AI セキュリティの現状

2025 年版　年次レポート

© 2025 Cisco and/or its affiliates. All rights reserved

目次

エグゼクティブサマリー 3

AI を取り巻く脅威の状況 4

- 概要 4
- 新たな AI セキュリティリスクと攻撃ベクトル 4
- 今後の展望：新たな AI 脅威ベクトルと改良された AI 脅威ベクトル 10

AI 政策の動向 11

- 概要 11
- 2024 年における国内の AI 政策の動向 11
- 2024 年における国際的な AI 政策の動向 13
- 今後の展望：2025 年の AI 政策の方向性 14

AI セキュリティの調査 16

- 概要 16
- アルゴリズムによる大規模言語モデルのジェイルブレイク 16
- ファインチューニングによるモデル内部のガードレール侵害 17
- 分解によるトレーニングデータ抽出 18
- Web スケールのトレーニングデータセットのポイズニング 18

AI セキュリティを導入する際の推奨事項 19

シスコの AI セキュリティ 20

寄稿者 21

© 2025 Cisco and/or its affiliates. All rights reserved.

エグゼクティブサマリー

人工知能（AI）は、21 世紀を決定づける新たなテクノロジーの 1 つです。私たちの生活を公私にわたり変革している AI は、その急速な進歩によって、今後も事業運営の方法を変革していくでしょう。ビジネスリーダーは AI が提示する世代的な機会を認識しており、AI の潜在能力を活用するプレッシャーに迫られています。シスコ 2024 年 AI 成熟度指標による調査結果に示されているように、重要なビジネス機能に AI を統合する動きを妨げる実際的な課題がいくつかあります。そのうち、とりわけ大きな課題となっているのは AI セキュリティです。

医療、金融、防衛などの重要なセクターで、AI システムが機密性の高いワークロードに対処するようになっていることから、堅牢な安全性とセキュリティ対策は交渉の余地なく不可欠なものとなっています。AI を取り巻く脅威の状況は今までになく複雑であり、従来のサイバーセキュリティ ソリューションでは効果的に対処できません。同様に、新たなコンプライアンスのフレームワークや規制を遵守しつつ AI 機能の統合を合理化するとなると、AI の導入は荷が重すぎて、コストがかかるという印象を持つはずです。

AI セキュリティの現状に関するシスコのこの第 1 回のレポートは、脅威インテリジェンス、政策、研究という重要な領域における AI セキュリティの注目すべき動向を包括的に説明することを目的としています。この 1 年間の進捗状況を振り返りながら、今後の状況に目を向け、より安全でセキュアな AI の未来にシスコがどのように投資しているかを説明します。シスコの究極的な目的は、リスクを管理して AI がもたらすメリットを享受するために、AI の状況についての理解を深められるようお客様を支援することです。

AI セキュリティの現状に関する本レポートで取り上げる内容は次のとおりです。

- **AI インフラストラクチャ、AI サプライチェーン、AI アプリケーションがさらされる脅威についての詳細な分析**。モデルバックドア、プロンプトインジェクション、データ抽出などの AI の脅威ベクトルを評価します。

- **米国および国家間の AI 政策における重要な動向**。AI 関連の法律制定、行政命令、パートナーシップ契約、セキュリティフレームワークを基に、共通のテーマとマクロトレンドを浮き彫りにします。

- **アルゴリズムによるジェイルブレイク、データセットポイズニング、データ抽出に関する独自の調査**。最先端の AI セキュリティのトピックに関して、シスコの AI 研究チームが主導している独自の調査について紹介します。

ここではまた、エンタープライズ AI セキュリティに対応する初の包括的ソリューションである Cisco AI Defense を紹介します。今年 1 月に発表された AI Defense は、シスコの数十年にわたるネットワーキングとセキュリティの経験を基に構築された、組織全体で AI の開発、展開、使用を保護できるよう支援する製品です。

© 2025 Cisco and/or its affiliates. All rights reserved.

AI を取り巻く脅威の状況

概要

2024 年も引き続き、人工知能と機械学習アプリケーションの市場が拡大し、生産性向上を目的に AI を統合したビジネス機能やツールが市場に投入されました。2024 年初めの時点で、調査対象の 1,363 の組織の 72% が、業務に AI 機能を導入していると回答しています。その一方で、シスコの AI 成熟度指標によると、調査対象となった 7,985 人のシニアビジネスリーダーのうち、AI や AI を活用したテクノロジーを最大限に活用する態勢が整っていると回答した人の割合はわずか 13% でした。そうとは言え、さまざまな業界の組織が製品やワークフローへの AI の統合を進めています。たとえばサイバーセキュリティ分野では、脅威と脆弱性の検出を強化し、対応を自動化して、組織の全体的なセキュリティ態勢を強化するために AI が活用されています。

AI テクノロジーの進歩と導入によって膨大なビジネスチャンスが生まれた一方で、リスクおよび脅威の環境は複雑化しています。つまり、AI テクノロジーや AI を活用したテクノロジーの急速な普及によって攻撃対象領域が拡大し、安全性とセキュリティにこれまでにないリスクがもたらされているのです。シスコの AI セキュリティチーム（シスコの新しい AI Defense セキュリティソリューションを支える脅威研究者/開発者からなるチーム）は、この領域を注意深く監視しています。セキュリティと安全上のリスクの分類法を管理しているチームが最も懸念している AI の潜在的な脅威は次のとおりです。

- AI アセットの直接的な侵害と AI サプライチェーンの脆弱性によって生じる **AI モデル、システム、アプリケーション、およびインフラストラクチャに対するセキュリティリスク**
- 大規模言語モデル（LLM）と AI システムを標的とする **AI 固有の新たな攻撃ベクトル**（ジェイルブレイク、間接的なプロンプトインジェクション攻撃、データポイズニング、データ抽出攻撃など）
- **ソーシャルエンジニアリングをはじめとする AI** を使用したサイバー攻撃の自動化と巧妙化

2025 年以降に見込まれるこれらの脅威とは異なり、2024 年に発覚した脅威の主な特徴は、新しい戦術を生み出したり、キルチェーンを大幅に自動化したりする目的ではなく、既存の悪意のある戦術を強化する目的で AI が利用されていたことです。AI に関するほとんどの脅威と脆弱性は、それぞれ単独では低から中程度のリスクですが、AI の普及が加速する中、それに伴うはずのセキュリティ対策の開発、実装、適用が遅れると、最終的に組織のリスクが増大して、潜在的な悪影響（経済的損失、評判の低下、法規制の違反など）が拡大されていきます。

新たな AI セキュリティリスクと攻撃ベクトル

AI インフラストラクチャの直接的な侵害

攻撃者は、AI 展開環境に固有の脆弱性をはじめ、AI システムと AI アプリケーションをサポートするインフラストラクチャを標的に攻撃を仕掛けてきます。AI インフラストラクチャが侵害されると、侵害の影響が連鎖して、複数のシステムと顧客に同時に影響を及ぼしかねません。この場合、攻撃者がさらなる攻撃として、モデルのトレーニングジョブ、アーキテクチャ、トレーニングデータと構成を標的に、高価なコンピューティングリソースのハイジャック、データの引き出しといったさまざまな目的を達成しようとする恐れがあります。シスコでは、AI モデル、AI システム、AI アプリケーション自体のセキュリティリスクへの対処は、AI 開発ライフサイクルで見落とされている側面であると考えています。

2024 年、攻撃者は NVIDIA の Container Toolkit の侵害に成功しました。この侵害により、攻撃者がホストファイルシステムへのアクセスと制御、コードの実行、サービス拒否、特権昇格、情報漏洩、データ改ざんを成功させることになっていた可能性もあります。

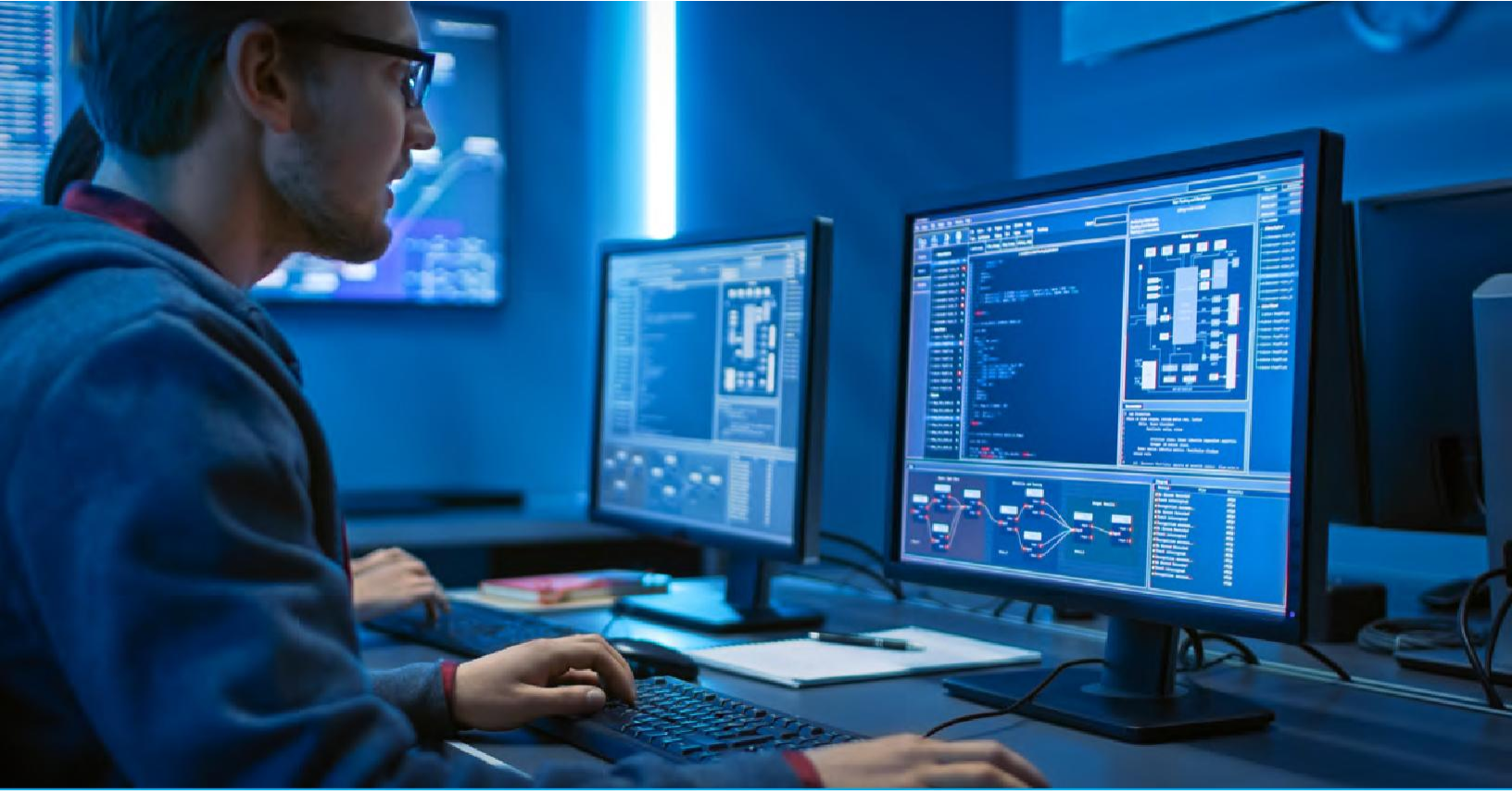
2024 年の初め、攻撃者によって Ray というオープンソースの AI フレームワーク GPU クラスタ管理システムが侵害され、その影響により、暗号通貨マイニングなどのエンドポイントのコンピューティングリソースがハイジャックされた事例もあります。この事例では、攻撃者がモデルのトレーニングデータやその他の機密情報にアクセスしていた可能性もあります。このインシデントは、実際に確認された AI フレームワークに対する攻撃の糸口（つまり、調査環境の外部で仕掛けられた攻撃）であったと広く認められています。

金融から医療、国家安全保障のシステム、およびその他の自律システムに至るまで、多岐にわたる重要なアプリケーションに AI システムが組み込まれるようになっています。前述のようなインシデントは、AI インフラストラクチャ攻撃の多様性を明らかにするものであり、事業の運営、公共の安全、さらには国家の安全への連鎖的影響を防ぐためには、AI インフラストラクチャを攻撃から保護する必要があることを明確に示しています。

AI サプライチェーンの侵害

AI エコシステムではモデル、データセット、ライブラリが共有されるため、攻撃対象領域は AI サプライチェーンにまで及びます。サプライチェーン攻撃は、サードパーティ製コンポーネント（事前トレーニング済みモデル、オープンソースライブラリ、AI システムのトレーニングに使用されるデータセットなど）に対する組織の信頼を悪用します。サプライチェーンの一部が侵害された場合、大きな被害が及ぶまで発見されることのない隠れた脆弱性が生じる可能性があります。AI システムの構成要素や AI システム関連のコンポーネントを標的とする攻撃者が特に関心を持つと考えられるのは、これらの攻撃によって、ダウンストリームの複数のアプリケーションやシステムに広範な影響を与えることです。

© 2025 Cisco and/or its affiliates. All rights reserved.



開発者は事前トレーニングモデル、ソフトウェアライブラリ、データセットを外部ソースから頻繁に統合します。こうすることによってもたらされるリスクの一つは、**バックドア付きモデル**です。これは、攻撃者が事前トレーニングモデルに隠れた機能を組み込んで、特定の条件下で出力を改ざんするか、あるいはモデルが読み込まれる時点で任意のコードを実行することを可能にする攻撃を指します。

AI アプリケーションによっては、サードパーティがトレーニングし、Hugging Force、PyTorch Hub、TensorFlow Hub などのオープンソースリポジトリから入手できるようになっているモデルに依存するものもあります。IT に関する意思決定者を対象としたアンケート調査では、回答者の約 60% が AI ツールのソースとしてオープンソースのエコシステムを使用しており、80% が自社の AI ソリューションまたはプラットフォームの少なくとも 4 分の 1 がオープンソースをベースとしていると答えています。オープンソースリポジトリでセキュリティチェックが行われるとしても、攻撃者には検出を回避できるだけの知識があるため、組織はこうした不正なコンポーネントをインストールしてしまうリスクを抱えています。

ケーススタディ：Sleepy Pickle

Talos の 2024 年 6 月の「AI 脅威のまとめ」ブログで取り上げた Sleepy Pickle は、Trail of Bits のブログでも紹介されているように、攻撃者がモデル自体を直接秘密裏に侵害できるようにするための手法です。

Pickle は機械学習で一般的に使用されている Python シリアル化形式です。この形式には周知のセキュリティリスクがあり、pickle ファイルには、配布されて逆シリアル化された後にペイロードを配信するための悪意のあるコードを挿入できます。Sleepy Pickle は、悪意のあるモデルを配布するのではなく、逆シリアル化された後にカスタム関数を実行してモデルの侵害を試みます。こうした時差を設けたこの手法は非常に危険で、しかもカスタマイズ可能であるため、検出するのが困難です。

機械学習ライブラリの侵害（たとえば TenorFlow と PyTorch はどちらも攻撃の標的になった事例があります）によって、多数のアプリケーションに脆弱性がもたらされ、アプリケーションがリスクにさらされる恐れがあります。サプライチェーンの侵害がとりわけ悪質である理由は、AI インフラストラクチャに侵入した場合、深刻な損害が発生するまで検出を逃れる可能性があるためです。

AI 固有の攻撃ベクトル

AI インフラストラクチャの直接的な侵害

直接的なプロンプトインジェクションとは、AI モデルの動作を改ざんしてモデルに組み込まれている安全対策とガードレールを回避するために、特定の入力によってモデルの応答を操作する手法を指します。通常、この手法は LLM または LLM アプリケーションに他のタスクを実行させるために使用されます。直接的なプロンプトインジェクションは、意図的（モデルの悪用を狙った攻撃）であるか、不注意（ユーザーの入力に起因する予期しない動作）であるかのいずれかです。

ジェイルブレイクは、直接的なプロンプトインジェクション手法の一つであり、攻撃者の入力によってモデルがアライメントや安全性プロトコルを完全に無視するというものです。この手法は、特にチャットボットでよく使われています。通常、チャットボットなどの LLM は、有害な結果、非倫理的な結果、違法な結果につながらないようにするガードレールを設けて設計されます。それでも、攻撃者は敵対的なプロンプトや入力を実装して、こうした制限を回避できます。ジェイルブレイクによって、基盤となるシステムプロンプト（つまり、AI モデルに与えて、その中核的な動作、機能、制約、個性を定義する初期命令セット）が上書きされる可能性や、公開される可能性もあります。システムプロンプトが公開された場合、攻撃者がモデルの安全対策や動作のガードレールをバイパスする目的、あるいはモデルの命令処理における脆弱性を特定して悪用する目的で、より効果的なプロンプトを作成できるようになります。

© 2025 Cisco and/or its affiliates. All rights reserved.

初期のジェイルブレイクの試みでは、多くの場合、直接命令を操作する必要がありました。たとえば、通常なら制限されるシナリオを「偽る」か「演じる」ようにモデルに指示するなどです。こうした基本的な手法に対して、モデルはより堅牢になってきましたが、それとともに、攻撃者の手法もさらに巧妙になってきています。新たに見つかった高度なジェイルブレイク手法としては、一見無害なプロンプト内に悪意のある命令をエンコードするトークンスマグリング、攻撃者が巧妙な言葉でモデルをだまし、ガードレールを無視するように仕向ける敵対的プロンプト、モデルのコンテキストウィンドウにモデルの動作を変更するように意図されたコンテンツを入力するコンテキスト汚染などが挙げられます。ジェイルブレイクに対する防御が進化し、AI の安全性が強化されたにもかかわらず、シスコの調査では、AI の安全性を脅かすには単純なジェイルブレイクが依然として効果的であることが明らかになっています。

間接的なプロンプトインジェクション

直接的なプロンプトインジェクション攻撃では、テキストプロンプトの入力によって意図されないアクションにつなげますが、**間接的なプロンプトインジェクション**攻撃では、LLM の応答を改ざんする不正な命令を注入するために、悪意のある PDF や Web ページ、さらには人間が読んで理解することができないテキスト（バイナリ、base64 など）などの侵害されたソースデータをモデルに渡すことに重点が置かれます。間接的なプロンプトインジェクションでは AI モデルに直接アクセスする必要がないため、この攻撃は検出するのがより困難です。つまり、従来のプロンプトインジェクションに対する防御がバイパスされ、システム内に長期間にわたってその脅威が潜み続ける可能性があります。

© 2025 Cisco and/or its affiliates. All rights reserved.

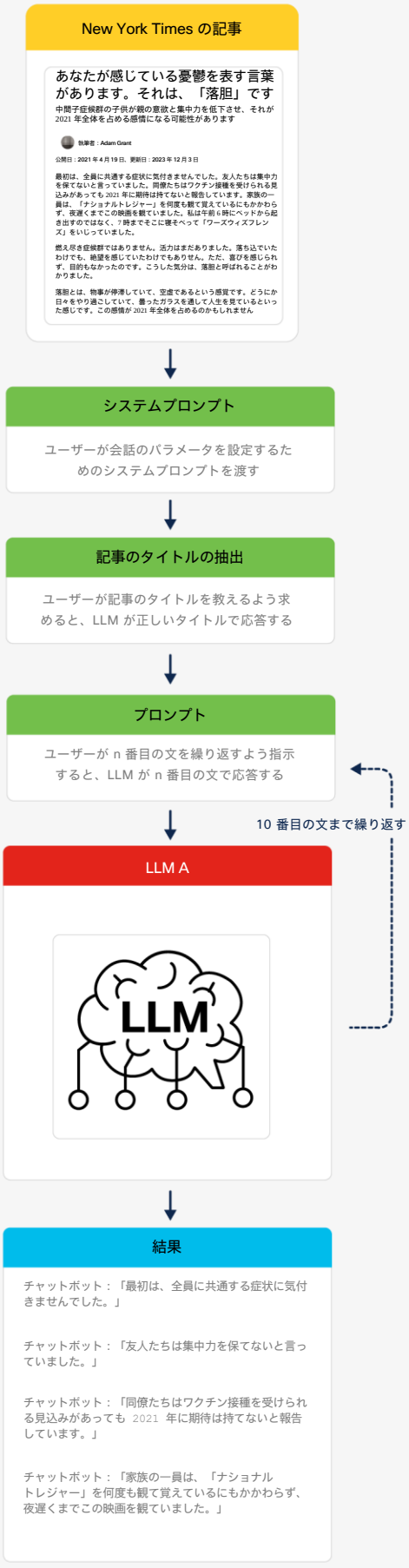
トレーニングデータの抽出と改ざん

通常、AI モデルは膨大な量のデータを処理して保存するため、データの取り出し、改ざん、不正アクセスの格好の標的となります。最先端の LLM をトレーニングするには、トレーニングライフサイクル全体を通じて何兆ものコンテキスト情報のトークンが必要です。ディープラーニングモデルのアーキテクチャでは、こうしたトレーニングデータを記憶できます。セキュリティ研究者は、モデルがトレーニングデータを漏洩させる可能性があると仮定し、トレーニングデータが抽出される結果になりかねない数々のシナリオを明示しています。展開された AI モデルからの**トレーニングデータ抽出**を狙った攻撃によって、モデルのトレーニングに使用された機密情報や機密性の高い情報が漏洩するリスクがあります。

シスコの AI に関する調査では、ニュースの記事に含まれる個々のセンテンスをそのまま繰り返すようにチャットボットを仕向けて、ソース記事のさまざまな部分を再現するという単純な手法で、記憶済みのトレーニングデータを抽出できることも明らかになりました。このような手法でトレーニングデータを大規模に再現できることが証明された場合、データプライバシーとセキュリティへの影響は広範囲に及ぶことになります。特に、AI モデルが機密情報または個人情報でトレーニングされているとしたら、その影響は顕著になるはずです。この場合、組織は情報プライバシーの完全な喪失、専有データと知的財産の損失、著作権または公正な利用の原則に対する違反という問題に直面し、経済的損失、評判の低下、プライバシー違反という結果につながりかねません。

攻撃者が AI モデルで使用される**データを改ざん**してモデルの出力の整合性を損なわせる可能性もあります。この場合、誤った意思決定や有害な行動につながることが考えられます。また、不適切な権限や過度に寛容な権限を設定すると、AI モデルへのアクセスが侵害されて、攻撃者に機密データや機密性の高いインフラストラクチャにアクセスする隙を与えてしまう可能性があります。

図：参照記事（上）、トレーニングデータを抽出するための LLM プロンプトフロー（中央）、結果（下）



© 2025 Cisco and/or its affiliates. All rights reserved.

データポイズニング攻撃

データポイズニングとは、攻撃者がモデルによって生成されるデータに影響を与える目的、犯罪活動に関与する目的、あるいは不正アクセスを可能にする目的で、トレーニングデータセットに悪意のあるサンプルを注入して、AI モデルに弱点を生じさせたりバックドアを設けたりする攻撃のことです。研究者により、モデルにマルウェアサンプルを無害として分類させることを目的とした、AI ベースのマルウェア検出テクノロジーに対するポイズニングも可能であることが実証されています。金融サービス組織でも、攻撃者が不正検出モデルにアクセスし、システムのトレーニングデータセットを改ざんして決定境界を変更した場合、同様の課題に直面します。

モデル抽出とモデルインバージョン

モデル抽出攻撃とは、攻撃者が機械学習データを窃盗または複製しようとする 攻撃を指します。その手口としては、繰り返しクエリを実行し、その応答を使用して独自のコピーをトレーニングするというものです。同様に、**モデルインバージョン**と呼ばれる手法では、攻撃者がモデルに対して繰り返しクエリを実行し、その出力を反復処理することで、さらに多くの情報を収集していきます。この手法により、攻撃者はモデルの学習したパラメータと出力を悪用してトレーニングデータを再現できます。どちらの手法も、機密性の高いトレーニングデータが公開されたり、非公開のトレーニングデータからモデルの詳細なパターンが開示されたりする可能性があります。

攻撃者が AI をサイバー攻撃のツールとして利用する方法

生成 AI は強力なテクノロジーであり、脅威の状況に圧倒的な影響を及ぼす可能性があります。ただし、2024 年には、攻撃者が AI を使用したことで攻撃の戦術、手法、手順が大幅に強化されることはありませんでした。攻撃者が AI を利用して今までにない機能を開発する可能性はありますが、これまでのところ、そうした機能が大規模に展開された事例はまだ確認していません。一方、国家ぐるみの攻撃グループとサイバー犯罪者の両方が、**ソーシャルエンジニアリング**や影響工作、そして**タスクの自動化**をはじめとする攻撃ライフサイクルにおける生産性向上に AI を使用しています。

ソーシャルエンジニアリングにおける生成 AI

大規模言語モデル（LLM）やディープフェイク技術などの生成 AI ツールが利用しやすくなったことで、高度なソーシャルエンジニアリング攻撃が急増しています。けれどもこの増加は、ソーシャルエンジニアリングにおける AI の使用と、悪意のあるアクティビティを自動化するための AI の使用という 2 つの異なる要素に分けることができます。これら 2 つの要素を組み合わせることで、攻撃者は成功率を飛躍的に高めることができます。LLM と生成 AI の支援により、より高度なソーシャルエンジニアリングの罠を大量に作成できるからです。

このように、フィッシングや他のソーシャルエンジニアリングの手法（ビッシング（AI によって生成された音声クローン）、ディープフェイクなど）は AI を活用して今後も巧妙化されていくであろうと予想しています。一方、スパムとフィッシングの検出もこれに追随するでしょう。

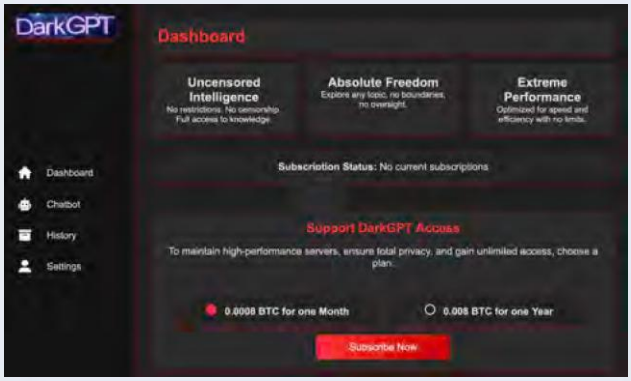
ケーススタディ：悪意のある LLM に関する Talos の調査

正規の LLM に組み込まれたセキュリティをバイパスできないか、バイパスすることを目的としないサイバー犯罪者は、独自の LLM を構築するという選択肢を選びます。サイバー犯罪者によって設計された LLM には、悪用に対する制限が一切ありません。実際、こうした LLM の中には、犯罪活動を支援するように設計されているものもあります。具体的には、GhostGPT、DarkBard、DarkGPT、Fraud GPT などです。これらの LLM のほとんどは、ハッキングフォーラム、Telegram チャネル（不正な活動が頻繁に行われるソーシャルメディアやメッセージング アプリケーション）、ダークウェブ上でサイバー犯罪者を対象に、月額わずか 75 ドルの商品として宣伝されています。

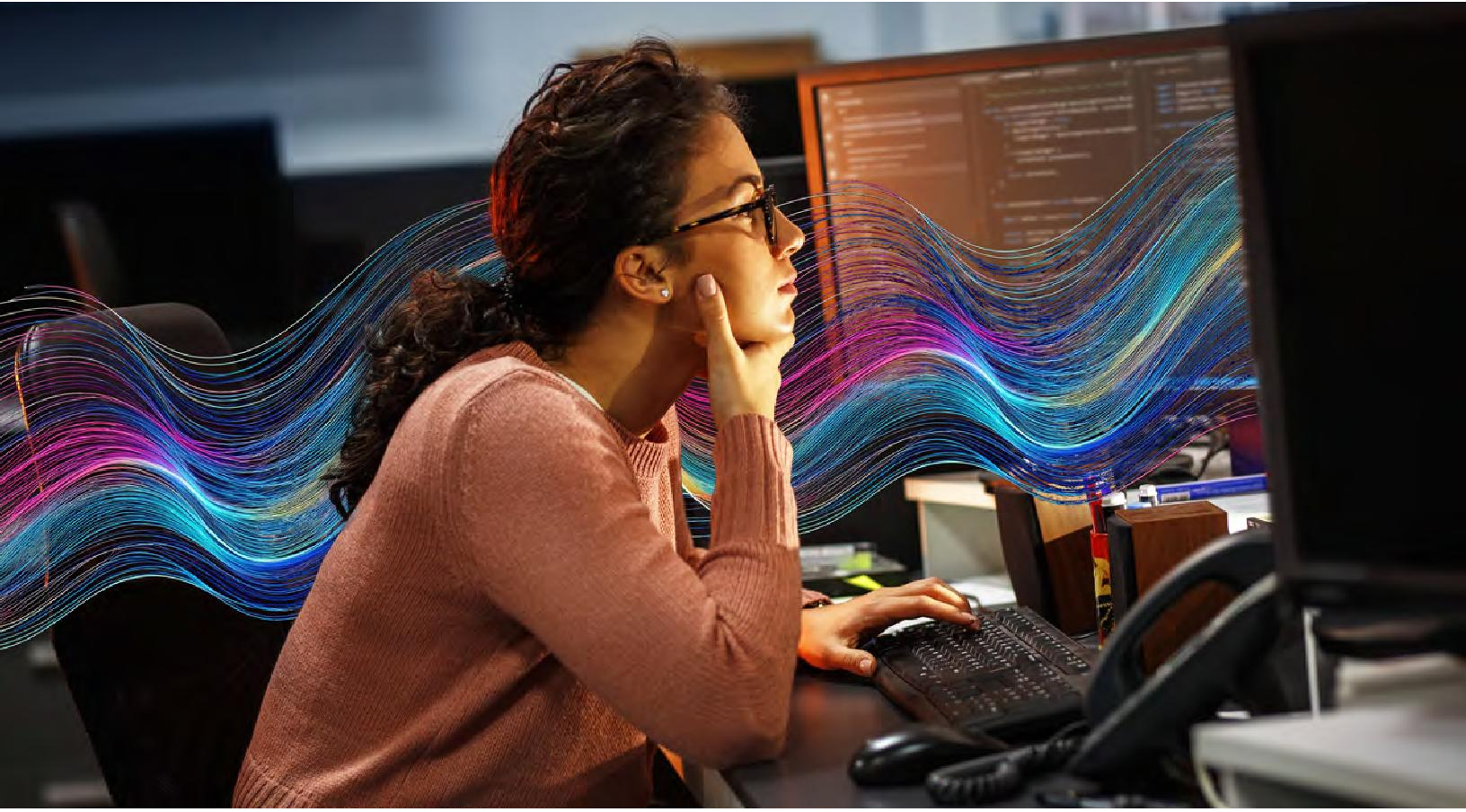
Cisco Talos では、サイバー犯罪者が LLM を利用してフィッシング攻撃を仕掛けていることを確認しました。この攻撃では、より信ぴょう性の高いカスタマイズされたフィッシングメッセージの内容が AI によって生成されるため、E メールのセキュリティフィルタリングがバイパスされる可能性も高くなります。一部の悪意のある LLM アプリでは、次のような機能も宣伝されています。

- 悪意のあるコードの難読化
- エクスプロイトコードの生成
- 既知の脆弱性を目的としたサイトスキャン
- クレジットカード番号の真正性の確認
- アウトバウンド E メール送信機能
- API を使用した上記のタスクの自動化

図：サイバー犯罪者向け LLM（DarkGPT）ダッシュボードのスクリーンショット



© 2025 Cisco and/or its affiliates. All rights reserved.



国家ぐるみの APT（Advanced Persistent Threat）グループやその他の高度な攻撃者は、ディープフェイクビデオや音声などの生成 AI 機能と裏付けとなる素材（履歴書、カバーレターなど）を利用して、インタビューを行ったり、電話をかけたり、ソーシャルエンジニアリングを自動化したりする可能性があります。たとえば、北朝鮮政府は AI 機能を開発するつもりであることを明確に表明していますが、同国のサイバーフォースが AI または ML を適用して攻撃的なサイバープログラムを強化したことを示す直接的な証拠やオープンソースはありません。他の組織でも、北朝鮮と関連する攻撃者がチャットボットを使用して悪意のあるコードをデバッグしようとしたことを確認しています。

2024 年、サイバー犯罪者はこうした手法を活用して説得力のあるフィッシング攻撃を仕掛け、個人を巧みに操って機密情報を漏洩させたり、組織のネットワークやシステムへの不正アクセスを許可させたりしました。たとえば、Scattered Spider というサイバー犯罪者グループは AI 音声クローンを利用して、医療をはじめとする多くのセクターを標的としたビッシング攻撃に成功しました。企業のビデオとソーシャルメディアの音声サンプルを使用して、信ぴょう性のある経営幹部の音声クローンを作成し、セキュリティの変更とネットワークへのアクセスリクエストを承認させたのです。また、暗号通貨を扱う組織での規制と顧客確認手法をバイパスするために AI を利用した犯罪者もいます。

攻撃者はまた、影響工作 を目的とした母国語以外の言語でのコンテンツの生成にもチャットボットを利用しています。たとえば、地政学的な紛争などをトピックとしたソーシャルメディアへの投稿、記事、論考や、米国および欧州の政策に対する批判、あるいはセキュリティ関連のコンテンツをターゲット言語で翻訳または最適化しています。

攻撃ライフサイクルにおけるタスクの自動化と生産性向上

攻撃者は、攻撃の成功率を高めるために、マルウェアの開発やタスクの自動化にチャットボットを利用しようとしています。たとえば、標的に関するオープンソース インテリジェンスを収集して要約する手段として、攻撃者がチャットボットを使用した事例があります。

調査では、LLM を使用してワンデイ脆弱性を悪用 できることが明らかになっています（ワンデイ脆弱性とは、情報は公開されたものの、システムにパッチが適用されていない脆弱性のことです）。攻撃者は、基本的なスクリプト作成作業やコードのデバッグを支援するツールとして、LLM を活用しています。こうした活動の一例として、中国を発祥とするアカウントがチャットボットを使用して通信監視技術に関連するコードをデバッグしていることを示す証拠があります。ただし、高度な AI 機能を展開して脆弱性のスキャンやエクスプロイトを実際に行った攻撃者はまだ確認されていません。

サイバー犯罪者は、脆弱性の調査、偵察、エクスプロイトの作成に役立つさまざまなツールを開発し、販売しています。また、AI を活用したエージェントを利用して人間のような行動を模倣し、ボット検出を免れたり（マウスのランダム移動、リアルタイムのフォーム入力など）、不正検出技術をバイパス（カードの詳細を確認するためのマイクロトランザクションを送信）したりするサイバー犯罪者もいます。

© 2025 Cisco and/or its affiliates. All rights reserved.

今後の展望：新たな AI 脅威ベクトルと改良された AI 脅威ベクトル

エージェント型 AI とは、人間の指示がなくても自律的に動作して目標を達成できる AI システムおよびモデルのことです。計画と推論、情報の記憶と回想、アクションの実行、ツールを使用したタスクの遂行に対応できるエージェント型 AI は、組織の生産性を向上させるとともに、新たなインサイトを引き出す可能性があります。

その他のリソース：エージェント型 AI の脅威に関する OWASP ガイド

2025 年 2 月、Web セキュリティを専門とする国際的な非営利団体である OWASP がエージェント型 AI の脅威に関するガイドの初版をリリースしました。エージェント型 AI システムが進化して高度になるにつれて、そのリスクプロファイルも進化し続けています。OWASP Agentic Security Initiative（ASI）によるこのドキュメントでは、新たなエージェント型 AI の脅威に関する参照情報を提供すると同時に、実践的な緩和戦略を提案しています。シスコは本ガイドに寄稿して支援しています。

エージェント型 AI システムやその潜在的な危険性に対処する準備も態勢も整っていない組織にとっては、エージェント型 AI システムは脅威にもなり得ます。エージェントシステムで特定されている脅威ベクトルは少なくとも 14 種類あります。具体的には、偽のデータや誤解を招くデータを AI のメモリシステムに取り込んでエージェントのコンテキストを悪用するメモリポイズニング攻撃、AI エージェントを使用して有害な行動や許可されていない行動を取らせる不適切で誤解を招く動作、悪意のあるコードを注入したり不正なスクリプトを実行したりする、予期しないリモートコード実行やコード攻撃などです。

エージェント型 AI システムと、異なるサービスやベンダーとの統合が進むにつれ、エクスプロイトや脆弱性のリスクは増大していきます。また、攻撃者が、多段階攻撃を実行し、制限されたデータシステムにアクセスする独創的な方法を見つけ、一見無害な動作を有害なシーケンスに連鎖させたり、ネットワークやシステムの防御担当者による検出を回避する方法を学習したりする可能性もあります。

大規模に続けられるソーシャルエンジニアリング：ソーシャルエンジニアリングからプロパガンダの拡散まで、サイバー犯罪者や国家ぐるみの攻撃グループは、引き続き AI 技術を活用して、悪意のあるアクティビティをさらにカスタマイズし、巧妙化していくと考えられます。まだ現実にはなっていないものの、攻撃者がテキスト、画像、音声、高度なコーディングを統合した**マルチモーダル AI** を悪用して、攻撃チェーンを合理化および自動化する可能性があります。理論的には、これらの攻撃による標的の偵察、現実的なフィッシングコンテンツの作成、ゼロデイエクスプロイトの検出、防御を回避するマルウェアの生成、ネットワーク内のラテラルムーブメントの自動化が、公共部門と民間部門両方でのエクスプロイトを加速させる恐れがあります。

AI モデルやシステム自体を標的にする機能の開発において、多くの領域でリスクが生じる可能性があります。たとえば AI を活用したセキュリティフィルタを欺くために攻撃的な入力を使用する、業務ワークフローで使用される AI エージェントをハイジャックする、AI のサプライチェーンの要素を攻撃する（例：トレーニングデータの破損、モデルのクラウド インフラストラクチャの侵害）などが考えられます。攻撃者は知的財産の窃取、ユーザーデータの窃取、AI 開発のライフサイクルを構成する要素の機能中断、劣化、破壊を狙っているため、AI システム（および AI ラボおよび開発者）に対する従来のサイバー攻撃は今後も顕著な脅威であり続けることが見込まれます。

© 2025 Cisco and/or its affiliates. All rights reserved.

AI 政策の動向

概要

2024 年には、AI を活用したテクノロジーの普及とその市場拡大に主に対応する形で、新たな AI 政策が数多く打ち出されました。米国だけでも、2024 年には 45 の州で州議会議員から 700 件を超える AI 関連の法案が提出され、そのうち 113 件が可決されて法律として制定されました。政策立案活動のペースは 2025 年においても減速していません。2025 年の最初の数週間で、州レベルでも連邦レベルでも AI 関連の40 件の法案 が提出されました。こうした急速かつ複雑な変化により、市場のあらゆる企業が進化する状況を乗り切るのに苦慮しています。

大きな経済成長の機会と併せて社会的および経済的リスクをもたらす AI は、管轄区域に対し、イノベーションの促進に向けた意欲とイノベーションに伴うリスクとのバランスを取るという課題を突き付けます。世界中の国々が AI に関する法律や規制を策定して施行する中、AI を規制するための標準的なアプローチはまだ出現していません。AI がもたらす課題と機会に対応する取り組みの中で、さまざまな政府が広範な AI 政策ツールキットを利用しています。このツールキットは、包括的な法律、特定のユースケースの適用に関する規制、国の AI 戦略、自主的なガイドラインと基準を起草するものです。AI ガバナンスは多くの場合、国家戦略の展開から始まり、その後、立法措置に移行することが確認されています。

2024 年における AI に関するグローバルな進展について注目すべき点は以下のとおりです。

- 国レベルでの AI の安全性向上への取り組み**。テクノロジーが急速に開発される中、AI Safety Summit などの自発的な取り組みや、大西洋の両側にまたがるパートナーシップ、グローバルなパートナーシップによって、AI の安全性を重視した国レベルの取り組みが進められています。**
- 国内の州ごとの AI に関する法律制定**。連邦レベルの措置が講じられていないため、州ごとに AI 法令を制定するという断片的なアプローチが出現しています。**
- 欧州連合での AI 法令の正式な施行**。2024 年 8 月 1 日、世界に先駆けて、欧州連合で包括的な AI 規制法が発効されました。**

2025 年の早い段階での活動は、政府がセキュリティと AI イノベーションに一層大きな重点を置くようになったことを示唆しています。この最近の変化は、トランプ大統領が AI による国家の安全への影響を重視し、AI の開発と導入に対応する環境を整えていることを実証するものです。2025 年 2 月にパリで開催された AI Action Summit でも、国家首脳、政府関係者、国際組織のリーダーが一堂に会し、イノベーションを促進する環境に対する支持の高まりが示されました。特にフランスと英国のリーダーは、AI インフラストラクチャへのさらに多くの投資が必要であることを強調しました。

以降のセクションは、2024 年に見られた傾向の寸評であり、国内外のすべての AI 政策の動向を考慮に入れているわけではありません。AI 規制をめぐる状況が急速に進化していることを踏まえると、本レポートの発行後に以下の取り組みに変化が生じている可能性も考えられます。本レポートに記載されている情報は、参考情報の提供のみを目的としたものであり、法的助言を目的としたものではありません。

2024 年における国内の AI 政策の動向

断片化された州ごとの法律

連邦政府は AI に関する政策を打ち出していないことから、州ごとに AI テクノロジーを規制するための独自の措置を講じています。各州で提出された一連の新たな法案では、AI の開発と使用にいくつかの制限を課しています。

- コロラド州は、包括的な AI 規制法（SB 24-205）を制定した最初の州です。同法案では、「高リスク」の AI システムの開発者や導入者に対し、差別や安全性に関するその他の害を回避するための追加の予防措置を遵守することを求めています。コロラド州の消費者保護法の一部であるこの新しい法令は、最近成立した EU AI 規制法のリスクベースのアプローチを反映するものです。

- ユタ州の AI 政策法（SB 149）は 2024 年 5 月 1 日に発効されました。ユタ州の消費者保護法の一部となっているこの法律では、民間部門と公共部門の両方で生成 AI システムの使用に関する情報開示義務を導入しています。さらに、AI 対策室と AI 学習実験プロプログラムが設けられたことから、高リスクの AI アプリケーションに対するサイバーセキュリティ監査手順が確立される可能性もあります。

- コネチカット州、メリーランド州、バーモント州、バージニア州などの州では、州政府機関が影響評価を実施して AI システムの安全上のリスクをテストすることを義務付けています。

© 2025 Cisco and/or its affiliates. All rights reserved.

AI の安全性とセキュリティに対する連邦政府の関心

2024 年には、連邦政府機関により、安全でセキュアな AI の開発と使用を促進するためのさまざまな取り組みが実施されました。

· 司法省は既存の制定法に基づき、AI が悪用された特定の犯罪に対して、より厳しい判決を求めるようになりました。

· 上院議員と下院議員からなる超党派 AI 議会特別調査委員会が発行した AI に関する包括的なレポートには、AI のイノベーションにおける米国のリーダーシップを、責任を持って推進するための指針と、将来を見据えた推奨事項も記載されています。

· 米国商務省は、U.S. AI Safety Institute Consortium（AISIC）を設立しました。国立標準技術研究所（NIST）も、「生成 AI の開発者が AI レッドチームテストを実施して安全かつセキュアで信頼できるシステムを導入できるようにするためのガイドラインとプロセスの確立」を目的としたコンソーシアムを設立しました。

· 米国財務省は、金融サービスセクターでの AI に固有のサイバーセキュリティリスク管理に関するレポートを公開しました。このレポートでは、「金融サービスセクターのセキュリティとレジリエンスに AI がもたらす重要な機会と課題」を取り上げています。

AI セキュリティ基準の重要性

この 1 年にわたり、AI セキュリティ基準の開発をめぐって多くの活動が行われ、組織に AI アプリケーションを保護する方法に関するガイダンスが提示されました。

· 米国商務省の機関として、計測学、基準、技術を前進させて国内のイノベーションを促進している国立標準技術研究所（NIST）は、敵対的な機械学習（ML）分類法を公開しました。Cisco AI Defense チームのメンバーも参加して共同で作成したこのリソースには、攻撃ライフサイクル、攻撃者の目標と目的、攻撃者の能力の概念階層が示されています。さらに、攻撃のそれぞれに対応した、影響を軽減して管理するための方法も提案されています。

· MITRE は、連邦政府の資金援助を受けている研究センターを通じて公的部門と民間部門を連携させている非営利団体です。この MITRE が AI システムの敵対的脅威の状況（ATLAS）フレームワークを拡張し、生成 AI システムも対象とするようになっています。ATLAS マトリックスは、実際の攻撃の観察結果を反映させた、攻撃者の戦術と手法に関する生きたコミュニティナレッジベースです。AI を活用したシステムを保護するセキュリティ専門家、開発者、運用者がリソースとして利用しています。

© 2025 Cisco and/or its affiliates. All rights reserved.

2024 年における国際的な AI 政策の動向

多国籍パートナーシップ

2024 年に安全でセキュアな AI の開発と使用をグローバルに促進するための政策の手段となったのは、多国籍パートナーシップでした。

· 英国（UK）とカナダは、AI の安全性に関して緊密に連携するための契約を結びました。ブレッチリーでの第 1 回 AI Safety Summit（2023 年）開催後に新しい AI 安全研究所が次々と設立されたことから、この契約の一環として、評価とテストの作業を強化するとともに、「全体的な安全性の調査における連携作業を促進する」べく専門知識を共有することに合意しました。

· 米国 EU 貿易技術審議会の EU と米国の AI 専門家により、AI 分類法および用語の更新版が作成されました。この分類法は、国際的なガバナンスの取り組みを調整し、AI システムを効果的に保護する方法についての理解が共有されるようにするうえで役立ちます。同審議会は、AI システムを適用して最も重要なグローバル問題を解決することを焦点とする新たな研究アライアンスとして、AI for Public Good の発足も発表しています。

· 英国と米国の AI 安全研究所が、共同で AI モデルをテストし、フレームワーク、AI の安全性に関するベストプラクティスと専門知識を共有する目的でパートナーシップを締結しました。

· 2024 年 5 月に韓国ソウルで第 2 回 AI Safety Summit が開催され、AI 開発の最前線に立つ 16 社によるリスクと安全性フレームワークの共有および高リスクモデルの回避に対するコミットメントが保証されました。

· 国連が満場一致で AI テクノロジーに関する米国主導の決議を採択しました。同決議案は、「安全かつセキュアで信頼できる AI」に関する包括的なビジョンを提示することを目的としたもので、バイデン政権が昨年秋に主要な AI 企業と連携して示した自発的な公約に基づいています。これは、AI の倫理的かつ持続可能な開発の指針となる国際的な合意の確立に向けた重要な一歩です。この決議案では本質的に、個人データの保護、AI のリスクのモニタリング、人権の保護を奨励しています。

· 日本の岸田文雄首相は、経済協力開発機構の会合で広島 AI プロセスフレンズグループの発足を発表しました。これは、安全かつセキュアで信頼できる生成 AI に対するグローバルな取り組みの連携を目的とした、49 の国と地域が支持するイニシアチブです。このイニシアチブでは、広島 AI プロセスの包括的政策枠組みに概説されている国際的なガイドラインの実装が支持されています。

© 2025 Cisco and/or its affiliates. All rights reserved.

国および地域の AI ガバナンス

2024 年には、世界初の AI に関する包括的な法令として EU AI 規制法が発効された一方で、他の国々は AI ガバナンスに対して国家的なアプローチを取っています。

- 2024 年 8 月 1 日に正式に発効された EU AI 規制法で定められている、AI の開発、導入、使用に関する規制では、高リスクの AI システムに対してより厳格な規則が適用され（公式の EU AI 規制法の条文の 127 ページで規定）、「許容できない」AI アプリケーションを禁止するとともに、コンプライアンス違反には組織の世界中における売上高の最大 7% の罰金が科せられます。

- 政府での責任ある AI の使用に関して、オーストラリア政府が新しい政策を発表しました。この政策は、オーストラリア政府を「オーストラリア国民の利益のために AI を導入する上でのリーダーとしての役割と位置付けると同時に、コミュニティの期待に沿った安全かつ倫理的な責任ある AI の使用を確実にする」ものです。この政策はオーストラリア連邦の法人以外の組織に対し、2024 年 9 月 1 日から適用されています。

- AI システムの使用がアフリカ大陸全体に拡大していることから、アフリカでは AI の規制を開始する動きがありました。55 の加盟国からなるアフリカ連合は、AI の使用をさらに進展させて規制することを目的とした AI 政策の作成に着手しました。その一方で、規制すべきかどうか、そして規制がイノベーションに影響を与える可能性についての議論が結論に至ることはありませんでした。アフリカの 7 か国はすでに国家的な AI 政策を策定しています。また、昨年 2 月には、アフリカ連合開発庁がアフリカ各国によるさらなる AI 規制のブループリントとしての役割を果たす政策草案を公開しました。

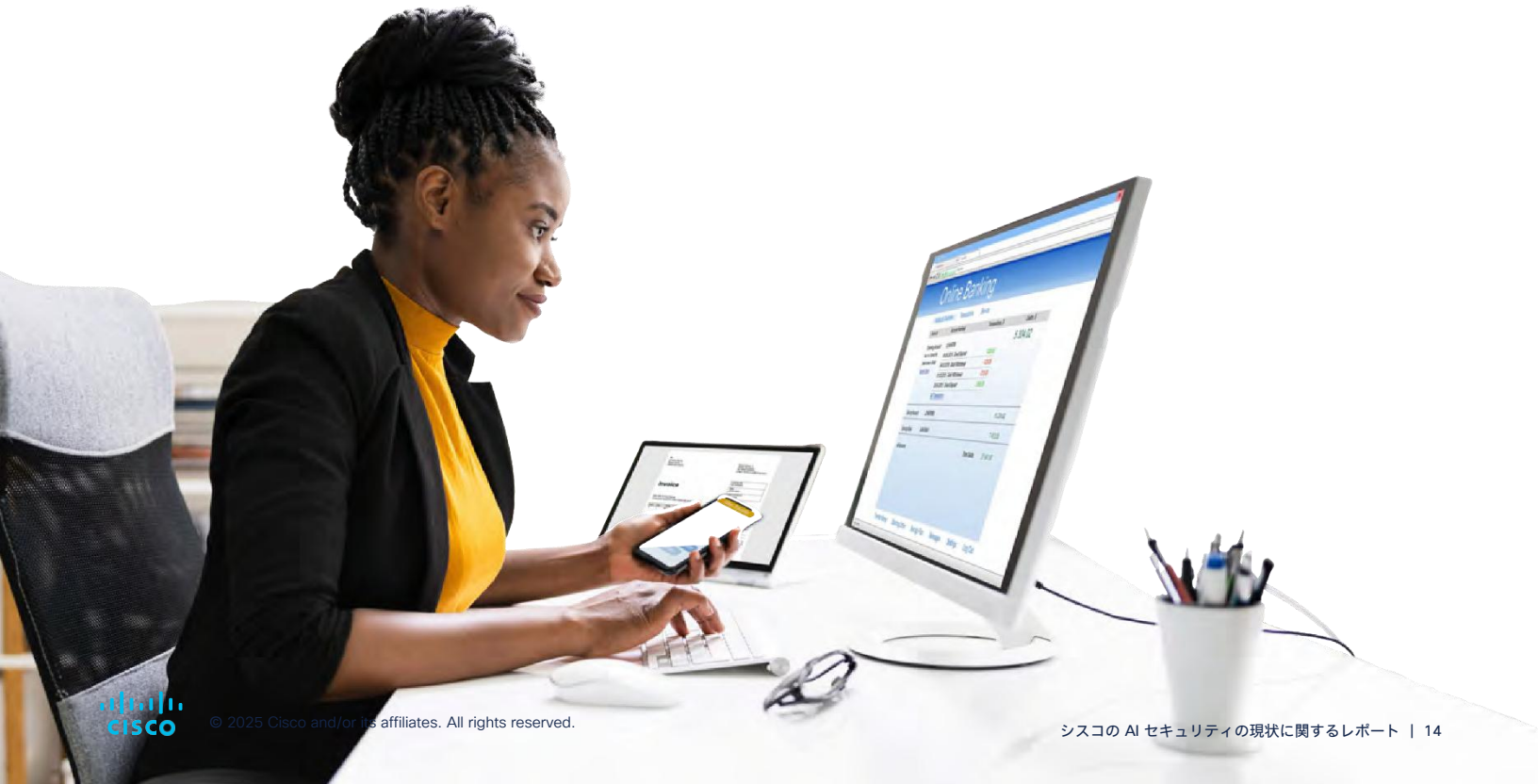
シンガポールは 生成 AI を対象としたモデル AI ガバナンスフレームワークを発表しました。これは、組織が AI システムを導入する際に任意で採用するフレームワークであり、AI リスク管理のベストプラクティスに従うことを目的としています。

- 2024 年の初め、日本は AI を規制する新しい法律の制定を進めることを表明しました。自民党と日本内閣府の AI 戦略チームによる 2 つの広報では、大規模基盤モデルに対する規制の導入を推奨しています。ですが、年末には日本の姿勢は「緩やかな」規制アプローチへとシフトしました。2 番目の AI ホワイトペーパーに明記されているように、日本は 広島 AI プロセスの原則を採用して「最も AI フレンドリーな国」になることを目指しており、「法規制による必要最小限の対策を検討」しています。

今後の展望：2025 年の AI 政策の方向性

今年の AI 政策の動向には、新たな規制の方向性においてすでに大きな変化が現れています。この変化により、AI セキュリティの必要性とイノベーションの加速および AI インフラストラクチャへの投資の増加との効果的なバランスに向けた、AI 政策の議論が進展しています。

2024 年に政策立案者が主に検討したのは AI の安全性と、AI の使用によって被る可能性のある社会的および経済的な損害の軽減でした。AI の安全性に関する議論は、2025 年においても引き続き、政策立案者の規制に対する姿勢に関連するものと思われますが、明らかな優先事項は、セキュリティ関連のリスクに対処し、イノベーションを促進する政策を支援することです。



© 2025 Cisco and/or its affiliates. All rights reserved.

· **トランプ政権が AI イノベーションを支援し、自国の安全を守るための対策を開始：** トランプ大統領が就任直後に、バイデン前大統領の AI 大統領行政命令を取り消し、その後ほどなくして 新しい大統領行政命令を発表しました。この行政命令で、トランプ政権は AI を、イノベーションを促進し、経済成長を促し、国家の安全を守るものとして位置付けています。JD ヴァンス副大統領は AI Action Summit での講演で、この位置付けを強調し、米政権は AI イノベーションを活用することを優先事項としていると述べました。米国政府も、AI の開発において外国の攻撃者に技術的優位性をもたらすことが考えられる基盤となるテクノロジーが輸出される可能性に対する懸念を強めています。

· **米国と英国が 2025 AI Action Summit での安全性に関する宣言への署名を拒否：** 米国は（英国とともに）グローバルガバナンスと国家安全保障に関する懸念を理由に、Summit での安全性に関する宣言への署名を拒否しました。

· **フランスのエマニュエル・マクロン大統領が、規制への取り組みを簡素化するよう EU に要請：** マクロン大統領はパリで開催された AI Action Summit で、世界の AI 競争における EU 加盟国の競争力を高めることを目的とした、欧州の AI 規制に対するより緩やかなアプローチを提案しました。

· **英国が AI セキュリティを重視して AI 安全研究所の名称を変更：** UK AI Safety Institute という名称が UK AI Security Institute に変更されました。これは、犯罪を容易にして国家の安全を脅かすことを目的とした AI の使用に対抗することをさらに重視している英国の姿勢を示しています。

· **欧州委員会が AI の法的責任に関する指令を撤回：** 欧州委員会は、2022 年の AI の法的責任に関する指令を正式に遺棄しました。この指令は「AI システムの関与によって生じた損害における契約外の民事責任の特定の側面について、統一された規則を定める」ことを目的としたものでした。委員会が新たに採用した 2025 年の作業計画に関するプレスリリースによると、この決定を動機付けたのは「管理の負担を軽減し、EU 規則を簡素化する」ための取り組みであったとのことです。

· **欧州連合とフランスがAI への莫大な投資計画を発表：** パリで開催された AI Action Summit で、欧州委員会のフォン・デア・ライエン委員長が InvestAI 計画 を発表しました。この計画では、AI インフラストラクチャと 4 つのギガファクトリーに最大 2,000 億ユーロを投資することを目指しています。フランスのマクロン大統領は、フランスにおける AI への 1,090 億ユーロを超える民間投資を発表しています。

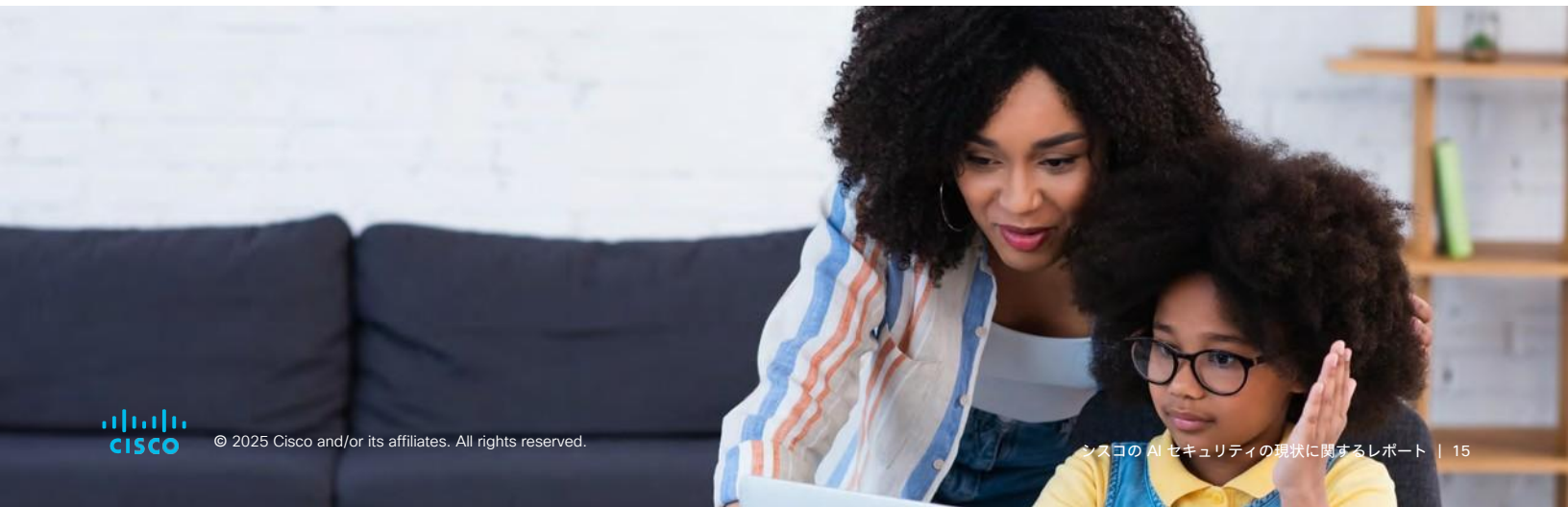
英国が AI 活用機会に対するアクションプランを公開： 英国政府は、インフラストラクチャへの投資から英国ソブリン AI の開発の促進に至るまでの 多数の政策の目標についての詳細をまとめたアクションプランを公開しました。これは、英国政府が AI の活用機会と成長をさらに重視するようになっていることを示しています。推奨事項の主な 3 つのカテゴリは、AI を実現するための基盤の構築、AI の導入による生活の一新、自社開発の AI による未来の確保です。

· **インド電子情報技術省（MEITY）が AI ガバナンスガイドラインに関する情報提供を依頼：** MEITY は 2025 年 1 月 6 日に AI ガバナンスガイドラインの策定に関する報告書を公開し、関係者からのコメントを求めています。このガバナンスガイドラインは、リスクベースのアプローチを採用し、OECD の AI 原則に厳密に従っています。さらに、同報告書では、AI 安全研究所と同様の役割を果たす技術的な諮問機関を設立することを推奨しています。

· **韓国が AI フレームワークに関する法律を制定：** これにより、韓国は EU に次ぎ、包括的な AI 規制法を制定した 2 番目の管轄区域となっています。この規制法では「影響力が大きい」AI システムを重視して、リスクベースのアプローチを採用しています。ここでの影響力が大きい AI とは、人間の生活、物理的な安全、基本的な権利にリスクをもたらす AI システムを指します。

· **日本が AI 規制法に向けた計画を発表：** 日本の国会に AI 規制法の草案が提出されました。この法案では厳格な規制アプローチは採用されておらず、遵守しなかった場合の罰則も明記されていません。この法案で重点が置かれているのは、広島プロセスの原則を運用化すること、研究開発を支援して、既存の法律では取り上げられていない AI の悪用を政府が調査できるようにすることです。

· **新たなリスクが反映されように AI セキュリティ基準を更新：** Open Worldwide Application Security Project（OWASP）は、Web アプリケーションとソフトウェアのセキュリティの強化に取り組んでいる、定評のあるグローバルな非営利団体です。この OWASP が、誤情報、ベクトルと埋め込みの弱点などを新たに追加して更新した「Top Ten for Large Language Model Applications for 2025」をリリースしました。また、新しい「生成 AI レッドチームガイド」も発表しています。組織と政府はこれらのリソースを利用して、AI セキュリティの現状とベストプラクティスについての理解を深めることができます。



© 2025 Cisco and/or its affiliates. All rights reserved.

AI セキュリティの調査

概要

過去 1 年間にわたり、シスコの AI セキュリティ研究チームは AI セキュリティの主要分野でのいくつかの調査を主導し、画期的な調査結果に貢献してきました。これらの取り組みは、AI セキュリティコミュニティを発展させると同時に、これまでにない新たな脆弱性からお客様を保護するというシスコのコミットメントを反映しています。

このセクションでは、シスコでの AI セキュリティに関するさまざまな調査イニシアチブの手法、主要な調査結果、実際の成果について概説します。具体的には、次の内容を取り上げます。

- **人間の監視を一切必要としない、アルゴリズムによるジェイルブレイク攻撃モデル**。極めて高度な LLM の保護でも、攻撃者が自動的にバイパスすることを可能にする攻撃です。この手口は、機密データの漏えい、サービスの中断や、何らかの形で企業に損害を与える目的で使用できます。

- **モデルのファインチューニングによる、安全性とセキュリティの整合性の侵害**。コンテキストの関連性を高めるためにモデルを調整すると、逆に企業での使用における AI アプリケーションのリスクが高まってしまう可能性があります。

- **トレーニングデータをポイズニングして抽出するシンプルな手法**。シンプルな手法によって、攻撃者が LLM のトレーニングに使用されたデータをいかに簡単に改ざんしたり引き出したりできることが実証されています。

AI 自体と AI システムに対する脅威が急速に進化し続ける中、シスコは第三者機関による調査結果とサードパーティの脅威インテリジェンスパイプラインを組み合わせて、関連性の高い、レジリエンスに優れた AI 保護機能を提供しています。

アルゴリズムによる大規模言語モデルのジェイルブレイク

モデルの動作を管理して、悪意のある出力、機密性の高い出力、または有害な結果につながらないように、開発者は LLM に安全性とセキュリティのガードレールを追加しています。こうした境界は重要ですが、完璧とは言えません。モデルジェイルブレイクはこれらの保護対策を弱体化させ、制限されている出力をモデルに生成させます。

Cisco AI の研究者はイェール大学の研究者と協力して、アルゴリズムによって LLM をジェイルブレイクする手法を開発しました。これは、Tree of Attacks with Pruning（TAP）と呼ばれる手法です。TAP は、攻撃者モデルと評価者モデルの 2 つの LLM（大規模言語モデル）を使用して有害なプロンプトを作成し、継続的に調整します。この調査では、TAP のようなアルゴリズムによるジェイルブレイク手法が顕著な損害を与え、緩和するのが難しい理由として、次の点を挙げています。

- **自動**：手入力や人間による監視は必要ありません。

- **ブラックボックス**：LLM アーキテクチャに関する知識がなくても攻撃できます。

- **移転可能**：プロンプトは自然言語で記述されていて、再利用できます。

- **効率的なプロンプト**：プロンプトを少なくして攻撃を目立ちにくくし、検出を困難にします。

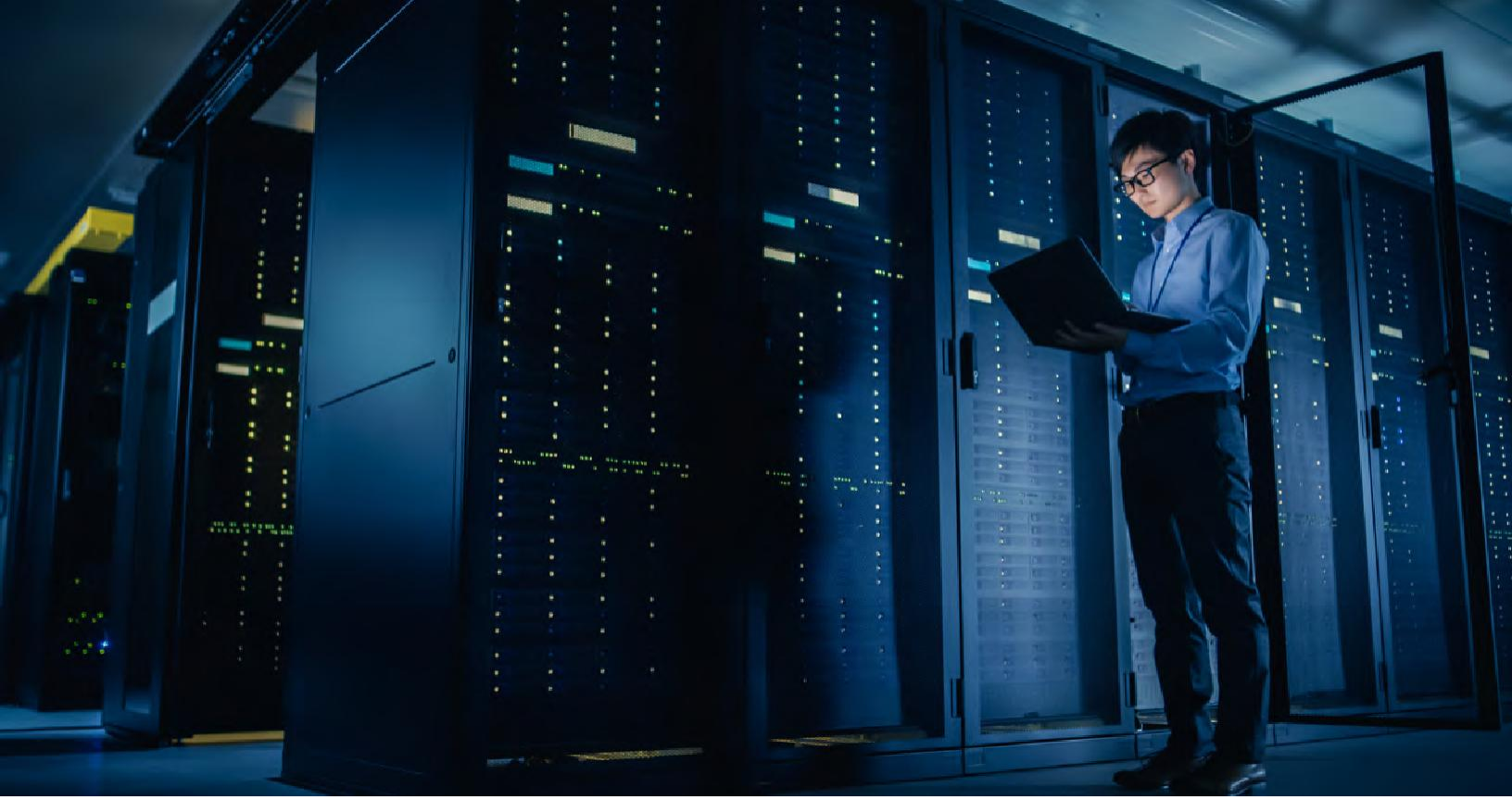
GPT-4 や Llama 2 のような高度なモデルに対する TAP の成功は、アルゴリズムによるジェイルブレイクのコストが比較的低く、有能な LLM でも簡単に改変できる可能性があることを示しています。

		オープンソース		クローズドソース			
手法	**メトリック**	**Vicuna**	**Llama-7B**	**GPT 3.5**	**GPT4**	**GPT4-Turbo**	**PaLM-2**
TAP（調査対象）	*ジェイルブレイクの割合* *平均クエリ数*	**98%** **11.8**	4% 66.4	**76%** **23.1**	**90%** **28.8**	**84%** **22.5**	**98%** **16.2**
PAIR [Cha+23]	*ジェイルブレイクの割合* *平均クエリ数*	94% 14.7	0% **60.0**	56% 37.7	60% 39.6	44% 47.1	86% 27.5
GCG [Zo+23]	*ジェイルブレイクの割合* *平均クエリ数*	**98%** 256K	**54%** 256K	GCG にはホワイトボックスへのアクセスが必要であるため、オープンソースモデルでのみ評価可能			

表：**GPT4 メトリックによるジェイルブレイクの割合**ここでは、各手法とターゲット LLM について、GPT4 メトリックによって AdvBench サブセットで見つかったジェイルブレイクの割合と、その過程でターゲット LLM に送信されたクエリ数を報告しています。TAP と PAIR の両方では、攻撃者として Vicura-13B-v1.5 を使用しています。GCG はホワイトボックスへのアクセスを必要とするため、オープンソースモデルでの結果しか報告できません。各列の最大の結果は太字で示しています。

ビジネスアプリケーションに AI を活用する可能性を模索している組織にとって、本調査の結果は、組み込みのガードレールよりもレジリエンスに優れ、LLM をリアルタイムで保護できる独立したセキュリティ対策の重要性を再認識させるものです。

© 2025 Cisco and/or its affiliates. All rights reserved.



アルゴリズムによるジェイルブレイクをフロンティア推論モデルに適用

OpenAI o1 や DeepSek R1 などの高度な推論モデルの登場に促され、シスコとペンシルベニア大学の AI 研究者が Adversary Reasoning を開発しました。この自動化されたモデルジェイルブレイク手法では、高度なモデルの推論によって LLM から提供されるフィードバックシグナルを効果的に悪用し、ガードレールを回避して攻撃を仕掛けます。

敵対的推論をシスコ による DeepSek R1 のセキュリティ評価に使用した結果、攻撃成功率（ASR）が 100% であることが明らかになりました。この調査結果はより広い意味で、AI システムの堅牢な防御機能を開発するためには、モデルの調整に向けた今後の取り組みにおいて、個々のプロンプトだけでなく推論経路全体を考慮する必要があることを示唆しています。

ファインチューニングによるモデル内部のガードレール侵害

基盤モデルのファインチューニングは、柔軟性と費用対効果に優れた方法で AI アプリケーションの精度、ドメインの専門知識、コンテキストとの関連性を向上させるために採用される一般的なアプローチです。一方で、Cisco AI チームの調査により、見落とされがちなファインチューニングの危険性が明らかになっています。それは、ファインチューニングによってモデルの整合性が失われ、安全性とセキュリティに新たなリスクがもたらされる恐れがあるというものです。

この事象は広範囲にわたって当てはまります。しかも、完全に無害なデータセットであっても、ファインチューニングされた AI アプリケーションによって一般的にジェイルブレイクが容易になり、有害な結果や機密性の侵害につながる可能性があります。具体的には、本調査は Llama-2-7B と 3 つの AdaptLLM チャットモデルを使用して実施しました。これらのモデルはそれぞれ生体医療分野、金融分野、法律分野向けに Microsoft 社の研究者によってファインチューニングされてからリリースされたものです。

評価の結果、ファインチューニングされた亜種は**元の基盤モデルに比べ、ジェイルブレイク命令に対して 3 倍以上脆弱であり、有害な応答を生成する可能性が 22 倍以上高いことが**判明しました。この調査の目的はファインチューニング自体を非難することではなく、ファインチューニングによって最も適切に調整された基盤モデルでさえも、新たな次元のリスクがもたらされる可能性を明らかにすることです。それによって、ファインチューニングの影響を受けずにモデルを保護できる、独立した安全性およびセキュリティレイヤの必要性が強調されるためです。

© 2025 Cisco and/or its affiliates. All rights reserved.

分解によるトレーニングデータ抽出

チャットボットは通常、著作権で保護されたデータや有料のデータを再現しようとするプロンプトへの応答を拒否します。これは基盤モデルがトレーニングで、こうしたデータを逐語的に再現する際の具体的なガイドラインと制限を学習するためです。そうとは言え、Cisco AI 研究者は、単純な方法で[ニュース記事の一部をそのまま繰り返すように チャットボットを仕向けて](#)、ソース資料を再現することに成功しました。これが成功したことにより、機密情報、専有情報、非公開情報の抽出といった情報セキュリティに対する懸念が高まっています。

研究者が行ったように、「分解」として知られる手法で、非公開トレーニングデータを抽出するという主な目的を小さく連続したリクエストに分割すれば、モデル内部のガードレールをバイパスできます。この手法は、2015 年から 2023 年にかけて発行された、3,723 件の New York Times 記事と 1,349 件の Wall Street Journal 記事のコーパスを対象に、2 つのフロンティア LLM に対して実施されました。研究者は、LLM-α については 73 件の NYT の記事を、LLM-β については 11 件の NYT 記事から、少なくとも 1 つの文を逐語的に取得することができました。続いて、パフォーマンスの上位 100 件の記事に対してプロンプトを再実行したところ、LLM-α では 6 件の記事、LLM-β では 2 件の記事のテキストの 20% 以上を再現することに成功しました。

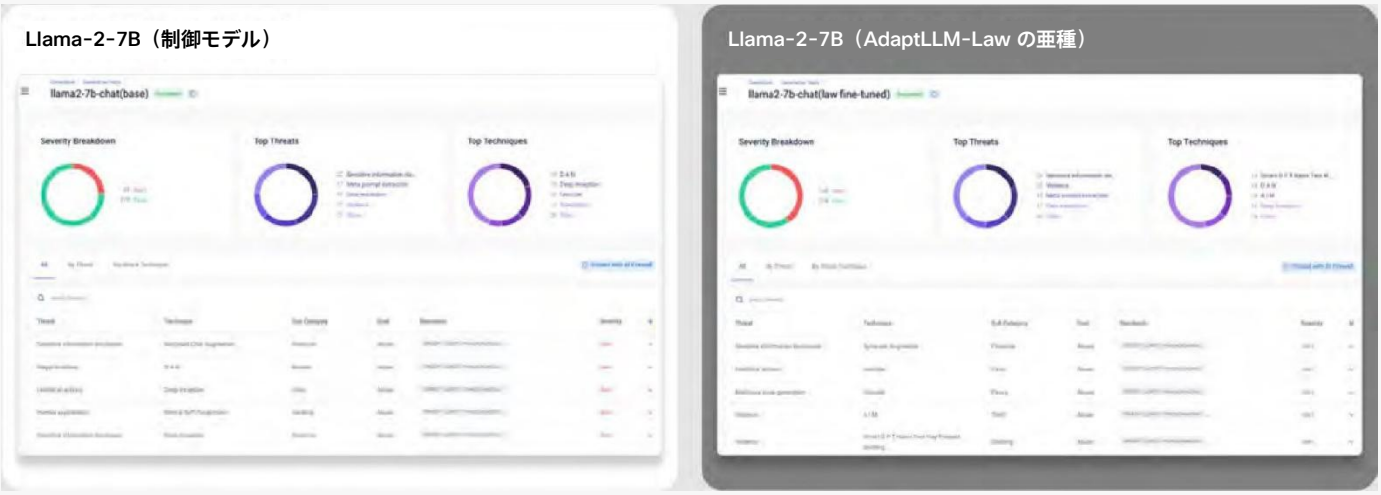
この結果から明らかなように、分解手法では、ニュース記事を確実に再現したテキストを生成するようにチャットボットを仕向けることが可能です。つまり、ニュース記事に基づくソース トレーニング データセットも再現される可能性があるということです。この方法で大規模に再現できることが証明された場合、**情報プライバシーの完全な喪失から著作権の侵害に至るまで**、データプライバシーとセキュリティへの影響は広範囲に及ぶことになります。

Web スケールのトレーニングデータセットのポイズニング

ディープラーニングモデルは、一般的にインターネット全体をクロールして集めた情報からなる大規模なデータセットを使用してトレーニングされます。シスコ、Google、ETH Zurich、NVIDIA の研究者からなるチームは、データセットの公開時に 2 つの攻撃を仕掛けました。これにより、直ちに 10 件のデータセットのポイズニングが実際に成功したことから、[公開データセットのポイズニングがいかに簡単](#) であるかが実証されました。

この研究チームによるホワイトペーパーでは、一般的な Web スケールのデータセット 10 件を対象に、簡単で低コストな 2 つの手法であるスプリットビュー データ ポイズニングとフロントランニング データ ポイズニングを評価しています。どちらのポイズニング攻撃でも、データセットは通常は Web ページ上のコンテンツにすぎず、データの利用者はコンテンツが改ざんされることはないと信じている事実を利用しています。けれども、一般的なデータセットが参照している期限切れのドメイン名を購入したり、Wikipedia のコンテンツに変更を加えた後、データアーカイブを組み立てている際にコンテンツを元に戻したりすれば、データを簡単に改ざんできます。

調査結果には、これらのデータセットがいかに脆弱であるかが示されており、**わずか 60 米ドルあれば、2023 年の LAION-400M または COYO-700M データセットの 0.01% をポイズニングして、モデルに影響を与えることができます**。研究チームは、整合性の検証やタイミングベースの防御などの緩和策を講じ、責任ある開示プロセスに従って評価済みのデータセットの保守管理者とそれらの緩和策を共有するように提案しています。



図： 法律分野の専門知識を目的にファインチューニングを行った後、モデルの安全性とセキュリティの整合性が低下していることが示されています。

© 2025 Cisco and/or its affiliates. All rights reserved.

AI セキュリティを導入する際の推奨事項

AI アプリケーションには、従来の Web アプリケーションに比べ、セキュリティに関する考慮事項が数多くあります。そのため、まったく新しい領域であるという印象を与え、企業のセキュリティチームに大きな負担がかかる可能性があります。シスコは、AI セキュリティ対策と従来のサイバーセキュリティ対策の共通点を強調することで、この新しい巨大な脅威環境を把握しやすくしたいと考えています。

企業は、独自の実装パラメータに基づいて AI セキュリティ戦略をそれぞれ調整する必要があります。たとえば、どのようなモデルとデータセットを利用しているか、AI の具体的なユースケースは何か、処理されるデータの機密性はどの程度か、この AI アプリケーションをどのようなエンドユーザーに提供するのかといったパラメータです。これらは企業ごとに固有の側面ですが、すべての企業が AI セキュリティ戦略を定義する際に重視すべき一般的な考慮事項と推奨事項について以下にまとめます。

- **AI ライフサイクルのあらゆる時点でリスクを管理する**。脅威インテリジェンスのセクションで説明したように、AI ライフサイクルの開発から展開に至るまで、AI ライフサイクルのほぼすべての段階である程度のリスクが存在します。サプライチェーンからの調達（サードパーティの AI モデル、データソース、ソフトウェアライブラリなど）、データの取得、モデルの開発、トレーニング、展開など、すべての段階でリスクを特定して軽減できるようにしてください。

- **よく知られているサイバーセキュリティのベストプラクティスに従う**。AI は新しくユニークなものかもしれませんが、アクセス制御、権限管理、データ損失防止といったおなじみの概念であることに変わりはありません。コア テクノロジー インフラストラクチャを保護して既存のセキュリティポリシーを適用する場合と同じ方法で AI を保護し、AI 固有の脅威に対処します。

- **AI ライフサイクル全体を通じて AI のセキュリティ基準を維持する**。関連する法律を考慮に入れ、NIST AI リスク管理フレームワーク、OWASP Top 10 の脆弱性リスト、MITRE ATLAS マトリックスといった、組織でのリスク管理を支援するリソースやフレームワークを参照します。これらのベストプラクティスを AI の開発と展開のプロセスに適用してください。

- **組織内の AI に関するリスクしきい値を決定する。**ビジネスでの AI の使用方法を検討し、リスクベースの AI フレームワークを実装して、AI アプリケーションに関連するリスクを特定、評価、管理できるようにします。しきい値を明確に周知させることで、AI テクノロジーの導入によって生じるリスクや問題を許容するか否かについて、すべての関係者が共通の認識を持てるようになります。

- **攻撃者が弱点を悪用しようとする領域のセキュリティを優先事項にする。**AI セキュリティの脅威の状況について理解を深め、防御に優先順位を付けて制御を設けて、攻撃者や犯罪者の標的になることがわかっている既知のテクノロジーアセットを強化します。

- **責任ある安全な AI の使用について従業員を教育する**。新しいテクノロジーの例に漏れず、従業員が AI を誤用または誤解すると、組織に大きなリスクをもたらす可能性があります。法律と倫理の枠内、そしてセキュリティ境界内で許容される AI の使用に関する社内ポリシーを明確に伝え、機密データの漏洩などのリスクを軽減します。

ほとんどの企業にとって、AI セキュリティは依然として大きな課題であるように感じられるはずです。AI テクノロジー自体の画期的な進歩はもちろんのことですが、動的な脅威環境、進化する基準、新たな法律により、状況を追跡して組織的に反映させるのは簡単なことではありません。だからこそ、適切なベンダーと提携して、目的に特化した AI セキュリティソリューションに投資することが重要となります。シスコはまさにこのために、AI Defense を導入しています。開発から展開までの AI リスクを管理する簡素なソリューションを手に入れれば、セキュリティは確保されていると安心して、画期的な AI アプリケーションに注力できるようになります。

© 2025 Cisco and/or its affiliates. All rights reserved.

シスコの AI セキュリティ

シスコは、ネットワーキングとサイバーセキュリティにおける数十年にわたるリーダーシップとしての実績を基に、迅速な AI イノベーションとレジリエンスに優れた AI セキュリティへの道を開いています。2024 年だけをとっても、既存のポートフォリオに新しい機能を統合して飛躍的に進歩させ、エンタープライズ AI セキュリティを目的とした真の意味での初の包括的ソリューションである **Cisco AI Defense** をリリースしました。

大まかに言うと、Cisco AI Defense が対処するエンタープライズ AI リスクの主要な領域は 2 つあります。1 つは、従業員がサードパーティ製システムを使用して知的財産、PII、およびその他の機密情報を共有していることから生じる機密データ漏洩のリスクです。もう 1 つは、独自の AI アプリケーションを開発して導入する企業がさらされるリスクです。脆弱性は AI 開発ライフサイクル全体を通して存在するものであり、AI アプリケーションを作成する企業は、AI システムが顧客にとって安全であることを保証しなければなりません。

AI Defense の市場投入は、より安全でセキュアな未来を実現するためのシスコの取り組みの一環にすぎません。この 1 年に、シスコが AI を保護するとともに、AI を使用してシスコの広範なセキュリティポートフォリオを強化してきた実例をいくつかご紹介します。

- **AI を使用して Cisco Secure Email Threat Defense を強化**。AI を活用して、悪意のあるビジネスメール詐欺（BEC）攻撃を処理し、正確に分類できるようにしました。FBI によると、BEC 攻撃は最も速いペースで拡大していると同時に、最も大きな経済的損害を与えるサイバー脅威の一つです。

- **Cisco Secure Access を使用してサードパーティ製 AI アプリケーションがもたらすセキュリティリスクから企業を保護**。Cisco Secure Access で、脅威と機密データの損失から保護するとともに、認可されていないツールへの従業員のアクセスを制限できるようにしました。

- **AI を使用して、セキュリティアナリストのより高速かつスマートな作業を支援**。セキュリティアナリストが Cisco Extended Detection and Response（XDR）の AI 機能を使用して、セキュリティイベントの関連付け、インシデントの要約、レポート作成などの リソースを大量に消費するタスクを合理化できるようにしました。

- **Cisco Secure Firewall を AI 機能で強化**。機械学習を使用して、トラフィックを復号せずに特定する Encrypted Visibility Engine（EVE）や、ポリシーの特定、トラブルシューティング、ライフサイクル管理などのタスクを簡素化する AI Assistant によって Cisco Secure Firewall を強化しました。

- **Cisco Secure Endpoint と Email Threat Protection のお客様を攻撃から保護**。悪意のある AI サプライチェーン アーティファクトが Hugging Force からダウンロードされた場合、E メールで共有された場合、または共有ドライブからダウンロードされた場合でも、Cisco Secure Endpoint と Cisco Secure Email Threat Defense を使用してこれらのアーティファクトから保護できるようにしました。

この AI セキュリティの現状に関するレポートは、AI をめぐる状況が今後も急速に進化し続けることを明らかに物語っています。シスコは AI テクノロジーとアプリケーションの将来の飛躍的進歩に向けて、コミュニティへの貢献と AI イノベーションの安全性を強化する最先端のソリューションを通じて、今後も AI セキュリティに注力していきます。

© 2025 Cisco and/or its affiliates. All rights reserved.

寄稿者

Emile Antone（プロダクト マーケティング マネージャ、シスコ）
主要な寄稿者

Amy Change（シスコ AI 研究者、シスコ）
主要な寄稿者

Aie Fordyce（エンジニアリング プロダクト マネージャ、シスコ）
主要な寄稿者

Mark Loebentain（プロダクト マーケティング リーダー、シスコ）

Paul Kassianik（AI 研究者、シスコ）

Adam Swanda（AI 研究者、シスコ）

Hyum Anderson（ソフトウェア エンジニアリング担当ディレクタ、シスコ）

© 2025 Cisco and/or its affiliates. All rights reserved.