

トレンドマイクロ AI セキュリティの 現状レポート（2025 年上半期）



はじめに 3

PART 1 : AI インフラに対する現在の攻撃 4

Pwn2Own ベルリンで注目された AI 4

AI アプリケーションで発見された主な脆弱性 4

Chroma DB の 익스프로イト 4

NVIDIA Triton Inference Server の 익스프로イト

5

Redis の 익스프로イト 5

NVIDIA Container Toolkit の 익스프로イト 6

その他の調査結果 7

PART 2 : AI 特有の脆弱性 8

複雑な LLM ベースのアプリケーションへの攻撃 8

プロンプトベースの攻撃の進化 9

PART 3 : サイバー犯罪における AI の活用 12

複雑な LLM ベースのアプリケーションへの攻撃 12

AI による自動翻訳の戦略的優位性 12

攻撃者による AI 採用の選択性 12

ディープフェイクがサイバー犯罪の参入障壁を下げた

13

eKYC 回避攻撃における AI 対決 13

PART 4 : これからの道筋 15

エージェント AI の複雑化が進む中で 15

デジタルアシスタントによる支援 16

結論 : AI セキュリティの最前線に立つベンダーたち 17

参考文献 19

はじめに

人工知能（AI）の幅広い活用は、企業の効率化に貢献する一方で、それを狙う攻撃者にとっても利便性をもたらしており、今後 3 年間で AI への投資を拡大しようとする組織が増加しています。2025 年にはセキュリティリーダーの 93%が AI を利用した攻撃を日常的に受けると予測しており、意思決定者たちは業務フローやセキュリティ体制を根本的に見直す必要に迫られています。世界経済フォーラムの「グローバル・サイバーセキュリティ展望」レポートによれば、調査対象となった組織の 66%が、今年最も大きな影響をサイバーセキュリティにもたらすのは AI であると回答しています。「AI モデルの使い方については、まだ多くの疑問が残っています」と、South London and Maudsley NHS Foundation Trust (SLAM) の CTO である Stuart MacLellan 氏はトレンドマイクロに語っています。「個人情報共有に関するリスクは私たちの領域では非常に深刻です。私たちは現在、トレーニングを進めるとともに、どのデータがどこに保管され、AI モデルにおいてどのように利用されるのかを明確にするルールを策定しているところです。」

AI の存在感はますます高まっており、日々のタスクを管理するデジタルアシスタント（DA）から、ビジネス判断を自動化する AI エージェントに至るまで、私たちの生活のあらゆる側面に浸透しつつあります。今年、世界有数のハッキングコンテストである Pwn2Own に AI 専用カテゴリーが新設されたことは、単なる流行への対応ではなく、AI がサイバーセキュリティの境界線を塗り替える重要な存在になっていることの表れです。

この新カテゴリーは、今年ベルリンで開催された OffensiveCon カンファレンス中の Pwn2Own イベントで初登場となり、防御側が常に後手に回るのではなく、設計段階からセキュアであることが求められる AI システムの必要性に改めて注目を集める結果となりました。

このレポートでは、トレンドマイクロが AI の活用に伴う期待とリスクの両面について幅広く考察し、Pwn2Own での初の AI 部門における成果や、最新の AI 関連調査に基づく専門的な見解を紹介しています。次世代のエージェントイック AI アプリケーションによってもたらされる新たな脅威の状況や、犯罪者たちが AI をどのように利用して自らのビジネスモデルを強化しているかを明らかにしながら、今後の見通しと、それに対応するためにトレンドマイクロが進めている取り組みについても解説していきます。

主なポイントとして、防御側が常に警戒を怠らず、AI システムのすべての構成要素を確実に保護する必要があることが挙げられます。たとえすでに安定していると考えられている部分であっても例外ではありません。そのためには、サードパーティ製ライブラリやサブシステムを含むすべてのソフトウェアコンポーネントのインベントリを維持し、これらの構成要素に対する定期的なセキュリティ評価を実施するなどのベストプラクティスが有効です。こうした取り組みにより、攻撃者に悪用される前に潜在的な脆弱性を特定し、対処することが可能になります。

PART 1：AI インフラに対する現在の攻撃

Pwn2Own ベルリンで注目された AI

Pwn2Own は、トレンドマイクロの Zero Day Initiative™ (ZDI) が主催するグローバルイベントであり、2007 年の開始以来、ゼロデイ脆弱性の発見と責任ある開示のための主要なプラットフォームとして機能してきました。これまでに、世界トップクラスのセキュリティ研究者とソフトウェアベンダーの協力を促進し、脆弱性が実際に悪用される前に防ぐ上で大きな役割を果たしてきました。その伝統に則り、攻撃対象を用いたオフensiveテストや脆弱性研究を目的としたこのイベントでは、AI カテゴリーが新たに追加され、AI エコシステムを支える基盤アーキテクチャへの攻撃が求められました。このカテゴリーでは、AI モデルの構築や実行に広く使われている開発ツールキット、ベクターデータベース、モデル管理フレームワークなど、6つの対象が設定されました。

このイベントには、イングランド、フランス、ドイツ、イスラエル、ポーランド、セルビア、シンガポール、韓国、台湾、米国、ベトナムといった国々から、世界中の参加者が集まりました。3日間にわたり、合計 28 件のユニークなゼロデイ脆弱性が発見され、そのうち 7 件は AI カテゴリーから報告されました。

AI アプリケーションで発見された主な脆弱性

Chroma DB の 익스プロイト

Chroma DB は、主に Python で開発されたオープンソースのベクターデータベースであり、検索拡張生成 (Retrieval Augmented Generation) に基づく AI エージェントで高い人気を誇っています。他のベクターデータベースと同様に、テキストの断片に紐づくベクターを保存し、大規模言語モデル (LLM) へのプロンプトに近い内容をデータベースから取り出して組み込むことで、より関連性の高い、そしてできる限り事実に基づいた応答を生成できるようにする仕組みです。

これまで Chroma に関する脆弱性はあまり知られていませんでしたが、今回の Pwn2Own イベントで、Summoning Team の Sina Kheirkhah 氏が初めての AI カテゴリーでの勝利となる 익스プロイトに成功しました。この攻撃は、開発段階で残されたままのアーティファクト (不要な構成要素) を悪用するものであり、本番環境に移行する前にシステムを徹底的に検証・整理しないリスクが浮き彫りとなりました。組織にとっては、本番環境への展開時に非本質的なアーティファクトを完全に削除し、セキュリティ基準に準拠するための定期的な監査を含む厳格な運用プロトコルの重要性を示しています。

トレンドマイクロの先進脅威リサーチチーム (FTR) は、昨年実施した AI システムの公開状況に関する調査を再度実施しました。昨年は合計で約 240 台のサーバを確認しましたが、2025 年 5 月の調査では、完全に無防備な Chroma サーバが 200 台以上観測されました。これらのサーバは認証なしにデータの読み取り、書き込み、削除が可能である状態でした。さらに 300 台のサーバは公開されていたものの保護されており、今回 Sina Kheirkhah 氏が発見したような 익스プロイトによって、データへのアクセスや、最悪の場合そのマシン全体への侵入が可能になる恐れがあります。

NVIDIA Triton Inference Server のエクスプロイト

Viettel Cyber Security、FuzzingLabs、Qrious Secure など複数のチームが、NVIDIA Triton Inference Server をターゲットに設定しました。使用された多くのエクスプロイトは、すでにベンダー側で修正対応中の既知のバグに基づくものでしたが、最終的に Qrious Secure が 4 つの脆弱性を組み合わせた手法によって完全なエクスプロイトに成功しました。Triton に対する攻撃の多くは、任意のデータをサーバに読み込ませるといったものであり、これらはデータの適切な検証、迅速なパッチ適用、ゼロトラストアーキテクチャの導入などによって防ぐことが可能だったと考えられます。これらの脆弱性は、多数の相互依存するコンポーネントからなる複雑なシステムを管理する難しさを浮き彫りにしています。また、既知の脆弱性でありながら修正されずに残っていたことから、パッチ管理と脆弱性スキャンの継続的な強化の必要性も明らかになりました。

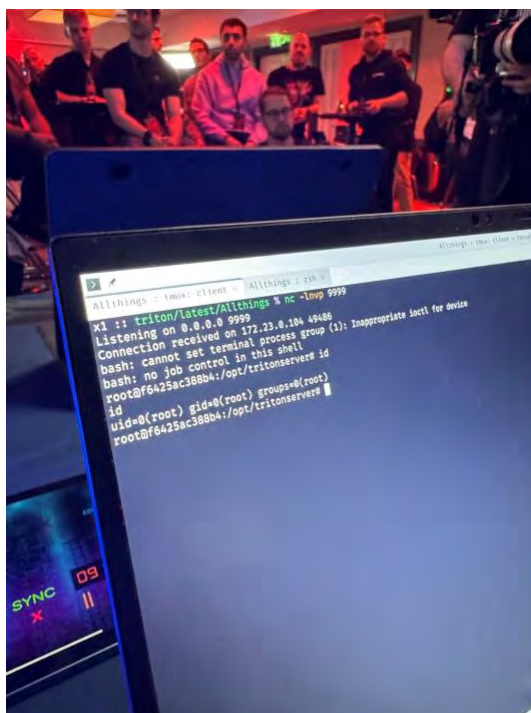


図 1. Qrious Secure は 4 つの脆弱性を連鎖させて NVIDIA Triton Inference Server をエクスプロイト

Triton は、CNCF の KServe サーバのように機能し、通常は Kubernetes インフラストラクチャの一部としてデプロイされます。公開されている KServe サーバは少数存在していますが、それが Triton であるかどうかは現在のところ判別できていません。いずれにしても、公開数はごくわずかにとどまっています。

Redis のエクスプロイト

Redis は、一般的なデータベースクエリに迅速に応答するためのキャッシュとしてコンテンツ配信アプリケーションで広く使用されているキーバリューストアですが、それ以外にも多くの用途があります。バージョン 8 以降では、ベクターの保存と比較をネイティブでサポートしており、それ以前のバージョンではインストール可能なモジュールを通じて対応していました。

Wiz Research は、Redis のベクターデータベースにおける use-after-free (UAF) 脆弱性を利用して勝利を収めました。このエクスプロイトは複数の脆弱性を連鎖させたもので、主に Lua サブシステムに関係しています。Redis 内で使用されていた Lua のバージョンが古く、これが攻撃成立の決定的要因となっており、成熟したシステムにおいても古いコンポーネントやサポート終了のソフトウェアを使用し続けるリスクを明確に示す結果となりました。安定していると思われている構成要素であっても、すべてのコンポーネントに対して警戒を怠らず、適切に保護する姿勢が求められます。そのためには、サードパーティ製のライブラリやサブシステムを含むすべてのソフトウェア構成要素のインベントリを常に把握し、定期的に更新・パッチ適用を行うことが不可欠です。さらに、すべての構成要素に対して徹底的なセキュリティ評価を実施することで、攻撃者に悪用される前に潜在的な脆弱性を見つけて対処することが可能になります。



図 2. Wiz Research が Redis に対する UAF 脆弱性を突いて成功した様子

現在インターネット上には 25 万台以上の Redis サーバが公開されていますが、その中でバージョン 8 のサーバに絞ると、およそ 2,000 台にとどまります。これらがベクターストアとして使用されているかどうかを断定するのは難しいものの、使用されているキーの種類などから判断すると、数百台程度はその可能性が高いと見られています。

NVIDIA Container Toolkit のエクスプロイト

Wiz Research は NVIDIA Container Toolkit も標的とし、Trusted 変数の外部初期化に関する脆弱性 (External Initialization of Trusted Variables) を突いたエクスプロイトに成功しました。この攻撃は、すべての外部入力の検証と、重要な変数の安全な初期化がいかに重要かを改めて示しています。また、AI や機械学習 (ML) の導入が進む中で利用が広がっているコンテナ環境において、包括的なセキュリティレビューの必要性も浮き彫りとなりました。同様の脆弱性を防ぐには、厳格な入力検証プロトコルと、コンテナ化された環境に対する定期的なセキュリティ監査が求められます。さらに、最小限のベースイメージの使用やランタイムセキュリティツールの導入など、コンテナ管理におけるセキュリティベストプラクティスの採用も推奨されます。

なお、NVIDIA Container Toolkit の使用状況は Docker や Kubernetes など様々なコンテナ環境に分散しているため、その普及度を観測することはできていません。

その他の調査結果

Ollama は AI モデルサーバであり、2024 年 11 月の時点で認証なしでインターネット上に公開されたサーバが 3,000 台以上確認されていましたが、現在では 1 万台を超えています。Ollama は Pwn2Own の競技対象の 1 つでしたが、誰もエクスプロイトを試みませんでした。すでに 4 件以上の CVE（共通脆弱性識別子）が知られており、ソフトウェア自体の成熟度もまだ低いことから、通常であれば多くの攻撃が試みられてもおかしくありません。しかし、調査したところ、出場者たちは実際には Ollama に対するさまざまなエクスプロイトを保有していたものの、実行には至りませんでした。その理由は、Ollama が非常に頻繁にアップデートを行っているため、高リスクの競技である Pwn2Own では安定性の面から魅力的な標的とは見なされなかったためです。ただし、実際の攻撃者にはこうした制約は存在しないため、Pwn2Own で攻撃が行われなかったことをもって、Ollama が安全だと判断するべきではありません。

また、PostgreSQL 向けの拡張機能である PGVector に対する攻撃も確認されませんでした。PGVector は 2021 年から利用可能になった比較的新しい機能ですが、PostgreSQL 自体は 1970 年代の Ingres を起源とする非常に成熟したオープンソース製品です。そのため、PGVector は比較的新しいとはいえ、非常に堅牢な基盤の上に構築されており、決して未熟なソフトウェアとは言えません。

PART 2 : AI 特有の脆弱性

Pwn2Own Berlin 2025 で発見されたすべての脆弱性は、従来のあらゆる種類のソフトウェアに対して見られてきた攻撃手法に基づくものであり、AI 固有のものではありません。例外は、攻撃対象として設定された AI 関連システムそのものだけです。しかし、AI エージェントに関する脆弱性はすでに顕在化しており、その一例が [CVE-2025-32711](#) です。これは Microsoft 365 Copilot に影響を与えたもので、CVSS スコアは 9.3 と非常に高く、深刻な脆弱性であることを示しています。この脆弱性は AI に対するコマンドインジェクションによって構成されており、悪用されれば攻撃者がネットワーク経由で機密データを盗み出す可能性があります。Microsoft は 2025 年 6 月にこの脆弱性を公表し、パッチを提供しましたが、その影響の大きさは、今後セキュリティ担当者が直面する課題の深刻さを物語っています。トレンドマイクロが実施した「2024 年 Risk to Resilience World Tour Survey」の結果でも、AI セキュリティがセキュリティオペレーションセンター (SOC) チームにとって優先事項になりつつあることが明らかになっており、これらの懸念は現実的かつ差し迫ったものとなっています。

複雑な LLM ベースのアプリケーションへの攻撃

LLM (大規模言語モデル) 単体では、信頼性のある処理に限界があります。しかし、ツールの呼び出しや推論といった複数のステップを組み込んだシステムの一部として使用されることで、その有用性は大きく高まります。現在、LLM はさまざまなアプリケーションに統合されつつあります。こうしたシステムがどのように攻撃され得るかを調査するために、私たちは独自に LLM エージェントシステムを構築しました。というのも、この技術はまだ登場したばかりであり、実際の事例が少ないためです。

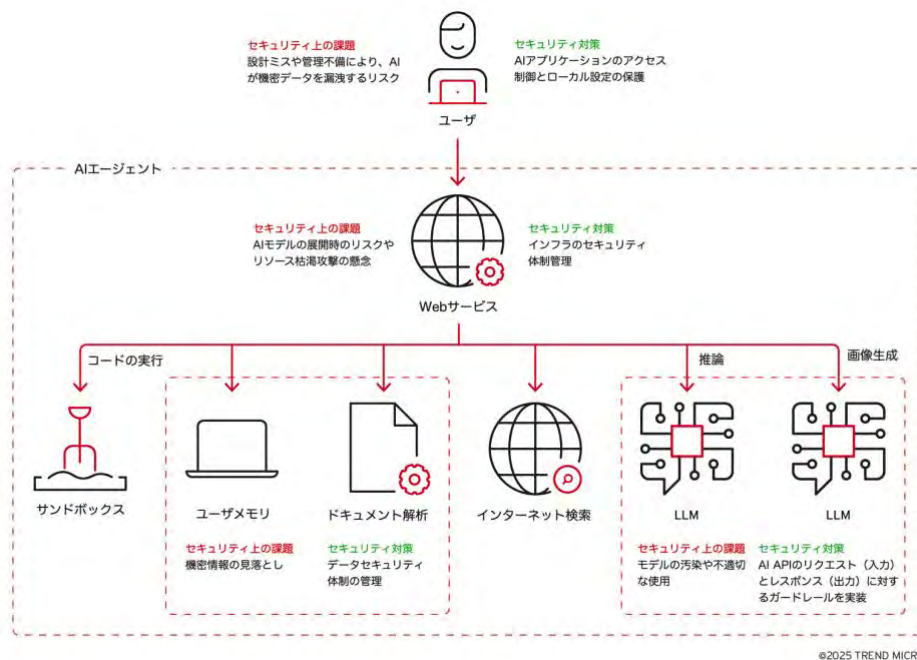


図 3. LLM 駆動の AI エージェントにおける典型的な構成要素と、それぞれのセキュリティ課題および推奨対策

Pandora は、攻撃者が AI エージェントを悪用する新たな手口を調査するために、トレンドマイクロの FTR チームが構築した 概念実証型の AI エージェント です。Pandora は、ChatGPT のような一般的な LLM 機能に加え、Docker 上で動作し、制限のないコード実行が可能な環境が整えられています。この Pandora を用いた研究により、Trend Research は間接的なプロンプトインジェクションを含む複数の攻撃シナリオを明らかにしました。攻撃者は、自身の指示をコンテンツ内に埋め込み、それをエージェントが読み込むよう誘導します。これにより、他のユーザーがアップロードしたチャット履歴やファイルが漏洩する可能性が生じます。たとえば、攻撃者が管理する Web ページの内容をエージェントが読み込む というシナリオでは、ユーザーからは見えないテキストや、攻撃に利用される画像がページに含まれ、最終的に悪意のあるペイロードが LLM プロンプトに挿入されてシステムが侵害される可能性があります。

また、Pandora にはデータベース検索機能も実装されており、これはエージェントが情報を取得するためによく用いられる構成を反映したものです。データには、複数のテーブルやリレーション、ビューを備えた実用的な構造を持つ公開サンプル「Chinook」データベースを使用しました。この環境を用いて、攻撃者が脆弱性を突いて、センシティブなデータを外部に持ち出したり改ざんしたりする可能性を検証しました。これは、ユーザーごとのクエリ制限や堅牢なガードレールを設けていたにもかかわらず発生したものであり、防御をかいくぐる方法が複数存在することを示しています。たとえば、AI エージェントがデータベースクエリを生成する過程で、安全でない SQL 文を生成してしまうことがあります。

さらに、ストアドプロンプトインジェクション攻撃という手口も存在します。これは、AI エージェントが後に読み込み、入力として処理するデータ内に、悪意あるプロンプトやペイロードを仕込むものです。この攻撃は保護機構をすり抜け、データ漏えいからエージェントの振る舞いの改変に至るまで、さまざまな影響を引き起こす可能性があります。

加えて、ベクターデータベースを狙った攻撃も考えられます。誤解を招く情報や有害な情報をベクターストアに注入することで、AI エージェントによるデータ読み込みと同様の結果、つまり誤動作や不正な情報の拡散といった影響を引き起こす恐れがあります。前のセクションで述べたように、ベクターストアが保護されていない場合にはそのまま悪用される可能性があり、仮に保護されていたとしても、脆弱性を突かれて回避される危険性が残っています。

プロンプトベースの攻撃の進化

こうした AI 特有の攻撃の多くは、プロンプト攻撃の手法に依存しており、現在それらはますます高度化しています。DeepSeek-R1 AI モデルは、AI 技術が意図せず攻撃者に新たな機会を提供してしまう一例であり、とりわけ Chain of Thought (CoT) 推論を通じてその傾向が表れています。他のモデルでは、プロンプトによって CoT の表示を促す必要があるのに対し、DeepSeek-R1 はデフォルトで CoT を特別なタグで明示的に表示するという珍しい仕様を持っています。これはユーザーがプロンプトを最適化するのに有用である一方で、攻撃者にとってはモデルの挙動を把握するための手がかりともなり得ます。

業界標準のツールを用いた分析により、このモデルが脱獄やプロンプト攻撃に対して脆弱であることが確認されました。特に CoT が可視化されることで、機密情報の窃取や誤った応答を強要される危険性が高まりました。CoT の存在は、攻撃者にとってより効果的なプロン

プトを作成するための材料となり得るため、エージェント AI の設計においては重要な考慮点となります。

このような悪用の具体例としては、リンクトラップ (Link Trap) があります。これは、攻撃者が GenAI モデルに偽装リンクを含む応答を生成させ、それをユーザーに送信させるという手法です。リンクは一見すると無害な参考資料に見せかけられており、ユーザーがクリックすると、データが攻撃者のサーバーへ送信される仕組みになっています。リンクトラップは、モデルのアクセス権限や外部通信の制限を回避する手段となるため、特に注意が必要です。

さらに、不可視プロンプトインジェクションと呼ばれる手法では、攻撃者が Unicode 文字として悪意ある内容を埋め込み、AI モデルの動作を変更することが可能になります。この種の攻撃は非常に巧妙で、他のプロンプトインジェクションと組み合わせることで強力な効果を発揮します。Unicode 文字に変換された英語テキストはユーザーインターフェース上では表示されないため、開発者はこのような不可視テキストを AI アプリケーションの入力として受け入れないようにし、ナレッジベースに使用するデータの内容を慎重に検証することが求められます。

「Do Anything Now (DAN)」や「前の指示を無視する」といった単純なプロンプト操作は、より高度な手法へと置き換わりつつあります。たとえば プロンプトリーク (Prompt Leakage, PLeak) 攻撃では、LLM のセキュリティ制限を回避して、システムプロンプトやファインチューニングの内容を漏洩させることが可能です。トレンドマイクロの調査では、この攻撃手法をもとに洗練された敵対的な文字列が、オープンソース・商用問わずさまざまな LLM のシステムプロンプトを脱獄するのに有効であることが確認されました。

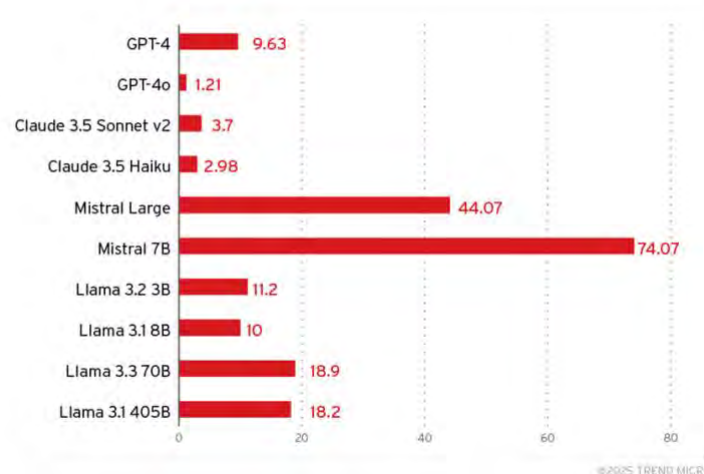


図 4. 主要な LLM モデルおよびサービスにおける PLeak 攻撃の成功率

プロンプト攻撃に対抗するためには、厳格な出力フィルタリングや定期的なレッドチーム演習が有効とされていますが、AI 技術の進化と並行してこれらの脅威も発展している現状においては、AI 開発全体に対するより大きな課題が浮かび上がっています。つまり、AI システムに対する信頼を育みつつ、セキュリティを確保するという微妙なバランスをどのように実現するかが問われているのです。

トレーニングデータの質の低さや保護機能の欠如といった内部的なミスも、AI が意図しない動作や不適切な応答をする原因となりますが、それ以上に、攻撃者がこうした状況を巧みに利用して自らの目的に活用する余地があることも明らかになってきています。彼らがこのような手法を使うかどうかは、自身の作戦上の欠落をどのように AI で補えるか、そして最も重要なのは、そのリスクに見合う価値があるかどうかにかかっています。

PART 3：サイバー犯罪における AI の活用

複雑な LLM ベースのアプリケーションへの攻撃

AI が保護すべき重要なインフラの一部になりつつある一方で、現在のサイバー犯罪を強化したり、新たな手口を模索したりするための基本的なツールとして、AI が犯罪者に利用されつつあることにも注目すべきです。

	犯罪に悪用される LLM	ディープフェイクの悪用
個人ユーザ	<ul style="list-style-type: none">- 海外の被害者を標的とする攻撃- フィッシングキャンペーン- 恋愛詐欺のスケール拡大- ビジネスメール詐欺 (BEC)	<ul style="list-style-type: none">- 仮想誘拐- セクストーション (性的脅迫)- 偽の広告- 恋愛詐欺- 児童ポルノ- 海外で足止めされた旅行者の偽装
法人	<ul style="list-style-type: none">- ビジネスメール詐欺 (BEC)	<ul style="list-style-type: none">- 経営幹部のなりすまし- KYC (本人確認) 回避- ビジネスメール詐欺 (BEC)- 採用詐欺

表 1. ジェネレーティブ AI の犯罪的用途の概要

AI による自動翻訳の戦略的優位性

ロシア語圏のサイバー犯罪者にとって、AI の有用性の多くは自動翻訳と文化的文脈の補完にあります。これにより、より信憑性のある誘導文を作成し、従来はアクセスできなかった地域の被害者にも攻撃の手を広げることが可能になりました。生成 AI の活用に加え、漏洩した生体認証データや二重恐喝型ランサムウェアによって流出した情報を組み合わせることで、暗号資産詐欺や恐喝行為を高度化させるためのデジタル ID を生成することもできます。

AI ツールはまた、大規模な攻撃キャンペーンの実行を容易にし、何百万もの被害者から少額ずつを抜き取るといった分散的な手法によって、法執行機関の目を逃れることにもつながっています。言語の壁に制限されない犯罪者は、防御側による言語ベースの帰属特定を妨げることができ、異なる言語の協力者との連携も容易になります。トレンドマイクロのリサーチによると、ロシア語圏のサイバー犯罪領域では、異なる言語や文化的背景を持つグループ間での協力が増加しており、犯罪歴はないが国家的な利害関係とつながりのある人物も含まれていることが確認されています。

攻撃者による AI 採用の選択性

「犯罪用 GPT」の広告も地下フォーラムで出回っています。WormGPT は 2023 年に登場したその一例で、地下マーケットで堂々と販売されていましたが、開発者の身元が広く知られていたことから、メディアやサイバーセキュリティ業界の注目を浴び、数か月で閉鎖されました。商用 AI よりも匿名性を提供するとされながらも、これらの犯罪向け LLM の採用は限定的でした。その多くは違法行為に特化して訓練されたわけではなく、BlackhatGPT のように商用 AI プラットフォームを脱獄 (ジェイルブレイク) しただけのものがほとんどです。EscapeGPT や LoopGPT といったモデルはその偽装すらなく、既存の GenAI プラットフォ

ームを脱獄したバージョンで、プライバシー保護をうたっていると明言しています。トレンドマイクロが調査した Ollama サーバー上では、これらのモデルは一切確認されませんでした。

「GPT」の名を冠して注目を集めようとしたツールが失敗する一方で、「ジェイルブレイク・アズ・ア・サービス」は地下フォーラムで人気を集めつつあります。有料で、最新の脱獄手法を利用して商用 LLM に組み込まれている倫理的な制限を回避し、制限のない回答を生成できるようにするというサービスです。

生産性の向上や言語・文化の壁を越えた翻訳の可能性とは別に、犯罪者は一般的に新技術の導入には慎重です。職業柄、非常にリスク回避的であり、導入するとしても急激な変化よりも段階的な変化を選びます。新技術によって明確な利益が得られると確信しない限り、武器や作戦の中核を刷新することはほとんどありません。そのため、たとえばエージェンティック AI の採用は、現時点ではまだ初期段階にとどまっています。

ディープフェイクがサイバー犯罪の参入障壁を下げた

現在、他のどの AI 技術よりも大きな影響を与えているのは、おそらくディープフェイクでしょう。実際に、トレンドマイクロの調査によると消費者の 36% がディープフェイクを使った詐欺の試みを受けたと報告しています。悪意のある手に渡れば、ディープフェイクは一般消費者を狙う大規模詐欺にも、特定企業を標的とした精密な攻撃にも利用される可能性があります。

犯罪者たちは、カスタム地下サービスから主流のディープフェイク作成プラットフォームへと移行する傾向を強めています。その理由は明白です。こうしたサービスは、リアルタイムの映像操作、多言語の音声クローン、画像のヌーディファイなどの高度な機能を低価格（中には無料）で提供しているからです。犯罪者向けの地下コミュニティでは、ディープフェイクを使った攻撃のための使いやすいツールやチュートリアル、サポートが豊富に出回っており、技術的スキルの乏しい者でも簡単に悪用できる環境が整っています。

こうした正規サービスの悪用に対する対策としては、利用履歴の追跡やウォーターマーキングなどが挙げられます。特にウォーターマーキングは、AI 生成コンテンツの識別に一定の効果がありますが、改ざんのリスクが残っており、信頼性のある標準化が難しいという課題も、ブルッキングス研究所などによって指摘されています。

eKYC 回避攻撃における AI 対決

ディープフェイク技術が犯罪者にとって重要な用途を持つのは、電子的な本人確認 (eKYC) システムを回避する手段としての役割です。これは主に、暗号資産プラットフォームで匿名アカウントを作成し、マネーロンダリングを行う目的で利用されます。攻撃者は AI によって生成したディープフェイク画像や動画を用いて、eKYC 側の AI モデルを欺くことが可能となり、AI 同士の対決が実質的に発生しているのです。

eKYC の認証モデルは、ユーザーのさまざまな端末に対応しつつ帯域や容量を節約するため、低解像度での顔の検出や向きを確認を行うことが多く、これにより基本的なラップトップ用 GPU と、Deepfake Offensive Toolkit (DoT) や Deep-Live-Cam といったオープンソース

ツールだけで、初期スキャンをすり抜けることが可能になります。成功には複数回の試行が必要な場合もありますが、それでも突破可能であるという事実が確認されています。

このような eKYC 回避手法は、悪意ある人物にとってはほぼ「簡略化された工程」となっており、その背景には、地下マーケットが幅広い攻撃者に対応したサービスを提供している実態があります。攻撃を自分で実行するためのオンラインディープフェイクサービスを利用する場合もあれば、US\$30 から US\$600 の価格帯で「バイパス・アズ・ア・サービス」を利用することも可能です。

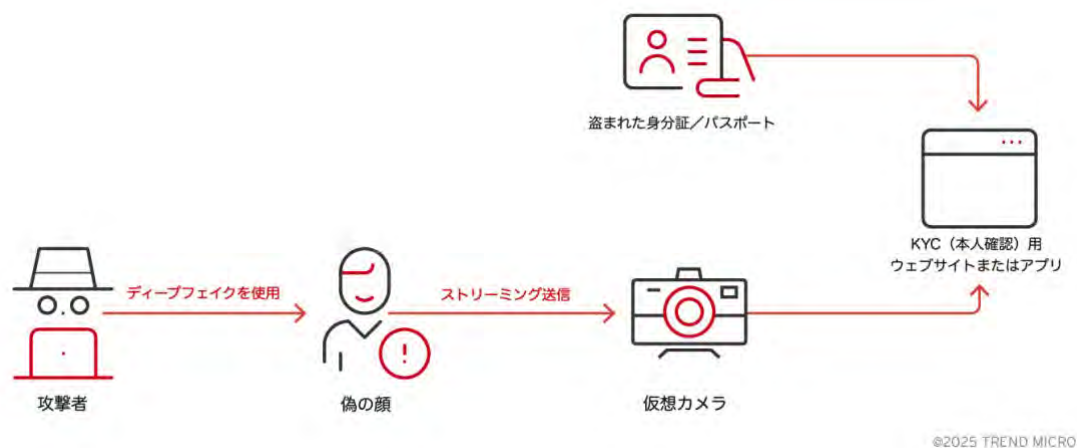


図 5. 一般的な eKYC バイパス手順

この問題に立ち向かうには、意識の転換が必要です。すべてのデジタルメディアを疑ってかかるゼロトラストの姿勢を持つことが、AI 強化された犯罪の武器を使いこなすベテラン・新興の攻撃者の双方に対抗するための戦略を構築する上で不可欠です。

PART 4：これからの道筋

エージェントAIの複雑化が進む中で

現在、市場には「エージェントAI」と称するソリューションが多数登場していますが、その多くは、エージェントAIを基本的な自動化と差別化するために必要な適応性や自律性を完全には備えていません。真のエージェントAIシステムは、目標志向であり、文脈を理解し、行動に基づいて駆動する能力を持ち、複数のステップを伴う推論や、新たなデータと経験に基づく自己学習も可能です。

このようなエージェントAIに固有のセキュリティ課題に対応するには、その本質的な特性を正しく理解することが不可欠です。大規模言語モデル（LLM）がテキストやメディアの生成に重点を置いているのに対し、AIエージェントはツールへ自律的にアクセスし、より複雑なタスクを実行できるため、まったく異なる次元のリスクが存在します。不正行為が発生した場合、複数のタスクや機能が連鎖的に影響を受ける可能性があり、初期段階ではその兆候を見逃してしまうこともあります。

エージェントAIの成熟とともに、現行の構成を超えて、サードパーティ製のツールマーケットプレイスやエージェントリポジトリなどのコンポーネントを組み込み、その柔軟性をさらに高めていくと予測されます。一方で、これにより新たなサプライチェーン攻撃のリスクも生まれます。

さらに、AIモデル、API、データストアなどのコンポーネントが不適切にパブリックネットワークに公開されている場合、攻撃者による機密データや主要機能の窃取・改ざんにつながる恐れがあります。AIエージェントは外部リソースとの頻繁なインタラクションを伴うため、従来の境界型防御のアプローチは、このように複雑化した環境ではもはや有効ではありません。

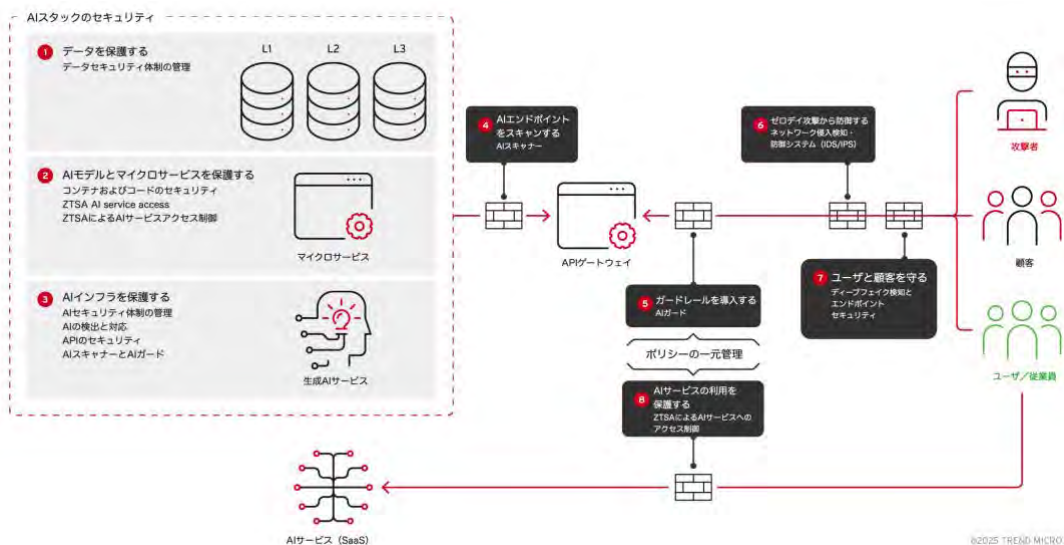


図 6. AI スタックを保護するためのステップ

将来的には、AI エージェントが自らのエージェントを作成することで、自律性をさらに拡張する可能性もあります。そのような時代には、ユーザーが AI システムを監督できる信頼構築メカニズムが不可欠となります。こうした未来に備え、トレンドマイクロのリサーチチームは、複数の接続された AI エージェントがユーザー定義の目標に対して協調して働くというアーキテクチャの概念「エージェントティックメッシュ」の進展を積極的に監視しています。

デジタルアシスタントによる支援

エンドユーザーが AI エージェントと直接やり取りするようになるとは、近い将来には考えていません。むしろ、エージェントティック AI を活用した一般向けアプリケーションの多くは、次世代のデジタルアシスタント (DA) によって制御され、その背後で複数の AI エージェントが連携しながら複雑なタスクを処理し、それぞれのエージェントが特定の役割を担う構成になると予想されます。次世代の AI 駆動型 DA は、大量のユーザー情報を扱い、これまでにない形でその情報とインタラクションできるようになるため、セキュリティチームにとっては、この領域の将来性を見据えた防御策の構築が早急に求められるようになります。

こうした新しいタイプのアシスタントが直面するセキュリティ脅威は、他の AI 技術と同様に、その進化と密接に関係しています。サードパーティ製のスキルやプラグインを利用することで攻撃対象の範囲も広がり、悪意あるコードが正規の拡張機能として偽装されることで、ユーザーに気づかれずに盗聴、情報窃取、さらには DA の乗っ取りが行われる可能性があります。さらに、今後エンドユーザーによる DA への依存が進めば、その傾向を逆手に取り、フィッシングやソーシャルエンジニアリング攻撃を DA に対して直接仕掛けるといった新たな手法も登場するかもしれません。

結論：AI セキュリティの最前線に立つベンダ ーたち

AI の進歩は必然的に攻撃対象領域を広げてしまうため、脆弱性や新たな攻撃経路が表面化する前に先回りするためには、AI のライフサイクル全体にセキュリティ対策を組み込むことが不可欠です。エージェンティックな推論は、AI およびサイバーセキュリティの双方における次なるフロンティアとなる多くのパラダイム転換の一例にすぎません。今年の NVIDIA グローバル・テクノロジー・カンファレンス (GTC) でも議論されたように、エンタープライズグレードの AI 向け「AI ファクトリー」の構築基盤はすでに整いつつあります。たとえば NVIDIA と連携した Dell AI Factory では、企業が AI 環境のトレーニング、導入、チューニングを効率的に行えるよう支援しています。さらに、NVIDIA の CUDA-Q 開発プラットフォームや Accelerated Quantum Research Center の設立は、量子耐性暗号、量子と古典のハイブリッド演算アルゴリズム、そして将来の量子コンピューティング応用を見据えた研究の推進に貢献しています。トレンドマイクロの研究チームも、暗号システム設計と実装における重要要素、量子コンピュータが解決可能な複雑な問題の種類、そして量子耐性アルゴリズムが現在の暗号規格に取って代わる可能性についての知見を提供しています。量子機械学習 (Quantum ML) はまだ発展途上にあるものの、現段階からその可能性と限界を理解しておくことは有益です。

AI を活用したサイバーセキュリティの最前線に立っているのが、Trend Cybertron です。これは 2025 年 3 月に発表された AI モデルおよびエージェントで、サイバー脅威に対して自律的に対応できるよう設計されています。生成 AI を活用して複雑な環境を分析し、インシデントの分析や意思決定を自動化、攻撃に対する組織の対応を調整することが可能です。

Cybertron は、トレンドマイクロの Agentic AI サイバーセキュリティプラットフォームである Trend Vision One™ に完全統合されているほか、その多くの構成要素がオープンソースとして提供されています。Trend Cybertron の「サイバーブレイン」に相当する部分を一般に公開することで、トレンドマイクロは世界中のセキュリティコミュニティとの連携を促進し、その能力を共同で強化しながら脅威インテリジェンスを共有しています。このようにして Cybertron は、開発者や研究者にとっても貴重なリソースとなり、学習データセット、サイバーセキュリティに特化した LLM、クラウドリスク評価 AI エージェント (Trend's Cloud Risk Assessment AI Agent) といったツールにアクセスする手段を提供しています。

こうしたリソースは、実装の現実に根差し、AI の進化の方向性を導くための広範なロードマップの一部として活用されるのが最も効果的です。リスクの認識と具体的な行動とのギャップを埋めるため、トレンドマイクロは Security for AI Blueprint (AI 環境向けセキュリティの目指すべき姿) (AI Security Blueprint) を公開しています。これは AI システムを堅牢化するためのアーキテクチャ上の推奨事項を提示するもので、データと AI ワークフローの運用的完全性の両方を守るための多層的なセキュリティメカニズムが含まれています。これにより、防御側はエッジ、クラウド、データセンターといった技術領域をまたぐ課題に対して、より備えのある状態で対応できるようになります。

このプロアクティブセキュリティ姿勢を支えるもう一つの取り組みとして、トレンドマイクロは2024年12月に、AI搭載デジタルアシスタント（DA）システムにおける脅威シナリオを予測するためのフレームワークを発表しました。DAが今後さらに多機能化し、他のデバイスと高度に接続されていく中で、その進化の追跡——特にユーザーが知覚する機能、インタラクション関連の能力、および本質的な性能の進展——を行うことは、セキュリティチームが早期に盲点を見逃されるリスクに対処する上で重要です。

AIエコシステムを堅牢なものにしていくには、専門知識と支援を結集できる業界リーダー間の継続的な連携も不可欠です。この目的のもと、トレンドマイクロは最近、MITRE ATLAS に対して初のケーススタディを提出しました。これはクラウドおよびコンテナベースのAIインフラに対する攻撃を扱ったもので、2024年8月にはSecure AIを推進する「Coalition for Secure AI (CoSAI)」のスポンサーも務めました。トレンドマイクロのデジタルセーフティに対する長年の取り組みは、AIシステムを守るための業界横断的な協力体制、標準化、ベストプラクティスの策定を進めるCoSAIの方針と深く一致しています。トレンドマイクロは、AIサプライチェーンの保護、AI防御者の育成、AIのリスクガバナンスの強化、エージェンティックAIの新たなセキュリティモデルの開発といった分野で、CoSAIに積極的に貢献しています。これらすべてのテーマにおける背景調査への貢献に加え、2025年上半期には業界メンバーと連携し、モデル署名、機械可読モデルカード、AIシステムに対するインシデント対応、AI向けのゼロトラスト実装、MCP（モデル制御ポリシー）セキュリティといった具体的なソリューションの開発にも取り組んでいます。

これらの取り組みは、組織が必要な投資の優先順位を明確にし、適切な緩和策を導入し、AIを利用したサイバー脅威に対する既存の防御体制をさらに強化するための実践的な指針となります。トレンドマイクロの研究者たちは、脅威とAIに関する数十年にわたる知見を活かし、CoSAIがこれらのフレームワークやツールを継続的に更新・維持していく上で不可欠な技術的支援を提供しています。これにより、防御の最前線に立つ人々が、AIへの依存が高まるシステム環境でも戦い続けることが可能になります。

防御側も攻撃側も、AIを力の増幅装置として活用しています。AIはビジネスの推進力となる一方で、サイバー犯罪者の能力を著しく高める手段にもなっています。この現実を踏まえ、組織は攻撃者より一歩先に行く必要があります。そのためには、すべての層でセキュリティを組み込み、AIシステムを厳密に評価・強化していくことが重要です。今後、より高度で自律的なAIが登場するにつれて、AIライフサイクル全体にわたってセキュリティを最優先に考える姿勢が求められます。多層防御、継続的な学習、そして業界横断的な協力体制によって、ビジネスリーダーやサイバーセキュリティの専門家、変化に強く柔軟なAIエコシステムを構築し、未知の脅威を上回る形で未来のAIを守ることが可能になります。

参考文献

1. Hannah Mayer, Lareina Yee, Michael Chui, and Roger Roberts. (Jan. 28, 2025). *McKinsey & Company*. "Superagency in the workplace: Empowering people to unlock AI's full potential." Accessed May 5, 2025, at: [Link](#).
2. Security Staff. (April 26, 2024). *Security Magazine*. "93% of security leaders anticipate daily AI attacks by 2025." Accessed May 5, 2025, at: [Link](#).
3. World Economic Forum. (Jan. 13, 2025). *World Economic Forum*. "Global Cybersecurity Outlook 2025." Accessed May 12, 2025, at: [Link](#).
4. Trend Micro staff. (June 18, 2025). *Trend Micro*. "What Is AI?" Accessed July 11, 2025, at: [Link](#).
5. Russ Meyers. (May 13, 2025). *Trend Micro*. "Trend Micro Puts a Spotlight on AI at Pwn2Own Berlin." Accessed 16, 2025, at: [Link](#).
6. Trend Micro staff. (June 18, 2025). *Trend Micro*. "What is Agentic AI?" Accessed July 11, 2025, at: [Link](#).
7. Dustin Childs. (Feb. 24, 2025). *Trend Zero Day Initiative*. "Announcing Pwn2Own Berlin and Introducing an AI Category." Accessed May 7, 2025, at: [Link](#).
8. Dustin Childs. (May 17, 2025). *Trend Zero Day Initiative*. "Pwn2Own Berlin 2025: Day Three Results." Accessed May 19, 2025, at: [Link](#).
9. Morton Swimmer, Philippe Lin, Vincenzo Ciancaglini, Marco Balduzzi, and Stephen Hilt. (Dec. 4, 2024). *Trend Micro*. "The Road to Agentic AI: Exposed Foundations." Accessed May 13, 2025, at : [Link](#).
10. CVE. (June 11, 2025). *CVE*. "CVE-2025-32711." Accessed June 24, 2025, at: [Link](#).
11. Microsoft. (June 11, 2025). *Microsoft*. "M365 Copilot Information Disclosure Vulnerability." Accessed June 24, 2025, at: [Link](#).
12. Trend Micro. (Nov. 4, 2024). *Trend Micro*. "SOC Around the Clock: World Tour Survey Findings." Accessed May 6, 2025, at: [Link](#).
13. Sean Park. (April 22, 2025). *Trend Micro*. "Unveiling AI Agent Vulnerabilities Part I: Introduction to AI Agent Vulnerabilities." Accessed May 9, 2025, at: [Link](#).
14. Sean Park. (May 13, 2025). *Trend Micro*. "Unveiling AI Agent Vulnerabilities Part III: Data Exfiltration." Accessed May 16, 2025, at: [Link](#).
15. Sean Park. (May 21, 2025). *Trend Micro*. "Unveiling AI Agent Vulnerabilities Part IV: Database Access Vulnerabilities." Accessed May 23, 2025, at: [Link](#).
16. Trent Holmes and Willem Gooderham. (March 4, 2025). *Trend Micro*. "Exploiting DeepSeek-R1: Breaking Down Chain of Thought Security." Accessed May 9, 2025, at: [Link](#).
17. Jay Liao. (Dec. 17, 2024). *Trend Micro*. "Link Trap: GenAI Prompt Injection Attack." Accessed May 20, 2025, at: [Link](#).
18. Ian Ch Liu. (Jan. 22, 2025). *Trend Micro*. "Invisible Prompt Injection: A Threat to AI Security." Accessed July 9, 2025, at: [Link](#).
19. Karanjot Singh Saggu and Anurag Das. (May 1, 2025). *Trend Micro*. "Exploring PLeak: An Algorithmic Method for System Prompt Leakage." Accessed May 9, 2025, at: [Link](#).
20. AI Team. (Sep. 3, 2024). *Trend Micro*. "How AI Goes Rogue." Accessed May 24, 2025, at: [Link](#).
21. Trend Micro. (April 8, 2025). *Trend Micro*. "The Russian-Speaking Underground." Accessed May 22, 2025, at: [Link](#).
22. Vincenzo Ciancaglini and David Sancho. (May 8, 2024). *Trend Micro*. "Back to the Hype: An Update on How Cybercriminals Are Using GenAI." Accessed May 19, 2025, at: [Link](#).
23. Trend Micro. (April 24, 2025). *Trend Micro*. "How Agentic AI Is Powering the Next Wave of Cybercrime | #TrendTalksAI." Accessed May 13, 2025, at: [Link](#).
24. David Sancho, Salvatore Gariuolo, and Vincenzo Ciancaglini. (July 9, 2025). *Trend Micro*. "Deepfake It Till You Make It: A Comprehensive View of the New AI Criminal Toolset." Accessed July 9, 2025, at: [Link](#).
25. Trend Micro. (July 30, 2024). *Trend Micro*. "Trend Micro Stops Deepfakes and AI-Based Cyberattacks for Consumers and Enterprises." Accessed May 19, 2025, at: [Link](#).
26. David Sancho and Vincenzo Ciancaglini. (July 30, 2024). *Trend Micro*. "Surging Hype: An Update on the Rising Abuse of GenAI." Accessed May 19, 2025, at: [Link](#).

27. Trend Micro. (July 30, 2024). *Trend Micro*. "AI-Powered Deepfake Tools Becoming More Accessible Than Ever." Accessed May 19, 2025, at: [Link](#).
28. Siddarth Srinivasan. (Jan. 4, 2024). *The Brookings Institution*. "Detecting AI fingerprints: A guide to watermarking and beyond." Accessed May 19, 2025, at: [Link](#).
29. Philippe Lin, Fernando Mercês, Roel Reyes, and Ryan Flores. (Nov. 28, 2024). *Trend Micro*. "AI vs AI: DeepFakes and eKYC." Accessed May 13, 2025, at: [Link](#).
30. Salvatore Gariuolo and Vincenzo Ciancaglini. (June 18, 2025). *Trend Micro*. "The Road to Agentic AI: Defining a New Paradigm for Technology and Cybersecurity." Accessed June 29, 2025, at: [Link](#).
31. AI Team. (Dec. 1, 2024). *Trend Micro*. "AI Pulse: The Good from AI and the Promise of Agentic." Accessed May 9, 2025, at: [Link](#).
32. Vincenzo Ciancaglini, Salvatore Gariuolo, Stephen Hilt, Robert McArdle, and Rainer Vosseler. (Dec. 6, 2024). *Trend Micro*. "AI Assistants in the Future: Security Concerns and Risk Management." Accessed May 9, 2025, at: [Link](#).
33. Shannon Murphy. (April 7, 2025). *Trend Micro*. "GTC 2025: AI, Security & The New Blueprint." Accessed May 13, 2025, at: [Link](#).
34. Dell Technologies. (March 18, 2024). *Dell Technologies*. "Dell Offers Complete NVIDIA-Powered AI Factory Solutions to Help Global Enterprises Accelerate AI Adoption." Accessed May 9, 2025, at: [Link](#).
35. Nicholas Harrigan. (March 18, 2025). *NVIDIA*. "NVIDIA Accelerated Quantum Research Center to Bring Quantum Computing Closer." Accessed May 10, 2025, at: [Link](#).
36. Morton Swimmer, Mark Chimley, and Adam Tuaima. (Sep. 12, 2023). *Trend Micro*. "Diving Deep Into Quantum Computing: Modern Cryptography." Accessed June 13, 2025, at: [Link](#).
37. Morton Swimmer, Mark Chimley, and Adam Tuaima. (Jan. 18, 2024). *Trend Micro*. "Diving Deep Into Quantum Computing: Computing With Quantum Mechanics." Accessed June 13, 2025, at: [Link](#).
38. Morton Swimmer, Mark Chimley, and Adam Tuaima. (July 4, 2024). *Trend Micro*. "Post-Quantum Cryptography: Migrating to Quantum Resistant Cryptography." Accessed June 13, 2025, at: [Link](#).
39. Morton Swimmer. (Oct. 28, 2024). *Trend Micro*. "The Realities of Quantum Machine Learning." Accessed June 13, 2025, at: [Link](#).
40. Trend Micro. (March 19, 2025). *Trend Micro*. "Trend Micro to Open-source AI Model and Agent to Drive the Future of Agentic Cybersecurity." Accessed May 21, 2025, at: [Link](#).
41. Dave McDuff. (March 27, 2025). *Trend Micro*. "Trend Cybertron: Full Platform or Open-Source?" Accessed May 21, 2025, at: [Link](#).
42. Trend Micro. (n.d.). *GitHub*. "Trend Cybertron - Cloud Risk Assessment Agent." Accessed May 21, 2025, at: [Link](#).
43. Fernando Cardoso. (n.d.). *Trend Micro*. "Security for AI Blueprint." Accessed May 19, 2025, at: [Link](#).
44. Trend Micro staff. (June 18, 2025). *Trend Micro*. "What is Proactive Security?" Accessed July 11, 2025, at: [Link](#).
45. Alfredo Oliveira. (May 27, 2025). *Trend Micro*. "Trend Micro Leading the Fight to Secure AI." Accessed June 12, 2025, at: [Link](#).
46. Trend Micro. (Aug. 6, 2024). *Trend Micro*. "Trend Micro Expands Partnership Focus to Secure Enterprise AI Use." Accessed May 25, 2025, at: [Link](#).
47. Trend Micro. (May 23, 2025). *Trend Micro*. "Frameworks for Safer AI with Josiah Hagen // #TrendTalksLife." Accessed May 25, 2025, at: [Link](#).

TREND MICRO

本書に関する著作権は、トレンドマイクロ株式会社へ独占的に帰属します。

トレンドマイクロ株式会社が書面により事前に承諾している場合を除き、形態および手段を問わず本書またはその一部を複製することは禁じられています。本書の作成にあたっては細心の注意を払っていますが、本書の記述に誤りや欠落があってもトレンドマイクロ株式会社はいかなる責任も負わないものとします。本書およびその記述内容は予告なしに変更される場合があります。

本書に記載されている各社の社名、製品名、およびサービス名は、各社の商標または登録商標です。

〒160-0022

東京都新宿区新宿 4-1-6 JR 新宿ミライナタワー

<https://www.trendmicro.com>

トレンドマイクロはサイバーセキュリティのグローバルリーダーとしてデジタル情報を安全に交換できる世界の実現に貢献します。私たちの革新的なソリューションはデータセンター、クラウド、ネットワーク、エンドポイントにおける多層的なセキュリティをお客様に提供します。

当社のリーダーシップの根幹であるトレンドマイクロリサーチは、多くのエキスパートに支えられています。それは最新の脅威を発見し、重要なインサイトを公に共有し、サイバー犯罪の防止を支援することに情熱を注ぐ人材です。当社のグローバルチームは、日に数百万もの脅威を特定し、脆弱性の開示を先導し、標的型攻撃・AI・IoT・サイバー犯罪等における革新的な研究結果を公表しています。私たちは次に来る脅威を予測し、セキュリティ業界が進むべき方向を示しうる示唆に富んだ研究成果を提供するため、継続的に取り組んでまいります。



Trend Micro
Research

© 2025 Trend Micro Incorporated. All Rights Reserved.