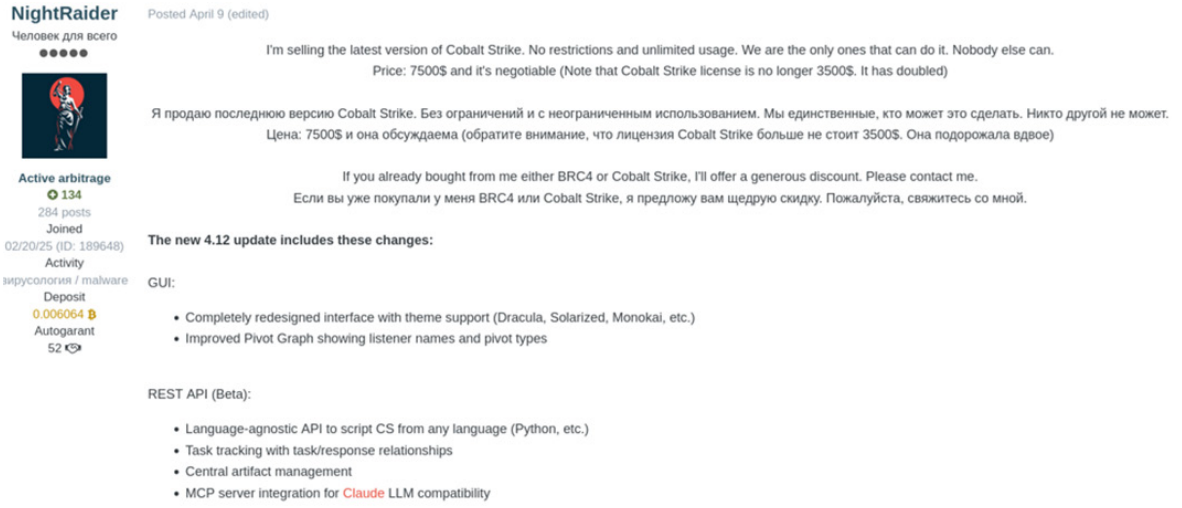


図 5：Cobalt Strike のアップデートを宣伝するサイバー犯罪フォーラムのユーザー

しかし、活発な実験や導入が行われていることを示す明確な証拠がある一方で、少なくとも CTU が調査したフォーラムでは、コミュニティ全体が一様に AI を歓迎しているわけではありません。ソフォスが 2023 年と 2025 年に犯罪者向けフォーラムで生成 AI に対する見方を調査した際の過去の調査結果と同様に、CTU は、AI によって仕事の機会が減少することへの懸念を示す人物を確認しました。特に、マルウェア開発やスクリプト作成の仕事が減ることを懸念していました。一方で、AI を完全に否定し、人間のスキルに頼ることを推奨する者もいました。

AI で強化されたソーシャルエンジニアリングとディープフェイク

AI によって、ソーシャルエンジニアリングキャンペーンの展開速度、多言語での文章の自然さ、そして制作コストが大きく変化しています。たとえば、Mandiant の「M-Trends 2026」では、画一的なフィッシングから、LLM を活用して標的に合わせたメッセージを作成し、関係を構築するソーシャルエンジニアリングへの移行が報告されています。ボイスフィッシングは初期感染経路として 11% を占め、2 番目に多い手法へと増加した一方、メールフィッシングはわずか 6% に低下しました。ConnectWise の「MSP 脅威レポート 2026 年版」では、ディープフェイクに対応した詐欺や LLM によって生成されたフィッシングキャンペーンが積極的に使用されていることが確認されています。AI は、新しい攻撃カテゴリを作成するのではなく、確立された戦術をより迅速にし、拡張性、説得力のあるものにしたと指摘しています。

ディープフェイクは、さまざまなタイプのサイバー攻撃者が実運用で利用するツールになっています。特に注目されるのが北朝鮮のサイバー攻撃者で、AI 生成の身元情報やディープフェイクを利用して欧米企業のリモート勤務職に就き、継続的な内部アクセスを確立しています。ソフォスは、北朝鮮の IT ワーカーによる手口を中心に解説した CISO 向けプレイブックを公開しており、検知の指標と組織として講じるべき対策をまとめています。

Sophos CTU は、AI をテーマにしたソーシャルエンジニアリングを用いた豚の屠殺 (Sha Zhu Pan) 詐欺を調査しました。英国在住の被害者は、LINE 上で数か月にわたって AI をテーマにした「レッスン」を受けるうちに、DAIQ Wealth という偽の AI 投資プラットフォームに誘い込まれました。複数の人物がグループチャットを通じて被害者との信頼関係を築いており、それぞれを別の人物が運用していたことから、あらかじめ定められたスクリプトに沿って活動する組織的な人員体制が存在していたことが示唆されます。この被害者は、数十万ポンドを失いました。被害者が自宅に約 2 万ポンドの現金を保管していることへの不安を伝えると、詐欺師は 24 時間以内に英国の住所まで現金を運ぶ回収役を手配しました。しかも、正規の金融機関のブランドを無断で使用した領収書まで用意していました。

要約

前述のとおり、STAC6994 の事例、アンダーグラウンドにおけるプロンプトエンジニアの募集、サイバー犯罪向けの有料サービスで AI がセールスポイントとしてうたわれていること、そして AI をテーマにした詐欺キャンペーンは、いずれも、サイバー攻撃者が AI を攻撃を補完および可能にする手段として導入するとともに、試行錯誤を重ねていることを示しています。

他のセキュリティ企業や研究機関の調査からも、同様の傾向が浮かび上がっています。Claude がメキシコ政府のネットワークを標的とした攻撃に利用された事例もあります。Mandiant の「AI のリスクとレジリエンス」レポートでは、実行中に LLM へ問い合わせを行うことが確認された 2 つのマルウェア系統が報告されています。PROMPTFLUX は、Gemini API を利用して実行時に自己改変を行う実験的な VBScript ドロッパーです。一方、PROMPTSTEAL は IRON TWILIGHT (APT28) によるものとされており、ウクライナに対する実際の攻撃活動において、国家支援型のマルウェアが LLM に問い合わせることが確認された初の事例となりました。Verizon DBIR では、VoidLink (AI エージェントによって 6 日間で構築されたマルウェアフレームワーク) と PromptLock (2025 年 8 月に ESET が初めて発見した AI を活用したランサムウェア) が報告されています。ただし、PromptLock は後に、実際の環境で悪用された悪意あるサンプルではなく、学術的な概念実証 (PoC) である可能性が高いことが明らかになりました。

しかし、AI が主導する完全自律型の侵入キャンペーンが大規模に実行されたことは、依然として確認されていません。GuidePoint GRIT の「ランサムウェアレポート 2026 年版」では、「技術力の高いアフィリエイトが完全自律型のランサムボットを展開する」という懸念は過大視されていると明確に指摘されており、成熟度の低いサイバー攻撃者における AI / LLM の利用も、依然として初歩的な段階にあるとされています。Recorded Future の AIM3 評価では、観測された AI マルウェアの大半が、AI マルウェア成熟度モデルのレベル 1 ～ 3 (実験、導入、最適化) に分類されました。また、被害者のホスト上でローカルモデルを実行する、AI をマルウェアに完全に組み込んだ事例については、確認されたものはないとしています。そして、「防御側は、SF 的なシナリオに過剰反応するのではなく、正規の AI サービスの悪用を監視し、既存のセキュリティ対策を強化するとともに、脅威を AIM3 のレベルにマッピングすることを優先すべきだ」と結論付けています。

Anthropic が報告した GTG-1002 の事例では、Claude Code が約 30 の組織を標的としたデータ窃取キャンペーンの 80 ～ 90% を自動化したと記載されており、公開情報で確認されている中では自律運用に最も近い事例ですが、依然として例外的なケースであり、トレンドを示すものではありません。

第 2 部：
エンタープライズ AI のリスク

AI はすでに企業内部に入り込んでいます。コーディングエージェントは開発者の認証情報を使ってコードを作成および展開し、エージェント型アシスタントはメールを読み取って API を呼び出し、複数のシステムが相互に接続されるようになっています。また、社内チームはマネージドクラウドインフラ上でオープンウェイトモデルを試しています。オープンソース / オープンウェイトモデルは、能力がますます向上している一方、一般にクローズドソース / クローズドウェイトモデルの数分の一のコストで利用でき、企業が推論処理を自社で管理し、データの管理権を維持できるという利点があります。一方、その大部分は中華人民共和国 (PRC) を出所としています。ガバナンス上の課題は、もはや AI の導入を許可するかどうかではありません。危険なデフォルト設定が定着してアタックサーフェスが固定化される前に、いかに AI を安全に導入および運用するかが問われています。

正式に認められた用途で AI をほとんど利用していない、より慎重な組織でさえ、シャドー AI の問題を抱えている可能性があります。そして、後ほど詳しく説明するように、シャドー AI やデータ漏えいに対する懸念は、現実に根ざしたものです。リスク許容度、管理策、可視性の具体的な内容については、第 4 部で詳しく説明します。

2026 年に見られる特徴的な傾向の 1 つが、AI に特化した開発インフラ、特に開発者が AI ツールを取り巻く中で築く信頼関係が攻撃対象となっていることです。

攻撃対象としての AI 開発インフラ

2026 年に見られる特徴的な傾向の 1 つが、AI に特化した開発インフラ、特に開発者が AI ツールを取り巻く中で築く信頼関係が攻撃対象となっていることです。大半の AI 関連の脅威が、いまだ新たに出現した段階にあるか、概念実証 (PoC) にほぼとどまっているのとは異なり、この領域では現在進行形で攻撃が発生しています。

SANDWORM_MODE npm ワームキャンペーンでは、正規の開発者向けユーティリティや AI コーディングツールを模倣するように設計された、少なくとも 19 個のタイポスクワッティングパッケージが確認されました。これらのパッケージは不正な MCP サーバーをインストールし、埋め込まれたプロンプトインジェクションによって正規の AI アシスタントを誘導し、ユーザーに気付かれないまま SSH キーやクラウド認証情報を取得して抽出していました。

Nx Console の VS Code 拡張機能の侵害では、HashiCorp Vault、npm、AWS、GitHub、1Password、Anthropic の API キーなどの認証情報が窃取されました。これは、より広範な TeamPCP / mini-Shai-Hulud キャンペーンの一部とみられています。開発者の IDE 内に侵害された拡張機能が入り込むと、重要な認証情報やリポジトリに直接アクセスすることができます。

2026 年 3 月に発生した **Anthropic Claude Code のソースコード流出** は、別の側面を示す事例です。このケースでは、ソースコードが誤って npm パッケージとして公開されました。報道によると、このコードには約 1,900 個のファイルと 50 万行のコードが含まれており、未リリースの機能に関する情報も含まれていました。Anthropic は顧客データが流出していないとしていますが、流出したコードを分析することで、今後バグや脆弱性が発見される可能性があります。AI ツール自体が、情報収集の標的になっています。

開発チームを持つすべての組織に対し、これを最優先で対処すべき緊急のリスクとして捉えることを推奨します。ここで必要となる対策は、主にプロセスに基づくものです。依存関係の慎重な管理、バージョンの固定、新たなオープンソースライブラリを導入する前のレビューを徹底するとともに、開発者のエンドポイントを綿密に監視することが重要です。(第 4 部では、エージェント型 AI の導入による影響範囲を抑えるために、防御側が今すぐ実行できる 7 つの対策を紹介します)。

リスクが集中する領域

AI サービスと企業システムをつなぐアイデンティティ基盤によって、既存のガバナンスでは対応できないリスクが生じています。IBM X-Force は、2025 年中に 30 万件を超える ChatGPT の認証情報がダークウェブで販売されていたと報告しています。また、2025 年 2 月のフォーラム投稿では、2,000 万件を超える ChatGPT アカウントが盗まれたとされていましたが、この数字は確認されていません。Salesloft/Drift のインシデントでは、具体的な攻撃手法が明らかになりました。Drift の OAuth トークンが複数の Salesforce 環境への侵入に悪用され、チャットボットに付与されたトークンベースのアクセス権が、そのままシステム侵害につながり得ることが示されました。

AI サービスと企業システムをつなぐアイデンティティ基盤によって、既存のガバナンスでは対応できないリスクが生じています。

BeyondTrust は、過去 1 年間で企業環境内のアクティブな AI エージェントが 466.7% 増加したと報告しています。こうしたエージェントは、プロンプトインジェクション、悪意のあるツール呼び出し、汚染されたデータ入力といった固有のアタックサーフェスを生み出します。Microsoft Copilot の脆弱性 CVE-2025-32711 (EchoLeak) は、不十分な入力サニタイズによって、ユーザーの操作を必要としないゼロクリック型のデータ漏えいが可能になることを実証しました。マシンアイデンティティに MFA に相当する仕組みがなく、長期間有効なシークレットを保持し、過剰な権限で動作している場合、それらは攻撃者にとって現実的で魅力的な標的になります。

攻撃者はこうした機会を十分に認識しており、実際にこれらのアタックサーフェスを積極的に洗い出しています。2026 年 1 月、GreyNoise は LLM の導入環境を対象としたマッピングを行うキャンペーンを報告しました。研究者は「Bizarre Bazaar 作戦」と呼ばれるキャンペーンを確認しました。このキャンペーンでは、外部に公開された LLM および MCP エンドポイントを乗っ取り、アクセス可能であることを確認した後、silver.inc に関連するインフラを通じてアクセス権を再販売していました。

ガバナンスのギャップ

業界全体、特に経営幹部では、AI がもたらすセキュリティ上の影響への懸念が広がる一方で、その影響に対処するための人員や能力は不足したままです。

Zscaler のデータでは、ChatGPT へのデータ送信に関連して、4 億 1,000 万件の DLP ポリシー違反が記録されました。Saviynt / Cybersecurity Insiders の「CISO AI リスクレポート 2026 年版」では、回答した組織の 75% がシャドー AI ツールを発見しており、95% がその悪用を検知して封じ込めることに自信がないことが明らかになりました。Pentera によると、AI セキュリティ専用の予算を確保している企業はわずか 1% です。また、Splunk の「CISO レポート 2026 年版」によると、AI が生産性向上や業務量への対応に役立つという点では意見が一致している一方、CISO の大半はデータ漏えい、シャドー AI、ハルシネーションを懸念しています。興味深いことに、Splunk の調査では、AI の成熟度が高い組織の CISO ほど、こうした懸念が強いことが分かりました。また、「生成 AI の導入が進んでいる組織ほど、トレードオフがあることに気付き始めている」と指摘しています。

こうしたトレードオフや懸念、特にシャドー AI をめぐる懸念には、現実的な根拠があります。2026 年 4 月、Vercel は、従業員がサードパーティ製の AI 生産性向上ツール「Context.ai」を使用したことをきっかけとする侵害を公表しました。この従業員は会社の Google Workspace アカウントで Context.ai に登録し、OAuth アプリに「すべてを許可」の権限を付与しました。Context.ai が侵害された際、攻撃者は OAuth トークンを引き継ぎ、Vercel の内部環境へと侵入しました。Context.ai の侵害は、報道によると、同社従業員のデバイスが Lumma Stealer に感染したことが原因とされています。

ガバナンス上の課題は、AI の導入スピードに、それを管理するための組織のプロセスが追いついていないことです。規制環境が厳格化する一方で、AI の導入スピードとガバナンスの成熟度とのギャップは、さらに広がっています。
EU AI 法では、2026 年から高リスク AI システムに全面的な遵守が求められ、違反した場合には全世界売上高の最大 7% に相当する制裁金が科される可能性があります。カリフォルニア州の透明性に関する要件は、2026 年 1 月 1 日に施行されました。しかし、Saviynt / Cybersecurity Insiders の調査では、大企業の 71% が基幹業務システムにアクセスする AI エージェントを導入している一方、そのアクセスを適切に管理できている企業はわずか 16% にとどまることが明らかになりました。

このギャップに対処する第一歩として、プロンプトの通信ログを重要なセキュリティログとして扱うべきです。プロンプトを記録および分析することで、インジェクションの試み、データの持ち出し、内部関係者による不正利用、予期しないエージェントの挙動を検知できます。クラウドホスト型モデルの場合はプロキシ層から、ローカル環境に導入されたエージェントの場合はハーネス層から始めることができ、モデルとの完全な統合を待つ必要はありません。ただし、プロンプトだけですべてを捉えられるわけではありません。ツールやスキルもアタックサーフェスの重要な構成要素になりつつあり、それらも同様に扱う必要があります。

AI の導入は、それを管理するために整備された組織のプロセスを上回るペースで進んでおり、規制環境も厳格化しています。

第 3 部：
AI が変えるセキュリティ運用モデル

トリアージと調査

プレイブックの実行、タイムラインの再構成、自然言語によるレポート作成は、推論モデルが実際のセキュリティ運用で真価を発揮できる領域です。たとえば、Kaspersky の MDR インフラでは、2025 年に約 40 万件のアラートが処理され、AI を活用した検知ロジックによって、アナリストが確認する前に誤検知が除外されました。そのうち 3 万 9,000 件が、さらなる調査のためにエスカレーションされました。AI モデルは、こうした調査にも対応できるようになりつつあります。Prophet Security は 2026 年 3 月、The Hacker News で、悪意のある活動がすぐには明らかにならなかったインシデントについて、AI モデルが文脈を踏まえた分析と調査を行った 2 つの事例を紹介しました。
1 つ目は、古い API キーを持つ休眠状態のユーザーが突然アクティブになったケースです。2 つ目は、ほとんど、あるいはすべてのセキュリティチェックや検証をすり抜けていた可能性のあるフィッシングメールから、悪意を見抜いたケースです。Prophet Security によると、どちらのケースでも、個々のデータポイント自体には不審な点はありませんでした。脅威が明らかになったのは、複数のデータポイントを文脈に沿って分析した結果であり、これは人間の調査担当者が十分な時間と余力がある場合に行うような分析です。もちろん、問題は調査担当者がすべてのアラートを同じレベルで分析できるだけの時間や余力を持ち合わせていないことです。そこで推論モデルが、すべてのアラートに同じレベルの分析を適用することで力を発揮します。

しかし、AI は万能ではなく、慎重なエンジニアリングが必要です。「過剰思考」に関する研究では、推論を一定以上重ねると、モデルが正しい回答を誤った回答に変えてしまうことが示されています。約 7,000 トークンを超えると、正解を否定する方向への回答の反転が、正解を肯定する方向への反転を上回り始め、12,000 トークンではその比率が 3 対 1 に達しました。したがって、SOC の実装では、推論の強度、リトライの深さ、ツール呼び出しのループを、調整可能なエンジニアリングパラメータとして扱うべきです。

総じて、AI の導入は一様ではなく、過度な期待は調整する必要があります。Splunk の CISO レポートによると、現在、セキュリティ業務に生成 AI を導入している CISO はわずか 40% です。エージェント型 AI は、現在導入している企業がわずか 6% にとどまる一方、39% が導入を検討しています。

さらに、防御側が利用する AI ツール自体が、新たな攻撃リスクを生む可能性もあります。たとえば、Cobalt の「侵入テストの現状」レポートでは、AI および LLM アプリケーションで高リスクの問題が見つかる割合は従来型ソフトウェアの 2.7 倍に上る一方、侵入テスト中に解決された問題は 38% にとどまることが明らかになりました。

メモリの問題

本番環境で導入されている AI エージェントの多くは、研究者がエージェントの成熟度の「第 2 世代」または「第 3 世代」と呼ぶ段階にとどまっています。つまり、ツールで機能を拡張しているものの、セッションごとに状態がリセットされるアシスタントです。こうしたエージェントは学習済みの知識は保持していますが、前日に同じアラートパターンが現れて無害だったことや、先月特定の IP 範囲をブロックしたことで VPN に障害が発生したことを記憶できません。

シニアアナリストとジュニアアナリストの違いは、必ずしも推論能力にあるわけではありません。それぞれのインシデントに対して、過去のインシデントから得た経験、環境に関する知識、そして何が有効な対応なのかを直感的に判断する力をどれだけ持っているかという点にあります。現在の AI アシスタントは、シニアアナリストというよりジュニアアナリストに近く、すべてのアラートを初めて見るかのように調査しています。研究者は、メモリをオペレーティングシステムのように扱うアーキテクチャ、外部メモリをミドルウェアとして利用するサービス、さらに過去と現在という 2 つの時間軸を考慮して推論するナレッジグラフのアプローチなどを研究しています。こうしたテクノロジーが成熟するまでは、AI システムは有用なアシスタントとしては機能しても、自律的に業務を遂行する存在にはならないでしょう。

検知の限界

振る舞い検知システムは、正常なデータが圧倒的に多いという極端なクラス不均衡の中で、リアルタイムのテレメトリデータを処理します。1 日に数兆件のイベントを処理するプラットフォームでは、イベントが集約されて数百万件の重複排除済みの検知となり、そこから優先順位付けによって数千件の高重大度アラートに絞り込まれ、さらに人間による確認後も重要と判断される調査案件が、1 日数百件生じます。最初のイベント数に対する最終的な調査案件の比率は、100 億分の 1 に近づきます。Stefan Axelsson が侵入検知におけるベースレートの誤謬について行った基礎的な研究 では、その数学的な仕組みが明らかにされています。攻撃の事前確率が十分に低い場合、特異度 99.9% の検知システムでさえ、真の脅威 1 件に対して数百万件もの誤警報を発生させる可能性があります。

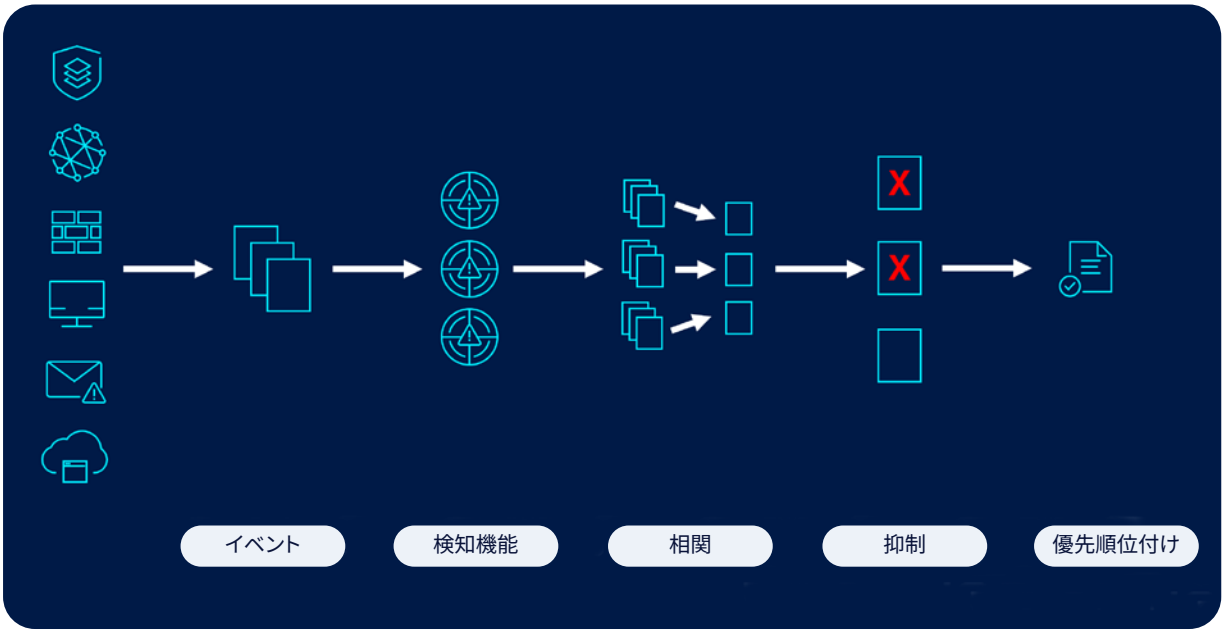


図 6：マルチステップパイプラインの概要

NorthSec 2026 で発表されたソフォスの研究は、なぜ多層的な検知が依然として重要なのかを示しています。2 週間分のテレメトリ 11.8 兆件は、単純な指標の照合、相関ルール、専用の機械学習モデル、重複排除、相関分析、抑制といった、複雑度を段階的に高めた検知処理によって絞り込まれ、最終的に高重大度および重大度の高いアラート 81,573 件となりました。これは 1 組織あたり平均 50 件未満です。その後も残ったアラートの中から、ソフォスは TamperedChef キャンペーンに関連するインフォスティーラーによる攻撃を特定しました。

推論モデルは、このレイヤーの上位に位置付けられます。推論モデルは、調査の仮説を立て、テレメトリを取得および統合して証拠に基づく要約を作成し、明示的なガードレールの下で、範囲を限定した対応アクションを調整します。推論モデルは、独自のインシデントデータや環境固有のベースライン、さらに汎用的な推論システムには利用できない継続的なフィードバックループを用いて訓練されたモデルに取って代わるものではありません。

McKinsey の「AI Trust Maturity Survey（AI 信頼成熟度調査）」は、組織が抱える制約を定量化し、回答者の約 3 分の 2 が、エージェント型 AI の導入拡大における最大の障壁としてセキュリティ上の懸念を挙げていることを明らかにしました。具体的なリスクとしては、「不正確さ」(74%) と「サイバーセキュリティ」(72%) が上位を占めています。AI ガバナンスにおいて、成熟度レベル 3 (「必要な責任ある AI の実践をすべて整備するための取り組みを進めている」) 以上に達している組織は、わずか約 30% にとどまります。責任ある AI への取り組みに対して明確な責任者を置いている組織の平均成熟度は 2.6 であるのに対し、明確な責任体制を定めていない組織では 1.8 にとどまります。

AI を活用した防御の制約となっているのは、組織の準備状況です。モデルには十分な能力があります。問われているのは、自律的なセキュリティ運用に求められるレベルの信頼性を確保したうえで、組織がそれらを運用できるかどうかです。

ソフォスの AI 研究：AI レイヤーの防御

前述の多層的な検知に関する研究に加えて、ソフォスの AI チームは、特定の失敗要因に対処するための、いくつかの新しい防御技術も開発しています。

- **Untrusted Data Scanner (UDS)** は、外部コンテンツに含まれるプロンプトインジェクションを検査し、LLM やエージェントに到達する前に検知します。すべての LLM のユースケースには、指示チャネル (アプリケーションがモデルに何をすべきか伝えるチャネル) と、データチャネル (モデルが処理する信頼できないコンテンツを受け取るチャネル) があります。プロンプトインジェクションは、データチャネルに命令を紛れ込ませます。UDS は、汎用的なジェイルブレイク検知機能ではなく、この問題に特化した設計になっています。初期の結果では、誤検知率が低いことが示されています。これは重要です。セキュリティデータには、インシデントレポート、マルウェアの分析記事、侵入テストの結果など、セキュリティに関する記述が大量に含まれています。脅威レポートに含まれるセキュリティ上の記述をことごとくプロンプトインジェクションと判定してしまうような検知システムでは、実用になりません。
- **LLM スコーピング**は、クエリが各機能の想定された用途の範囲内にあるかどうかを判定し、対象範囲外のリクエストを LLM に到達する前に拒否します。評価した手法の中では、スコープ判定モデルがシステムプロンプトベースの防御を大きく上回る性能を示し、複数の公開ベンチマークにおいて攻撃成功率をほぼゼロまで低減しました。また、ソフォスのチームはマルチラウンド自動レッドチーミング (MART) を適用し、攻撃者による適応に対してスコープ判定モデルの堅牢性を高めました。

- 「3 つのヘッド」を持つ BERT ベースのモデルである **CerBERTus** は、バイナリジェイルブレイク検知が抱える脆弱性に対処します。1 つの共有エンコーダーから 3 つの分類ヘッドに入力し、それぞれ「有害性」（主分類）、「目的カテゴリ」（ユーザーが何をしようとしているか）、「表現スタイル」（リクエストがどのように提示されているか）を分類します。補助タスクが帰納バイアスとして機能し、モデルが「目的」と「それを包む表現」を分離して捉えられるよう促します。チームは、この分離を学習させ、ストレステストするために、目的 (サイバー攻撃、詐欺、爆発物、薬物合成、プライバシー、過激派プロパガンダなど) と、敵対的なフレーム (ロールプレイ、フィクション、緊急性、学術的な口実、難読化など) の組み合わせを体系的に網羅したプロンプトコーパスを構築しました。
- **異常検知**：Black Hat USA 2025 で発表された研究では、チームは異常検知と LLM を組み合わせました。その結果、この手法の有効性は悪意のあるコマンドラインを見つけることではなく、多様な正常なコマンドラインを生成することにあると判明し、コマンドライン分類器の誤検知率を大幅に低減できました。
- CAMLIS 2025 で発表された **LLM ソルト処理**は、パスワードのソルト処理から着想を得た手法で、LLM に意図的に特定の振る舞いの違いを持たせることで、事前に用意されたジェイルブレイクを同一構成の LLM 環境間で再利用できないようにします。

自律型攻撃者に対する欺瞞

この 20 年間、サイバー攻撃への欺瞞が主要なセキュリティ制御として位置付けられることはありませんでした。自律型攻撃者を相手にする場合、この前提は変わります。AI エージェントは休むことことなく、考えられるあらゆる経路をマシン速度で列挙できます。Palisade LLM Agent Honeypot や MANTIS フレームワークで提案されているような欺瞞環境は、エージェントのコンテキストを誤解を招く情報で満たし、推論に割り当てられたリソースを浪費させ、実際の経済的コストを負わせる可能性があります。その一例である HoneyTrap では、従来の防御と比較して攻撃者のリソース消費量が 149% 増加することが実証されました。この分野はまだ未成熟です。しかし、欺瞞を主要な防御手段に押し上げる条件を、まさに攻撃側の AI 活用の拡大が生み出しつつあります。

時間差がもたらす優位性

AI を利用する攻撃者には、調達サイクルもコンプライアンス審査も、変更諮問委員会 (CAB) もありません。攻撃者は、ツールを導入し、その成果をわずか数日で実際の攻撃活動に組み込むことができます。FortiGate デバイスを標的とした AI を利用したキャンペーンでは、55 か国で 600 台を超えるファイアウォールがわずか 5 週間で侵害されました。これは、商用 AI ツールを利用した攻撃でした。

防御側では、検知が作動すると、AI を活用した調査によって数分で封じ込めの推奨策が提示されます。しかし、顧客側の修復プロセスでは、変更チケットの起票、メンテナンス時間の確保、48 時間以内の SLA が求められます。調査は速くなりましたが、対応は速くなっていません。これは、昨年相次いで急速に悪用された 2 件の脆弱性にも当てはまる可能性があります。前述のとおり、CVE-2026-10520 (Ivanti Sentry) は PoC の公開から 24 時間以内に悪用され、CVE-2026-42208 (LiteLLM) は 36 時間以内に悪用されました。

AI を活用したトリアージは、間違いなく防御側の対応速度向上に役立ちますが、セキュリティの基本をあらかじめ徹底しておくことも、この速度差への対処に寄与します。具体的には、アタックサーフェスを縮小し、MFA を徹底し、予防策が機能しなかった場合でも攻撃の内容を把握できるだけの十分なテレメトリを維持することが重要です。

第 4 部：
セキュリティリーダーが取り組むべきこと

本レポートで示したエビデンスからは、組織の現状に応じて、いくつかの判断を行うべきことが分かります。

AI 導入におけるリスク許容度を見極める

本レポートで示したエビデンスは、組織が解決しなければならない、あるジレンマを浮き彫りにしています。AI の導入には大きなリスクが伴いますが、導入の遅れにもリスクがあります。攻撃が加速する中で慎重になりすぎれば、本レポートが本質的な課題として指摘している防御側の対応速度で後れを取ることになります。リスクを受け入れることは、その一環として不可避です。重要なのは、取るべきリスクを見極めることです。

本レポートで示したエビデンスは、組織が解決しなければならない、あるジレンマを浮き彫りにしています。AI の導入には大きなリスクが伴いますが、導入の遅れにもリスクがあります。

ソフォスが社内で利用しているリスクフレームワークでは、リスクの上振れ余地と下振れの抑制という 2 つの軸に基づいて、取るべきリスクと避けるべきリスクを区別しています。取るべきリスクには、十分な効果が期待できること、影響範囲を限定するための対策が講じられていること、パイロット導入やロールバック、代替策など、問題が発生した場合に元に戻せる手段があること、責任の所在が明確であること、そして明確なフィードバックループが定められていることが求められます。避けるべきリスクには、被害が際限なく拡大する可能性があること、譲れない原則 (セキュリティ、プライバシー、法令、コンプライアンス) に違反すること、復旧計画がないこと、責任の所在が不明確であること、あるいは顧客に直接影響を及ぼす変更について人間による判断を省略することなどが挙げられます。

特に AI については、このフレームワークを導入前に検討すべき実践的な問いに落とし込むことができます。最悪の場合、現実にどのような事態が起こり得るのか？影響範囲はどの程度に及ぶ可能性があるか？元に戻すことはできるか？所有者は誰か？などの問いです。AI エージェントを活用したトリアージツールを分離されたデータセットに対して読み取り専用で試験導入するなど、リスクを限定できる AI 導入は、積極的に進めるべきです。一方、AI エージェントに本番環境のインフラへの書き込み権限を与えるなど、リスクが大きい導入については、実施に先立ってリスクを低減するか、対象範囲を限定する必要があります。メリットが小さい一方でリスクが大きい導入は、避けるべきです。

どのような状況でも、譲れない原則は守る必要があります。それは、顧客からの信頼、安全性、セキュリティとプライバシー、法務やコンプライアンス、そして運用の安定性です。AI によって変化が加速する環境における CISO の役割は、組織が賢明にリスクを許容し、無謀なリスクを避けられるようにすることです。

エージェント型 AI を導入する組織：被害の影響範囲を縮小

ソフォスは、AI エージェントの導入を検討する組織が影響範囲を縮小できるよう、今後 1 ～ 6 か月で実行可能な 7 つの対策を提案しています。

1. **エージェントのサンドボックス化によって、**エージェントのプロセスを制御された境界内に置きます。ほとんどのエージェントハーネスには何らかのサンドボックス機能が備わっていますが、通常は明示的に有効化する必要があります。可能であれば、プロセス分離を強化するため、ローカルサンドボックスよりもリモートまたはクラウドベースの実行環境を選択します。

2. **認証情報の分離**により、エージェントがシークレットを直接扱わないようにします。別のプロセスがボールト (保管庫) から認証情報を取得して API 呼び出しに注入し、サニタイズした応答だけを返すことで、長期間有効な認証情報が LLM のコンテキストウィンドウに入るのを防ぎます。

3. **シールドされたツールエンドポイントは、**さらに一歩進んだ対策です。エージェントはツール名を指定してツールを呼び出しますが、認証情報はブローカープロセスが保持し、固定されたスキーマに従って実際の API 呼び出しを行い、ツールごとに外部通信の許可リストを適用します。エージェントに許される操作は、どのツールを呼び出すか、そしてどのパラメータを渡すかを選択することだけです。

4. **外部通信の制限とネットワーク監視では、**侵害されたエージェントをデータ窃取の経路として捉えます。具体的には、外部通信に含まれるシークレットを検知し、外部へのリクエストに対して容量のしきい値を設定するとともに、Web カテゴリ分類によって未分類ドメインや新規登録ドメインへのリクエストをブロックします。

5. **EDR の適用範囲は**エージェントホストの動作まで対象にする必要があります。コンテナ、一時的な VM、管理対象外の開発者マシンは、可視性の一般的な盲点となっています。

6. **人間による承認**を必須とすることで、エージェントが自律的に実行する実行プレーンと、人間が管理する制御プレーンを分離します。取り消しのできない操作には、エージェントが偽装できるような単なる「OK」ボタンではなく、暗号学的な承認プロセスを要求します。

7. **インジェクションの伝播境界**は、インジェクションがセッション単位からエージェント単位、さらにエージェント間へと広がるにつれて高まるリスクに対処します。メモリへの書き込みやエージェント間の引き継ぎを、セキュリティイベントとして扱います。

2026 年初頭に OpenClaw が注目を集めるようになった際、3 万を超えるインスタンスがインターネットに公開されており、攻撃者はすでに、モジュール式の「スキル」をボットネット攻撃に悪用する方法について議論していました。ソフォスのレッドチームは、これを直接テストしました。対策につなげられる指摘が 23 件、Active Directory (AD) の偵察に要する時間が 3 日から 3 時間に短縮され、さらに、人手では現実的な時間内に得ることが難しいほど詳細な監査結果が得られました。しかし、チームはテストを実施するよりも、エージェント型 AI の行動範囲を安全に制限する仕組みの構築に多くの時間を費やしました。教訓は明確です。エージェント型 AI は導入する価値があるほど強力ですが、その場合セキュリティフレームワークを先に構築することが条件です。

可視性、アイデンティティ、パッチ適用

まず、AI サービスの利用状況を可視化することを推奨します。Salesloft/Drift および Vercel のインシデントで明らかになった AI のアタックサーフェスを踏まえると、これが最初に取り組むべき、最も効果の大きい施策です。環境内の AI ツールのインベントリを作成します。そのうち、エンタープライズシステムに接続する OAuth トークン、API キー、サービスアカウントを保持している AI ツールを特定します。AI ツールを通じてどのようなデータが流れ、どのような権限が付与されているのかを把握します。可視化できないものは統制することができません。

この可視化によって、AI レジストリを整備します。そこには、すべての AI ツール、エージェント、モデル、それらからアクセス可能な認証情報、サンドボックス設定、想定される接続要件を記録します。このレジストリがなければ、認証情報の分離、外部への通信制限、サンドボックス化といった対策を一貫して適用することはできません。他の SaaS アプリケーションと同じように、厳格にアイデンティティを管理します。条件付きアクセスポリシーを適用し、AI サービスのアカウントには MFA を必ず適用し、AI ツールが利用する送信経路に DLP 制御を適用します。また、AI と接続する連携機能の認証情報は、他の特権アクセスと同じスケジュールでローテーションします。IBM X-Force による認証情報データの調査結果と、Nx Console のサプライチェーン攻撃はいずれも、AI サービスの認証情報が窃取の標的になっていることを裏付けています。

パッチ適用のサイクルがいまだに数週間単位である組織では、リスクがますます高まります。CISA BOD 26-04 で定められた 3 日間という期限は最低限の基準です。AI を活用した攻撃によって、脆弱性の公開から悪用までの時間が数時間にまで短縮されているためです。インターネットに公開された重大な脆弱性を数日以内にパッチ適用できないのであれば、その遅れには無視できないリスクがあります。

AI を活用した攻撃によって、脆弱性の公開から悪用までの時間が数時間にまで短縮されています。

エビデンスに基づいて評価する

アナリストが使用する AI ツールからログやテレメトリを取得できるようにします。検証にかかる負担、つまりアナリストが AI の出力を確認するのに費やす時間を測定します。プロンプトの効率、つまり有用な結果を得るまでに何回試行する必要があるかを測定します。さらに、アナリストがシステムを過信したり、逆に過小評価したりしていないか、適切な信頼度で利用できているかを測定します。一般的なベンチマークではなく実際の SOC 業務からテストセットを構築し、モデルのアップグレード後も再現性を確保できるよう、手順をバージョン管理します。

AI サプライチェーンを保護する

AI は、従来のソフトウェア依存関係にとどまらないサプライチェーンリスクをもたらします。モデルの重み、学習データの出所、MCP サーバーの完全性、そして推論インフラ自体が、いずれも信頼境界となります。マネージドクラウドプロバイダーを通じてオープンウェイトモデルを導入する組織は、非公開のコードやデータを送信する前に、正確なモデルバージョン、提供するリージョン、データ保持期間、学習への利用に関するポリシー、ログ取得の範囲、契約条件を確認する必要があります。DeepSeek の API のように、中国を拠点とする SaaS エンドポイントを介して提供されるモデルについては、より厳格な境界を設ける必要があります。このようなモデルは、公開データや合成データを用いた機能テストに使用し、独自のリポジトリや機密性の高いエンジニアリング情報には使用すべきではありません。

まとめ

サイバー攻撃者「STAC6994」が使用したフレームワークには、Cobalt Strike のプロファイル、EDR 回避モジュール、認証情報窃取ツール、ラテラルムーブメント用ユーティリティ、そして正規のクラウドインフラの背後に隠された C2 チャネルが含まれていました。すべてのコンポーネントは馴染みのあるものでしたが、例外はその開発プロセスです。AI エージェントは、人間の開発者なら不可能なほどの速さで回避技術を反復改良し、構造化されたラボ環境で特定の製品に対してテストを行い、その結果を Git でバージョン管理したうえで、ランサムウェアを展開してデータを窃取するオペレーターに成果物を引き渡していました。

Sophos MDR およびインシデント対応の事例、CTU のインテリジェンス、SophosLabs の分析、Sophos AI の研究、そしてセキュリティ業界全体から得られた知見を総合すると、AI はセキュリティの攻守双方で、従来から存在するさまざまな業務を加速させています。AI は、攻撃者が既知の攻撃チェーンをより速く進めることを可能にしています。防御側は、トリアージと調査をこれまでより短い時間で行えるようになります。一方で、企業における AI 導入に関するアイデンティティ管理とガバナンスのレイヤーに、新たなリスクが生じています。AI によって、既存の検知アーキテクチャでは対処できない、根本的に新しい侵入手法が生まれたわけではありません。英国 AI セキュリティ研究所 (UK AI Security Institute) は、18 か月間で自律型攻撃能力が約 6 倍向上したと測定しています。モデルが世代を重ねるごとに、以前の世代では実行できなかった多段階の侵入シナリオの次のステップへ進めるようになっています。この傾向を踏まえれば、決して油断はできません。

明確な結論につながる 3 つの重要なテーマがあります。

1. **分類**は正確でなければなりません。AI の脅威をすべて一括りにしてしまうと、それぞれに必要な防御態勢が見えなくなるためです。
2. **評価**は厳密でなければなりません。モデルのルーティング、アラートの抑制、エージェントの推論の深さのいずれが問題であっても、その答えは測定結果に基づくべきです。
3. **透明性**は、具体的な行動を伴うものでなければなりません。成功だけでなく、そこから得られた有益な失敗も含めて、取り組みの内容を明らかにする組織こそが信頼を築くからです。

ソフォスは、自律型エージェントが既知のプレイブックを実行する段階から、大規模に未知の攻撃経路を発見する段階へと進むのか、アンダーグラウンド経済が、サービスとしてのランサムウェア (RaaS) が恐喝攻撃の参入障壁を引き下げたのと同じほど大幅に、攻撃のハードルを引き下げる、製品化された AI 攻撃ツールを生み出すのか、企業の AI ガバナンスが、現在の AI 導入の波によって生じたアイデンティティ管理上のリスクを解消できるほど速く成熟するのか、そして、悪用からパッチ適用までの時間がさらに短縮され、いまだに対応サイクルを数週間単位で考えている組織に、対応体制の見直しを迫ることになるのかについて今後 1 年にわたって注視していきます。

もう、以前と同じ状況ではありません。対応のスピードを変えなければ、防御側は自分たちが想定していた以上に後れを取ることになります。

投稿者

Nash Borges
エンジニアリングおよび
AI 担当 SVP

Adarsh Kyadige
シニアマネージャー
AI リサーチ

Ross McKerchar
CISO

Rafe Pilling
脅威インテリジェンス担当ディレクター
X-Ops Counter Threat Unit

Matt Stec
ディレクター
Sophos AI

Ryan Westman
脅威調査担当シニアマネージャー
X-Ops Insights

Matt Wixey
シニア脅威リサーチャー
X-Ops Insights

ソフォスチーム

Sophos X-Ops は、最先端のサイバーセキュリティイニシアチブです。脅威インテリジェンス、人工知能、MDR 運用、社内セキュリティ運用など、ソフォス社内のさまざまな専門セキュリティ分野から 1,000 人を超えるエキスパートを結集しています。

この部門横断型タスクフォースは、急速に進化し、高度化する今日のサイバー脅威に対する組織の防御力を強化します。

このタスクフォースの専門知識を結集することで、Sophos X-Ops はサイバー攻撃に対して多面的な対応を提供し、包括的な防御、検知、対応能力を実現します。

ソフォスは、この協調的かつ革新的なアプローチにより、比類のない脅威対応を実現しており、これが、卓越性のベンチマーク、そしてサイバーセキュリティのリーダーとしてソフォスが位置付けられている理由となっています。

Sophos X-Ops は、部門横断的なタスクの専門知識を組み合わせています。Sophos X-Ops の部門横断型チーム間の連携がインテリジェンスの共有を促進し、新たな知見に迅速に適応できるようにすることで、検知と対応のスピードを高めるとともに、ソフォスのお客様に対する全体的な保護能力を強化しています。

AI セキュリティに関するニーズや、ソフォスの支援についてご相談は、ソフォスの Web サイトにアクセスいただくか、当社のアドバイザーまでお問い合わせください。

ソフォス株式会社営業部
Email: sales@sophos.co.jp

2026-07-20 JP (NP) E-5820