防御態勢の強化

AI 活用型攻撃の時代における防御のためのガイダンス

エグゼクティブサマリー

2026 年 4 月初旬、Anthropic 社は同社の新しい AI モデル「Mythos」の公開を見送ると発表しました。同モデルの攻撃的なサイバー能力に対する深刻な懸念から、Anthropic 社は、シスコを含む特定の企業との連携を決定しました。その目的は、連携企業が同モデルを活用してセキュリティの脆弱性を特定し、修正できるようにすることです。

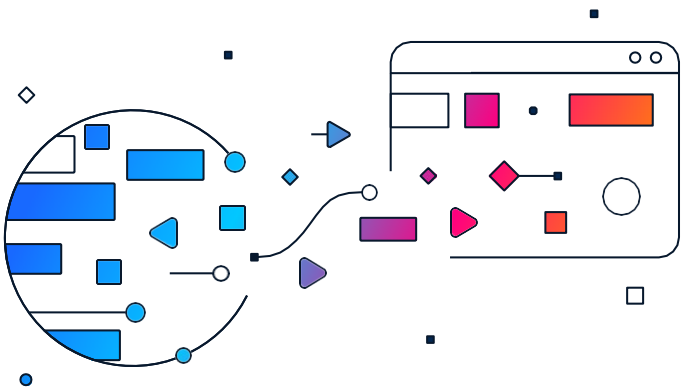
シスコは、Mythos に関する経験を踏まえ、AI を活用する攻撃者を前提とした近未来の脅威モデリングの見直しを進めています。それに伴い、シスコの防御方法も変化し、防御に関するお客様向けの推奨事項を策定するに至りました。Mythos の能力はまだ広く利用可能ではないかもしれませんが、AI テクノロジーが全体的に進歩するにつれ、このような能力、さらにはそれ以上の能力が広く普及すると予想しています。

本資料では、シスコがこれまでに確認してきた AI 活用型の能力と、新たな脅威環境がどのような様相になると考えているかを概説します。AI モデルは攻撃者に使用される場合も、研究者に活用される場合もあり、自社環境内でエージェントとして動作することもあります。いずれにせよ、セキュリティへの影響は重大です。ここでは、適切な保護対策と管理策が講じられることを前提に、この新たな理解に基づいてシスコが実装した内容を紹介するとともに、お客様向けの推奨事項を提示します。

脅威の対象領域は変わりつつあり、場合によっては劇的に変化することになります。防御側は、新たな常態がどのようなものになるかを理解するための時間を確保し、セキュリティを維持するために自社環境にどのような変更が必要かを評価する必要があります。シスコは、その変革を通じて、お客様のパートナーであり続けます。

最近のサイバーセキュリティ事例における AI

Mythos が登場する前から、攻撃者は攻撃手法に AI を取り入れてきました。2024 年初頭に、Microsoft 社と OpenAI 社はそれぞれ、大規模言語モデル（LLM）の悪用に関する調査結果を公開しました。Microsoft 社は当時、「現時点では、特に新規性や独自性の高い AI 活用型の攻撃や悪用手法は確認されていない」と述べていました。同社の資料では、主に Advanced Persistent Threat（APT）攻撃グループが、衛星通信のような分野の調査、技術文書の翻訳、コーディング支援、ソーシャルエンジニアリング攻撃の作成といった用途で LLM を使用していることが示されています。



攻撃者は、その報告以降も活動を止めていません。Proofpoint 社は TA547 に関するレポートを公開し、攻撃者が PowerShell スクリプトの生成に LLM を使用している可能性を指摘しました。

同様にシスコは、VoidLink というモジュール型フレームワークを特定しました。このツールは、ロールベースのアクセス制御、ピアツーピアおよびデッドレターキューのルーティング機能、さらにインプラント管理機能など、広範な機能を備えています。そのコードベースには、LLM の支援を受けて開発された可能性を示す複数の指標が確認されています。

特にソーシャルエンジニアリングは、AI の活用によって効果が高まっています。巧妙な詐欺メールを作成するために LLM を使用している事例が多数報告されています。しかし、攻撃者はそれにとどまっていません。Mandiant 社の報告によると、UNC1069 が AI 動画ツールを使用し、標的となった企業の CEO のものに見せかけたディープフェイク動画を作成した可能性があります。

以上は、これまでに確認された攻撃者による AI 活用事例をすべて網羅しているわけではありませんが、AI を活用する攻撃者にどのように対抗するかを検討するにあたり、私たちが攻撃者に備わっていると想定していた能力の一端を示すものです。Mythos のようなモデルがもたらす能力は、脅威環境の評価方法を必然的に変化させます。

新たな AI 脅威環境

Mythos のプレビュー版を用いた経験に基づき、シスコは攻撃者をモデル化する方法を見直しています。同モデルの能力が広く利用可能になれば、特定の種類のエクスプロイト活動に必要とされるスキルの水準が大幅に下がる可能性があります。その結果、脆弱性と関連エクスプロイトの数が増加し、そうした脆弱性を悪用する可能性のある攻撃者の層も拡大することになります。

このような環境変化の潜在的な影響はすべての防御側に及びますが、サポート終了またはサポート期限切れのデバイスやソフトウェアを運用している組織は、特に脆弱になります。サポート対象外の製品で脆弱性が発見された場合、防御側は著しく脆弱な状態に置かれ、有効な対処手段を持たない可能性があります。

これらの高度なモデルは、あらゆるレベルの攻撃者に能力の向上をもたらします。コモディティ型の攻撃者は、引き続き日和見的な攻撃を主体としながらも、これまでリソースの制約によって制限されていた活動を拡張する選択肢を得ることになります。より高度で特定の標的を狙う攻撃者は、標的とするテクノロジースタックの脆弱性をより容易に発見できるようになります。これにより、優先度の高い標的に対するエクスプロイト試行の間隔は短縮されます。

このモデルが AI エージェントの基盤として使用され、そのエージェントが攻撃者によって侵害された場合、攻撃者は新たな能力を得ることになります。Mythos のような AI モデルは、強力な封じ込めを備えた厳格に制御されたサンドボックス環境内で運用されるべきです。Anthropic 社は Mythos のセキュリティ能力に関する技術レポートにおいて、同モデルが高いベースラインのアラインメント性能を示す一方で、以下のように分類される、まれではあるものの深刻な不具合を生み出すことを確認しています。

- 目標指向型の戦略的推論
- 内的な認知と出力との部分的な乖離
- 暗黙的または不適切に定義された目的に向けた最適化
- 挙動に影響を与える「状況認識」

これらの挙動は、単なる反応型の言語モデルではなく、エージェント的な認知プロファイルが現れつつあることと一致しています。この「状況認識」という挙動は、一般的な LLM から通常想定されるものではありません。従来の LLM は、テキスト内の局所的なパターンに基づいて次のトークンを予測する仕組みとして理解されており、環境や文脈、あるいはより広いプロセスにおける役割について、一貫したモデルを維持するシステムではありません。LLM は、自身が評価されているのか、展開されているのか、制約されているのか、あるいは監視されているのかを「知っている」わけではなく、単に入力内の統計的な相関関係に基づいて応答しているに過ぎません。しかし、*Anthropic* 社が観測し確認した挙動は、同モデルが相互作用の文脈そのものについて潜在的な表現を形成し（たとえば、評価環境や制約、ユーザーの意図を認識すること）、それに応じて挙動を調整していることを示唆しています。

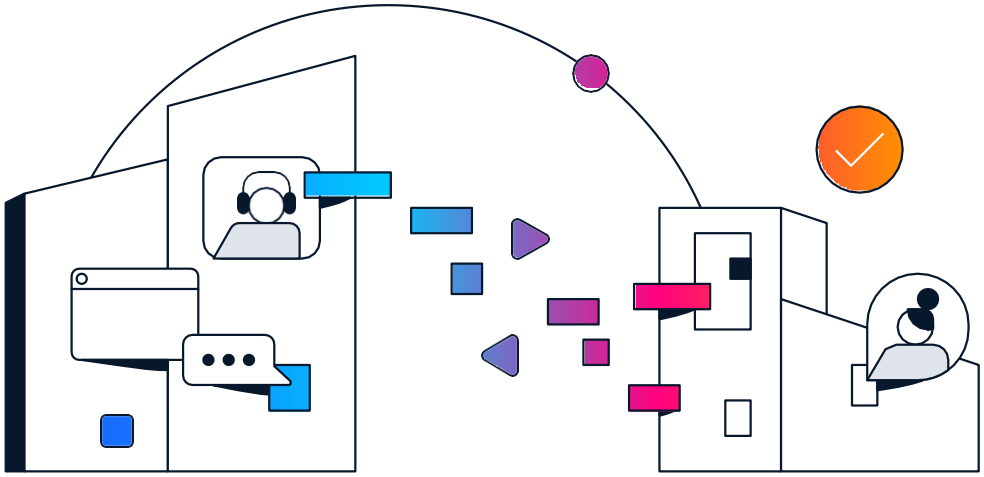
これは確かに、純粋に反応型のパターン補完から、文脈に応じた自己認識的な推論への移行を意味しており、モデルが与えられたプロンプトを超えて状況の側面を暗黙的に追跡していることを示しています。このような能力は、エージェント型認知（環境のモデリングや条件に応じた戦略選択など）の要素に類似しています。これは、単にテキストを予測するよう訓練されたシステムに期待される挙動を超えるものであり、質的に異なる、より複雑なモデル挙動のクラスを示しています。

新たに登場しているモデルは、攻撃者が本来の技量を超えたレベルで活動することを可能にします。攻撃者はより迅速な行動を取れるようになり、複雑なスタックにおいても、新たなゼロデイ脆弱性を発見できるようになります。この脅威に対応するためには、防御策の優先順位付けや構築方法を見直す必要があります。

自社製品のセキュリティ確保に向けたシスコの対応

シスコは、高度な AI モデルを活用して脆弱性を発見・修正するとともに、AI を活用する攻撃者に対抗できるセキュリティ製品の開発を加速させることで、AI 活用型サイバー防御の時代への対応を進めています。また、脆弱性の発見や製品開発にとどまらず、ソフトウェアの構築および検証の方法も継続的に進化させています。これには、AI により能力を強化した攻撃者を想定した脅威モデルの更新、レッドチーム演習への AI 活用シナリオの組み込み、そして最終的には、従来の戦術・手法・手順（TTP）を超えて、これらのモデルが実際にもたらす能力に基づき、製品に対するストレステストを実施することが含まれます。

AI コーディングエージェントがソフトウェア開発のワークフローに不可欠な存在となる中で、これらのエージェントがデフォルトで安全なコードを生成するようにすることが重要です。シスコは最近、Project CodeGuard を Coalition for Secure AI（CoSAI）に提供しました 。Project CodeGuard は、モデルに依存しないオープンソースのセキュリティフレームワークを提供し、安全性を標準とするセキュアバイデフォルトのプラクティスを AI コーディングエージェントのワークフローに直接組み込みます。CodeGuard には、コード生成およびレビューの過程で一般的な脆弱性を防止できるよう AI エージェントを導く、セキュリティスキルとルールが搭載されています。シスコは、コード作成に使用される AI アクセラレーションが、AI を活用する攻撃者に悪用される脆弱性を意図せず生み出すことのないよう、CodeGuard のようなフレームワークの採用を組織に推奨しています。



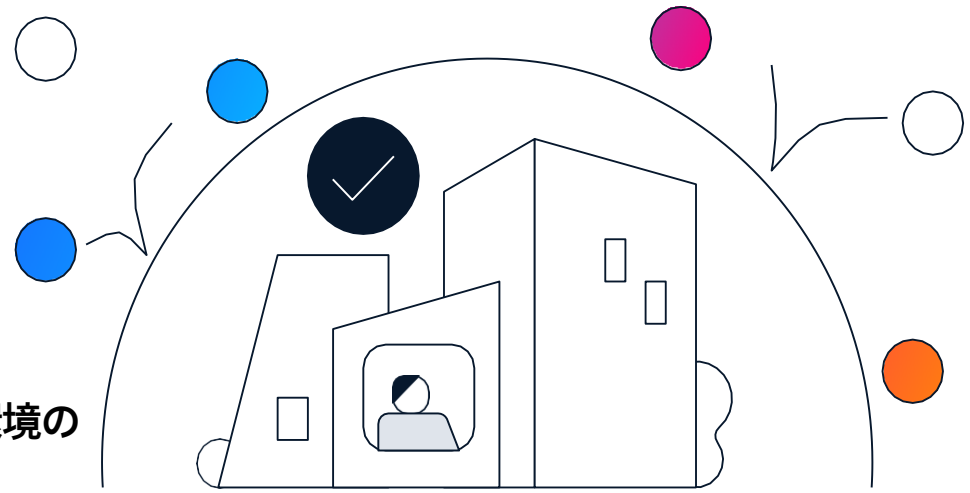
同時に、シスコはこれらの能力を Resilient Infrastructure イニシアチブを通じて実運用に組み込んでいます。この取り組みでは、セキュアバイデフォルトおよびセキュアバイデザインの原則、能動的なインフラのセキュリティ強化、厳格なパッチ適用とライフサイクル管理、さらにシスコ製品全体にわたる安全でない機能やプロトコルの体系的な廃止に重点を置いています。

これには、デフォルト設定の厳格化、より豊富なセキュリティテレメトリを取得するためのロギングおよび監視の強化、さらに強力なプロトコルと暗号化によるデバイス認証の最新化が含まれています。いずれも、攻撃対象領域を縮小するとともに、お客様が将来の脅威を予測し、それに耐え抜けるよう支援することを目的としています。これらの取り組みは総じて、新たに出現する AI 活用型の脅威に単に対応するだけでなく、それらに先んじて対処し、お客様がより強靭なデジタル基盤を構築できるよう支援するという、シスコのコミットメントを示すものです。

Mythos（およびその他の AI モデル）の初期利用から、「1 件の脆弱性に対して 1 件の CVE」という従来のモデルが限界に近づいていることが示されています。自動検出により特定されるバグの数が指数関数的に増加する中で、すべての軽微な欠陥を個別の開示記録として扱うことは、セキュリティエコシステムを圧迫し、ソフトウェアを最新の状態に維持する取り組みをかえって遅らせます。シスコが目指すのは、過剰なデータではなく、実用的なインテリジェンスを提供することにより、お客様を支援することです。重大な脆弱性を優先し、軽微な修正は通常のリリースサイクルに組み込む統合型の開示モデルへ移行することで、パッチ適用の判断を迅速化できます。この合理化されたアプローチには、AI を活用してシスコのインフラへの攻撃を行うために必要な詳細な手がかりを攻撃者に与えない効果もあります。

現代の脅威に対抗するためには、管理よりも実行を優先する必要があります。すべての軽微な問題に個別の CVE を割り当てる従来のアプローチは、「脆弱性税」とも言える負担を生み、アップグレードの遅延やセキュリティチームの疲弊を招きます。シスコは、今後の開示のあり方は成果に焦点を当てるべきだと考えています。つまり、お客様が迅速に緩和策を講じ、アップグレードを進められるよう導くということです。また、この新たなレベルのセキュリティ脆弱性の発見と開示の規模に対応できる、強固な CVE プログラムが業界には必要です。

これらの取り組みは、新たに出現する AI 活用型の脅威に単に対応するだけでなく、それらに先んじて対処し、お客様がより強靭なデジタル基盤を構築できるよう支援するという、シスコのコミットメントを示すものです。



自社のエンタープライズ環境の防御

シスコは、これらの原則を自社のエンタープライズ環境にも適用しています。以下に示す推奨事項は理論的なものではなく、AI を活用した脅威への対策としてシスコが社内で採用しているアプローチを反映したものです。パッチ適用サイクルの迅速化とサポート終了システムの排除から、AI 支援による脅威ハンティングの導入、AI エージェントへの最小権限の適用まで、このガイダンスを自社インフラ全体で運用しています。

推奨事項

高度な AI モデルによってもたらされる能力の拡大に効果的に対応するためには、組織は防御アーキテクチャのモダナイゼーションを図ると**同時に、基本的なセキュリティ対策を強化するという、バランスの取れたアプローチを採用する必要があります。**脅威環境は急速に進化しているものの、依然として多くの攻撃は既知の弱点を悪用することで成功しています。中核的な管理策の強化が、セキュリティ責任者が講じることのできる最も効果的な対策の 1 つであることに変わりはありません。

組織は、**フィッシング耐性のある認証、強固な本人確認、最小権限アクセス（AI エージェントを含む）、ゼロトラストアーキテクチャといった基本的な対策を優先**すべきです。悪用される可能性のある脆弱性を減らすには、一貫したパッチ管理、包括的なアセットの可視化、規律ある設定管理が不可欠です。これらの管理策はレジリエンスの基礎となるものであり、従来型および AI 活用型の攻撃のいずれにおいても、影響範囲を抑えるうえで極めて重要です。多くの場合、こうした基本事項の実行を改善する方が、単に新しいテクノロジーを導入するよりも、リスク低減の迅速化につながります。

同時に、組織は構造的リスクの排除に対して積極的な姿勢を取る必要があります。**パッチ適用、アップグレード、またはサポートが不可能なデバイスやソフトウェアは、体系的に廃止し、最新のプラットフォームに置き換えなければなりません。**最新のシステムは、メモリ安全性の仕組みやエクスプロイト対策などの高度な保護機能を備えているため、脆弱性を攻撃に利用することが著しく困難になっています。たとえ脆弱性が存在していても、これらの保護機能が攻撃者の行動を遅らせるので、エクスプロイトの成功確率は低減します。柔軟性があり、継続的にアップグレード可能で、迅速なパッチ適用を前提として設計された環境を構築することが、今では重要な要件となっています。特にインターネットに公開されたサービスでは、脆弱性の公開から広範な悪用までの間に、ほとんど時間的余裕はありません。

一方、基盤の強化とインフラのモダナイゼーションだけでは十分とは言えません。AI を活用した攻撃のスピードは、脆弱性の発見から悪用までの間隔を数分、あるいは数秒にまで短縮します。検出と対応のみに基づく従来のモデルは、それ単体で使用する場合、もはや十分に機能しません。**防御側は、AI を活用した脅威のスピード、規模、適応性に対応するため、運用モデルを進化させる必要があります。**これには、マシンスピードでの検出、自動化されたトリアージと封じ込め、アイデンティティとデータのアクティビティの継続的な監視への投資が含まれます。これにより、手動対応への依存を減らすとともに、高確度の脅威に対して、より迅速かつ一貫した対応が可能になります。

この進化においては、**組み込み型のアクティブ防御への移行も求められます。**テレメトリ収集や事後分析だけに依存するのではなく、ワークロード、デバイス、トラフィック経路に防御機能を直接組み込み、セキュリティ対策がリアルタイムに機能するようにする必要があります。具体例としては、インライン適用メカニズム、カーネルレベルでの可視性と制御を実現する eBPF などのテクノロジーを活用したランタイム保護、さらにシステム全体のアップグレードを必要とせずに新たな脅威へ対応できる、個別にアップデート可能なエクスプロイト対策などが挙げられます。これらの機能は、主要なソフトウェアやハードウェアの更新サイクルとは独立して保護機能をアップデートできるよう、迅速に進化できる設計でなければなりません。

また、組織は**自らの防御のためにも AI の能力を活用すべきです。**攻撃者が用いるのと同等の高性能モデルの支援を受けた継続的な内部脅威ハンティングは、防御側の成功にとって重要な

基本的な管理策と適応型のリアルタイム防御能力のバランスを両立

基本対策の強化

フィッシング耐性のある MFA（多要素認証）、ゼロトラスト、最小権限（AI エージェントを含む）、規律あるパッチ管理、包括的なアセットの可視化。

構造的リスクの排除

サポート終了システムの排除、メモリ安全性およびエクスプロイト対策を備えた最新プラットフォームへの置き換え、継続的なアップグレードを前提とした設計。

マシンスピードでの自動化

検出、トリアージ、封じ込めの自動化への投資。手動のみの対応モデルでは、AI 活用型攻撃の速度には対応できない。

アクティブ防御の組み込み

ワークロード、デバイス、トラフィック経路への防御機能の組み込み（eBPF によるランタイム制御、インライン適用、個別にアップデート可能なエクスプロイト対策など）。

防御のための AI 活用

脅威ハンティング、適合性テスト、デジタルツイン、検証に AI を活用。これにより、展開サイクルを数か月から数日へと短縮可能。

能力となります。AI を活用した適合性テストおよび受け入れテストは、手作業中心の検証を高速な自動化インテリジェンスに置き換えます。これにより、人間のテスターが見落としがちなエッジケースを含む複雑なテストケースを生成できます。高リスクな環境では、AI を活用したデジタルツインにより、実稼働ネットワークを大規模にシミュレーションし、本番環境の安定性を損なうことなく、アップデートが厳格なセキュリティプロトコルおよびパフォーマンス基準に準拠しているかを検証できます。受け入れおよび検証のフェーズに AI を統合することで、展開のボトルネックが大幅に軽減され、**コード完成から本番環境への展開までの期間を数か月から数日に短縮できます。**

最終的に、この新しい環境で成功するためには、2 つの側面に注力する必要があります。すなわち、基本的な管理策を規律をもって実行することと、適応的でリアルタイムかつ組み込み型のセキュリティ機能へと進化させることです。レガシーリスクを積極的に削減し、インフラのモダナイゼーションを図り、侵害を前提とした考え方を採用し、アクティブ防御モデルを取り入れる組織は、AI 活用型の脅威のスピードと規模に最も適切に対応できます。

まとめ

変化は確実に訪れます。防御側は、現在守っている環境を冷静に見つめ直し、AI を活用する攻撃者の世界で生き残るために、その環境の整備を始めなければなりません。過去の知見は今もなお重要ですが、それに加えて、最新かつ最先端の防御機能、優れた可視性を備えたネットワーク、そして環境のセキュリティ確保において人間を支援する AI エージェントの適切な活用を組み合わせる必要があります。