


# AI セキュリティは ここから始まる

すべての組織が知るべき実践と禁止事項

Written by: Dave McDuff and Andre Fernandes

OCTOBER 14, 2025



はじめに：AIセキュリティが経営レベルの最優先事項となった理由 3

AIセキュリティのビジネス上の意義 4

設計段階からのAIセキュリティ：実施すべきことと避けるべきこと 6

人とガバナンス：AIセキュリティの人的側面 7

AIフレームワークとコンプライアンス：先を行くか、取り残されるか 8

AI脅威への将来備え 10

結論：戦略的な拡張力としてのAIセキュリティと競争優位性 12

## はじめに：AI セキュリティが経営層の最優先事項となった理由

次の競争優位の波は、より速いコードやより大きなデータセットから生まれるのではなく、安全性、倫理性、そして法令順守を確保できる信頼性の高い AI システムから生まれます。

生成 AI が製品設計を加速させ、自律型 AI エージェントが業務プロセスを実行することで、効率性と革新性の向上はかつてない規模で進んでいます。その一方で、次のような深刻なリスクも拡大しています。



- プロンプトインジェクションによる操作
- 規制対象や機密情報のデータ漏えい
- 学習データや検索データセットの汚染
- ディープフェイクを用いた詐欺や偽情報の拡散
- 事前学習済みモデルや AI API に潜むサプライチェーンの脆弱性
- AI システムが参照する外部データソースに対する悪意ある SEO 汚染 (Agentic Malicious SEO)

調査機関 Forrester のレポート「The Top Cybersecurity Threats in 2025」<sup>1</sup>によると、「DeepSeek が有害なコンテンツを生成するために行ったテストのうち 45%が安全プロトコルを回避した」と報告されています。この結果は、AI モデルのガードレール(防御機構)に重大な弱点が存在することを示しており、攻撃者がセキュリティの甘い AI モデルを容易に悪用できる危険性を浮き彫りにしています。さらに、EU AI 法、中国の「生成 AI サービス管理暫定弁法」、米国の州レベルでの AI 関連法など、規制の強化が進んでおり、法令順守を怠ることがユーザーや機密データを重大なリスクにさらす可能性が高まっています。

結論は明確です。AI セキュリティはもはや選択肢ではありません。セキュリティを組み込まずに AI を構築することは、基礎のない高層ビルを建てるようなものです。短期間で形にはなるものの、極めて不安定です。初日から適切に設計することが、AI を脆弱な実験から持続的な競争優位へと変える鍵となります。

## AI セキュリティのビジネス上の意義

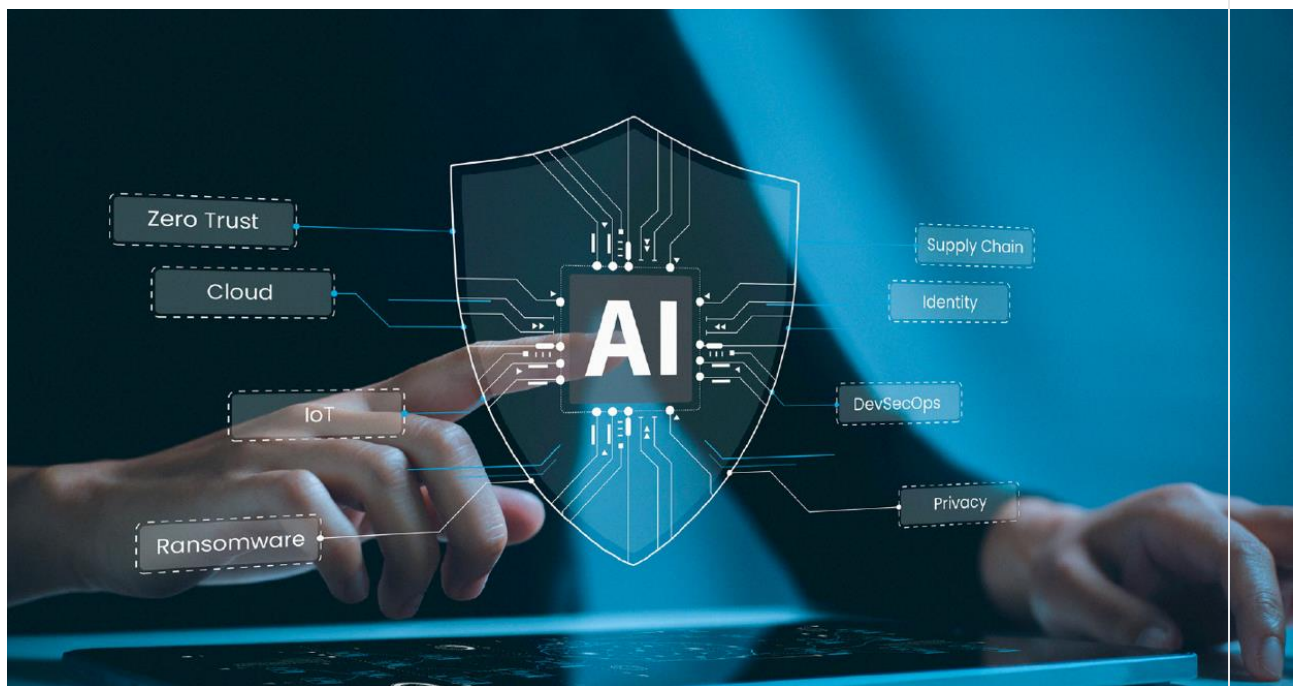
AI プロジェクトの初期段階からセキュリティを設計に組み込むことで、組織は被害を回避するだけでなく、競争優位を獲得することができます。

### イノベーションの加速とリスク低減

- 開発初日からセキュリティを統合することで、高額な手戻り対応、侵害対応、コンプライアンスによる遅延を防ぐことができます。DeepSeekのテストでは、「78%が安全プロトコルを回避して悪意のあるコードや安全でないコードを生成した」と報告されており、最先端のシステムであっても、厳密なセキュリティ設計がなければ脆弱になり得ることを示しています。モデルのテストと適切なガードレールの早期導入は、リスクを抑えるだけでなく、プロジェクトを予定通り進行させ、競合よりも先行するための重要な要素となります。

### 予防による TCO（総保有コスト）の削減

- AI に関するセキュリティインシデントは、システム構築そのもののコストを瞬時に上回る可能性があります。2024 年には、「ディープフェイクによって作成された“CFO”が、財務担当者に 2500 万ドルを詐欺口座へ送金させる」という事件が発生しました。敵対的テスト、モデルの監視、ユーザ認証などの予防策は、重大な侵害からの復旧費用に比べ、はるかに低コストで実施できます。



## 顧客の信頼を差別化要素とする

- あらゆる AI 製品の成功は、ユーザの信頼に支えられています。現在では、画像改ざん、音声クローン、精巧な偽動画といった高品質なディープフェイクが、顧客、パートナー、従業員を対象とする認証システムを侵食しています。透明性のある AI セキュリティの実践を導入し、それを明示することで、組織は関係者の安心感を高め、導入の障壁を下げ、ブランドへの忠誠心を強化することができます。

## 組み込み型の法令遵守

- Forrester の「Business Risk Survey 2024」によると、「企業リスク管理の意思決定者の 61%が、2025 年には法令遵守のための支出が増加すると予想している」と報告されており、変化する法律や基準への対応に組織が直面する圧力の高まりが示されています。
  - このような規制リスクを効果的に管理するために、組織は NIST AI RMF、OWASP Top 10 for LLM Applications 2025、CSA AI Control Matrix といった AI サイバーセキュリティフレームワークを採用することができます。これらはリスクと統制管理のための実践的な指針を提供します。
  - 国際規格である ISO/IEC 42001:2023 は、AI マネジメントシステムに関する正式な要件を定めており、認証の取得を可能にします。
- これらのフレームワークや標準を整合させることで、組織は EU AI 法などの規制により適切に対応できるようになり、罰金、製品修正、または市場での制限といったリスクを軽減しながら、AI の取り組みをグローバルに展開できるようになります。

## 将来への備え

- AI 関連の脅威は急速かつ低コストで進化しています。Forrester の報告によれば、「詐欺師は 5000 ドル未満で音声や映像のディープフェイク‘パペット(操り人形)’を作成・操作できる」とされており、オープンソースのアルゴリズム、安価な GPU、容易に入手できる音声データの普及がその背景にあります。継続的に脅威インテリジェンスを更新する AI セキュリティプラットフォームを導入することで、新たな攻撃手法への適応力を高め、攻撃者の戦術が変化した際に発生する高額な再設計コストを抑えることができます。

# 設計段階からの AI セキュリティ： 実施すべきことと避けるべきこと

領域 (Domain)	実施すべきこと (Do)	避けるべきこと (Don't)	ビジネス上の効果 (Business Impact)
戦略と設計 (Strategy and design)	プロジェクトの初期段階から AI 脅威モデリング、コンプライアスマッピング、ゼロトラストアーキテクチャを統合すること。  AI サイバーセキュリティフレームワークに準拠し、組織の利用方針とガバナンスポリシーを策定すること。	監視メカニズムや「キルスイッチ(緊急停止機能)」を持たない AI を構築すること。	再設計や無秩序な AI 利用を防止できる。
サプライチェーンセキュリティ (Supply chain security)	モデル、データセット、パッケージ、オープンソースライブラリ、API に関する完全なソフトウェア部品表 (SBOM) を維持し、透明性と追跡可能性を確保すること。	バイアスやデータ汚染の検証を行わず、未検証のモデルやデータセットを使用すること。	データ汚染、知的財産の盗用、隠れた脆弱性を防止できる。
アクセスと制御 (Access and control)	最小権限アクセスと多要素認証 (MFA) を適用すること。  実行時にプロンプトと応答を検査すること。  過度に自律的な行動を取る AI エージェントや ID を監視すること。	必要のないユーザやシステムに広範な AI アクセス権を付与すること。	悪意あるプロンプトや権限の不正拡大リスクを最小化できる。
運用とレジリエンス (Operations and resilience)	主要な AI 脅威に対するブレイブックやレッド/ブルーチーム演習を実施すること。  すべての AI 入力を検証・サニタイズ(無効化)し、ベクトルデータベースを保護し、異常を記録・確認すること。  モデルおよびデータの来歴を維持すること。	埋め込みの異常を無視し、AI 依存コンポーネントの修正を怠ること。	持続的な攻撃を防ぎ、フォレンジック上の明確性を維持できる。
人とガバナンス (People and governance)	従業員に AI リスク、ディープフェイク、データ取り扱いに関する教育を実施すること。  人間が介在する監督体制 (human-in-the-loop) を義務化すること。  レッドチーム演習に「いたづら (prank) テスト」を含めること。	シャドーAI を許可し、セキュリティ教育を省略すること。結果として、悪用を検証する文化を醸成してしまうこと。	現実世界での AI の誤用に対する耐性を強化できる。

## 人とガバナンス：AI セキュリティの人的側面

テクノロジーだけでは AI を守ることはできません。誤用やバイアス、ソーシャルエンジニアリングを防ぐためには、人による監督、明確な方針、責任の文化が不可欠です。AI リスクと規制の圧力が高まる中、強固なガバナンスと訓練されたチームを持つ組織は、脅威をより早く察知し、信頼を維持することができます。

この戦略を日々の実践へと落とし込むためには、次の点を優先することが重要です。

- **AI セキュリティ研修**：スタッフが AI ツールの使い方だけでなく、それがどのように悪用され得るかを理解できるようにします。
- **利用ポリシー**：社内および顧客向けの AI アプリケーションにおいて、許可される行為と禁止される行為を明文化します。
- **人間が介在する監督 (Human-in-the-loop)**：重要な AI 出力や意思決定には、人間による確認を行い、安全性と倫理基準を満たすようにします。
- **体系的な AI レッドチーム演習**：従来のペネトレーションテストに加え、敵対的プロンプト、プロンプトインジェクションの連鎖、データ汚染の試みを模擬します。
- **いたづら (Prank) テスト**：ドライブスルーAI (ファストフード店などのドライブスルーの注文処理用の AI システム) に対して「タコス を 1000 個注文する」などの不合理またはソーシャルエンジニアリング(心理的な隙)を伴う誤用を再現し、運用上のレジリエンス(耐性)を検証します。



# AI フレームワークとコンプライアンス：先手を取るか、後手に回るか

新たなフレームワーク、標準、法規制が AI の環境を再構築する中で、コンプライアンスを最優先とする姿勢が極めて重要です。初期段階からコンプライアンスを組み込むことで、法的小よび財務的なリスクを軽減し、顧客や規制当局との信頼関係を構築できます。コンプライアンスを重視する組織は、AI を安全に拡張し、後々高くつくことになるビジネスの中断や損害を回避することができます。

組織の法的管轄に影響を与える AI サイバーセキュリティフレームワークや標準、法規制を採用し、遵守することが求められます。

## 主要な AI サイバーセキュリティフレームワークと標準

フレームワーク／標準	説明
<a href="#">OWASP Top 10 for LLM Applications 2025</a>	大規模言語モデルにおける主要なセキュリティリスクの一覧（例：プロンプトインジェクション、機密データ漏えいなど）を示します。
NIST AI Risk Management Framework (AI RMF 1.0)	セキュリティ、バイアス、レジリエンスなどを含む AI リスク管理のための米国フレームワークであり、世界的に広く採用されています。
ISO/IEC 42001:2023	初の AI マネジメントシステム標準であり、AI ライフサイクルにおけるガバナンスとセキュリティ管理を含みます。
Cloud Security Alliance AI Controls Matrix (CSA AICM)	ベンダー中立のフレームワークで、モデルセキュリティや脅威管理など 18 の領域にわたり 243 の AI セキュリティ管理項目を定義しています。
<a href="#">MITRE ATLAS</a>	AI システムに対する敵対的脅威の全体像を示し、AI モデルに対する攻撃手法とその緩和策を体系的にマッピングしています。

## 制定済みの AI 関連法

法律・法案	詳細
<a href="#">EU Artificial Intelligence Act (Europe)</a>	リスクベースの AI 法であり、有害な利用（例：社会的スコアリング）を禁止し、高リスク AI システムに対して厳格な規制を設け、透明性と人による監督を義務付けます。
Transparency in Frontier AI Act (TFAIA) (USA - California)	大規模 AI 開発者に対し、安全計画の公開と深刻なリスクの報告を義務付け、AI による重大な障害を防止します。
Colorado AI Act (CAIA) (USA - Colorado)	「高リスク」AI を利用する企業に対し、バイアス防止の責任と、意思決定に AI を使用する際の通知義務を課します。
Section 103-E Artificial Intelligence (AI) Inventory (USA - New York)	州機関に対し、使用しているすべての AI ツールをリスト化し、AI が関与する業務で労働者の権利を保護することを求めます。
ELVIS Act or Ensuring Likeness Voice and Image Security Act (USA - Tennessee)	AI による音声や画像の無断クローン作成を禁止し、個人の肖像権を保護します。

法律・法案	詳細
Interim Measures for Generative AI Services (China)	生成 AI プラットフォームに対し、コンテンツ管理、セキュリティ審査、アルゴリズム登録を義務付けます。
Act 927 (GenAI ownership) (USA – Arkansas)	AI が生成したコンテンツの所有者を明確化し、AI システムによる著作権侵害を防ぐための規則を定めます。
Right to Compute Act (USA – Montana)	合法的なコンピューティング資源へのアクセスを保護し、重要インフラにおける AI のリスク管理に関する規定を設けます。

注：一部の法域（例：シンガポール）では、法的拘束力のある AI 法令ではなく、影響力の高いフレームワーク採用していません。をそれらは「標準／フレームワーク」の表に記載されています。

## 審議中／施行待ちの法案

法案名 (国／州)	提案内容 (Proposed Coverage)
Federal AI Safety Bill (USA)	高度な AI システムに対し、重大なリスクを防ぐためのテストおよび報告を義務付ける内容です。
Texas Responsible AI Governance Act (USA – Texas)	操作的な AI の利用や社会的スコアリングを禁止し、透明性に関する規定を設けることを目的としています。
UK AI Regulation Framework (United Kingdom)	各分野の規制当局が安全性と公平性の原則を適用するための指針を提供します。
Brazil AI Bill (PL 2338/2023) (Brazil)	倫理的な AI と消費者保護に重点を置き、上院で可決済みで最終承認を待っています。

## 今後の AI 脅威への備え

新たに出現する AI 脅威は、複雑さと影響の両面で加速しており、モデルの整合性への攻撃、生成機能の悪用、機密データを露出させる無許可の AI 導入など、法的・財務的リスクを引き起こす可能性があります。

新たに出現する AI 脅威	説明	リスク	事例
間接的プロンプトインジェクション	攻撃者がメール、文書、ウェブページなどの信頼できる情報源に悪意のある指示を隠し、AI システムを欺いてガードレールを回避させ、ツールの呼び出しやデータの漏えい・移動を誘発します。	企業向け LLM やエージェント型システムでは、信頼されたコンテンツに依存するため深刻であり、ユーザ操作なしで静かに実行される可能性があります。	AIM Security の EchoLeak は、Microsoft 365 Copilot において、単一の細工されたメールによって間接的プロンプトインジェクションが発生し、機密データが流出するゼロクリック脆弱性を明らかにしました。
汚染された学習データ	攻撃者は学習データセットのごく一部を操作することで、隠れた挙動やバックドアを仕込むことができます。	特定の文字列を含むプロンプトが提示された際にモデルが誤作動を起こし、サービス拒否やデータ流出攻撃を可能にします。	カーネギーメロン大学の CyLab 研究者は、学習データセットのわずか 0.1% を変更するだけで AI モデルを侵害できることを実証しました。汚染されたデータは特定の条件下でバックドアを起動させることができ、最小限の変更でも重大なセキュリティリスクを引き起こすことを示しました。
ディープフェイクによる音声・映像詐欺	攻撃者は AI 生成の音声や映像を用いて人物を巧妙に偽装し、通話や会議中にソーシャルエンジニアリングや身元詐称を行います。	内部アクセスの悪用、データ窃取、マルウェア感染、制裁対象者の雇用による法的リスクなどを引き起こします。	2025 年 7 月、米司法省は北朝鮮の組織を摘発しました。この組織は、ディープフェイクを使って偽の IT 労働者を装い、リモート勤務の職を得て得た収益を政権に送金していました。
モデル窃取／抽出攻撃	攻撃者が API 経由でモデルの挙動を模倣したり、学習データを抽出したり、モデルの重みを盗用・複製する攻撃です。	知的財産と競争優位の喪失、学習データからのプライバシー侵害、法的リスクの発生などが含まれます。	2023 年、Meta の LLaMA モデルの重みが BitTorrent 経由でオンラインに流出し、モデル全体のアーティファクトが一般に公開されました。
シャドーAI の導入	従業員やチームが無許可の AI ツールを利用したり、ガバナンス外でチャットボットを導入したりすることで、機密データを流出、セキュリティ対策を回避します。	データ漏えい、コンプライアンス違反、知的財産の喪失、評判や法的リスクの発生につながります。	医療従事者が患者データを AI ツールや個人クラウドにアップロードし、HIPAA コンプライアンスの違反を引き起こしました。

次に示すのは、これらの新たな脅威から組織を守るためのベストプラクティスです。

1. **継続的な脅威インテリジェンスの活用**：MITRE ATLAS、OWASP Top 10 for LLM Applications 2025、業界の AI 脅威グループを定期的に監視し、新たな発見を実行可能なチケットに変換します。
2. **動的ガバナンス**：ISO/IEC 42001:2023 および NIST AI RMF に沿って、AI 資産、リスク、ポリシーを四半期ごとに見直します。責任者、緊急停止機能、承認済みツールリストを明確に設定します。
3. **ガードレールとプロンプトインジェクション対策**：プロンプトの前後でフィルタリングを行い、最小権限アクセスを適用し、ジェイルブレイクや間接的インジェクションへの敵対的テストを実施します。
4. **RAG (Retrieval-Augmented Generation) および外部コンテンツの保護**：許可されたソースのホワイトリストを作成し、HTML やリンクを無効化し、取得時のポリシーを適用してベクトルストアの整合性を監視します。
5. **設計段階でのデータ保護**：(生成 AI による)推論前に情報漏洩対策とプロンプトの編集を行い、推論後の出力を検査するとともに、ポリシー判断が可能なゲートウェイを適用します。
6. **サプライチェーン保証**：モデルやデータセットに署名済みアーティファクトを要求し、導入前にデータ汚染やバックドアの有無を確認します。
7. **モデル窃取の防止**：クエリ数を制限するとともに、学習データにウォーターマーク(デジタル透かし)を付与、重みを分離して、漏洩につながるパターンを監視します。
8. **ディープフェイク対策の強化**：機密に抵触する指示については別ルートで確認をするように求め、ディープフェイク発生時の対応プレイブックを継続的に運用します。
9. **シャドーAIの管理**：安全な承認済みツールのカタログを公開し、無許可のエンドポイントを遮断、従業員に AI セキュリティ教育を実施します。
10. **SOC (セキュリティオペレーションセンター) との統合**：AI を活用した自動インシデント対応プラットフォーム、プレイブックを導入します。さらに、SOC の運用フローに AI による脅威ハンティングを組み込み、ガードレールび削除、異常な RAG 呼び出し、拒否回避といった試みを検知します。
11. **利用状況とコストのテレメトリ**：トークン、エンドポイント、利用コストを追跡し、無許可での利用やデータ流出の試みを検出します。



# 結論：戦略的な拡張力としての AI セキュリティと競争優位性

AI は今後 10 年のビジネスイノベーションを形作る中心的な要素になります。安全性に欠ける場合、信頼を損ない、法令違反を引き起こし、業務の中断を招く可能性があります。成長を促進する AI の取り組みと、監査や規制の中で停滞する取り組みの差は、AI の構築方法、そしてセキュリティを「成長初期からの戦略的推進力」として扱うかどうかにかかっています。

安全性と透明性を備え、ガバナンスが確立された AI を採用する組織は次のような成果を得ることができます。

- コンプライアンス上の停滞を回避し、再作業を減らすことで、より迅速にイノベーションを推進します。
- 対応に高額を要する侵害が発生する前に防止することで、総保有コスト (TCO) を削減します。
- 顧客、パートナー、規制当局から継続的な信頼を獲得します。
- 大規模な再設計を行わずに、新たな脅威にも柔軟に対応します。

AI の初期段階から堅牢な保護策を組み込み、運用パイプラインを強化、サプライチェーンを保護するとともに、モデルを検証し、新たなリスクを継続的に監視することで、組織はブランド、収益、評判を守りながら AI の可能性を最大限に引き出すことができます。

Trend Vision One による支援(一部日本では提供を開始していない製品、サービスもございます)

## ガバナンスとコンプライアンス

- CREM –コンプライアンス管理：組織が NIST RMF や独自のフレームワークおよび標準に基づいて、セキュリティ体制を評価、カスタマイズ、監視、報告できるようにします。
- ZTSA AI アクセス制御：多要素認証 (MFA) およびプロンプト検査を実施します。
- エンドポイントディープフェイク検出：有効化することで、各種コンプライアンス基準に準拠します。

## 運用のレジリエンス

- AI Application Security：AI モデルおよびアプリケーションを脆弱性、悪意あるプロンプト、データ漏えいから保護します。AI Scanner と AI Guard の 2 つの機能により防御を実現します。
- コンテナセキュリティおよびコードセキュリティ：AI ワークロードをスキャン、修正、保護します。
- ファイルセキュリティ：学習または導入前に、データセットやアーティファクトをスキャンして脅威を検出します。
- レッド/パープルチーミングサービス：実際の攻撃を想定したシミュレーションを行い、検知および対応力を強化します。

## 脅威防御

- AI Application Security および AI-DR：LLM、RAG パイプライン、ベクトルデータベースを保護します。
- ディープフェイク検出：AI による音声や映像のなりすまし詐欺を防止します。
- TippingPoint IDS/IPS：AI を悪用したエクスプロイトをネットワークの境界で遮断します。
- AI App Guard：ユーザのワークステーション上で AI アプリケーションおよび関連ファイルを保護するための防御を提供します。
- Trend Cybertron：AI に関連するサイバー攻撃経路を含む脅威を予測し、優先順位を付けて対応します。
- Trend Companion AI (生成 AI)：複雑な脅威インテリジェンスを経営層向けの実行可能なアクションプランに変換します。

## 取り組みを始めるための5つの行動

- √ 使用中のすべてのAIツール（承認済み・未承認を含む）を把握する。
- √ すべてのAIシステムへのアクセスに多要素認証（MFA）を導入する。
- √ 開発チーム向けにAIセキュリティ研修を計画する。
- √ AIモデルのサプライチェーン（SBOM）を確認し、文書化する。
- √ [トレンドマイクロに問い合わせ](#)て、AIリスクアセスメントを開始する。

## 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。



© 2025 Trend Micro Incorporated. All Rights Reserved.

---

<sup>i</sup> Allie Mellen et al. (April 14, 2025). Forrester. "The Top Cybersecurity Threats In 2025." Accessed on Oct. 27, 2025, at <https://www.forrester.com/report/the-topcybersecurity-threats-in-2025/RES182329>.