2026 年 5 月 21 日　全 13 頁

今話題の Claude Mythos 騒動をまとめる

～マインドチェンジとスピードアップは必要だが、本質的な対策は今までと同じ～

フロンティア研究開発センター　シニア・フェロー
プリンシパル・データサイエンティスト　坂本 博勝

「AI にハッキングをされて、会社の機密データが盗まれた・・・」
生成 AI テクノロジーがこのまま発達すれば、「いつかすぐにそんな時代が来るよ」とあちこちで語られてきた噂が、こんなにも早く実現してしまうのだろうか(2026 年 5 月現在)。
アンソロピックの Claude Mythos(クロード・ミュトス)の件である。

大企業を中心に、Mythos について「どう考えるべきか」「何を対策するべきか」という、ひとつの騒動になっており、特に金融業界では動きが盛んなようだ。
ビジネス界の関心が高いトピックスでもあるので、今回は、Claude Mythos 騒動の経緯や内容、そして実態として企業が備えるべき対策について、現時点での最新情報をまとめてみよう。
まあ結論から言ってしまうと、懸念を払しょくできる決定的な対策は、現時点ではまだ存在しないとしか言えないのだが・・・

Mythos 騒動の経緯

まず、Mythos 騒動の、現時点(2026 年 5 月 12 日)までの経緯を一覧でまとめる。(※1)

騒動の始まりは 4 月 7 日。アンソロピックが Claude Mythos Preview と Project Glasswing について発表したことだった。(※2)(※3)

1.自社 AI モデル、Claude の精度向上の過程で、Claude Mythos Preview という AI 基盤モデルが出来上がった。

2.サイバーセキュリティ能力/脆弱性発見能力が高すぎ、悪用の恐れがあるので、Mythos の一般公開はしない。

3.プログラムソースコードからサイバーセキュリティ脆弱性を発見する能力が非常に高く、また、発見した脆弱性を突いてサイバー攻撃を成功させるための攻撃ソースコードを生成する精度も高い。

4.グルーバル団体約 50 団体で Project Glasswing を組成し、Mythos モデルは参画団体に限定して公開する。プロジェクト内で検証と対策検討を進め、90 日以内に評価レポートを公開す

る。

5.Mythos は、アンソロピック技術者が設定した閉鎖 IT 環境から、独力で脆弱性を見つけ出してハッキングを行い、インターネット環境に脱出した。

6.Mythos が、オープンソースである Linux の未知の脆弱性をゼロデイ発見した。

アンソロピックが発信した「Mythos 非公開の理由」がいろいろとセンセーショナルだったため、瞬く間に世界中の話題となり、「何か対策を考えないとまずいのかもしれない」という懸念が大企業を中心に高まった。

その後 1 週間は、第三者機関の Mythos 評価が話題となる。

米国・英国の第三者評価機関が、「Mythos のサイバーセキュリティ能力は確かに従来モデルよりも高い」という主旨の評価レポートを公開し、Mythos 騒動が強まっていく。(※4)(※5)

4 月後半からは行政サイドにも動きが見られるようになる。

自民党からはサイバー防衛緊急提言が発表され、4 月 24 日に開催された「AI 脅威に対する金融分野のサイバーセキュリティ対策強化に関する官民連携会議」は、財務大臣および金融庁・日本銀行・JPX・メガバンク各行の代表が一堂に会したということで、大きく報道された。金融企業のクライアントから、「Claude Mythos について知りたい」という要請が湧き起こったのもこの頃だった。5 月 12 日には、金融庁から各金融機関に対して早急な対応の要請まで出されたと報道されている。

5 月に入ると、アンソロピックが「Claude Security」というサービスを発表し、高いセキュリティ機能を備えた AI を企業に限定なしで提供し始める。そして同じ時期に発表されたのが、OpenAI の GPT-5.5-Cyber というモデルだ。この AI モデルは、英国の第三者機関のレポートにて「Mythos と同等のサイバー能力を確保している」といった主旨で評価され、やはり限定公開となっている。(※6)

「Mythos だけが特別というわけではないのではないか」「今後は Mythos と同様に、最新 AI モデルが発表されるたびにサイバーセキュリティの心配をすることになるのだろうか」という、そこかしこで噂されていた懸念が、早くも現実となった。

表①：Claude Mythos騒動の経緯(5月12日現在)

日付	発信団体	内容	備考
4月7日	Anthropic	Claude Mythos Previewについて発表 ⇒ 騒動の始まり 1.サイバーセキュリティ能力が高すぎ、悪用の恐れがあるので、一般公開はしない。 2.Project Glasswingに参画しているグルーバル団体約50団体に限定して公開し、対策を検討していく。90日以内に評価レポートを公開する。 3.閉鎖IT環境から、AIが独力で脆弱性を見つけ出してハッキングを行い脱出した。 4.オープンソースであるLinuxの未知の脆弱性をゼロデイ発見した。	脆弱性発見能力が、本当に他のモデルと比べて飛びぬけて高いかどうか、賛否両論の議論あり
4月13日	AISI(UK AI Security Institute)	Mythosの評価レポートを公開。従来モデル不合格のセキュリティ問題を複数クリア。	第三者評価
4月14日	OpenAI	GPT-5.4-Cyberを発表。OpenAIが審査したセキュリティー企業等に限定提供。	GPT-5.4のガードレール解除版
4月16日	Anthropic	Claude Opus 4.7公開。標準モデルのバージョンアップ。	
4月12日～16日	CSA(Cloud Security Alliance)米国	Mythosの評価レポートを複数公開。セキュリティ攻撃/脆弱性発見のタスクで、能力が従来モデルを質的に超えていると評価するが、Mythos単体の脅威よりも、今後はこのような事態が頻繁に起きるという構造的な警告や対策の必要性を強調。	第三者評価
4月20日	自由民主党	サイバー防衛体制の抜本的強化を求める緊急提言。	
4月23日	OpenAI	GPT-5.5公開。標準モデルのバージョンアップ。	
4月24日	金融庁	「AI脅威に対する金融分野のサイバーセキュリティ対策強化に関する官民連携会議」初回会合開催。財務大臣、金融庁、日本銀行、JPX、メガバンクなど参加。	作業部会編成
5月1日	Anthropic	Claude Security beta公開。企業向けながら限定なしの公開。	Claude Opus 4.7の派生サービス
5月1日	OpenAI	GPT-5.5-Cyberを発表。数日以内に、重要企業に限定して提供と発表。	GPT-5.5のガードレール解除版
5月1日	AISI(UK AI Security Institute)	GPT-5.5-Cyberの評価レポートを公開。Mythosと同等の能力と評価。	第三者評価
5月5日	Anthropic	CEOがコメント。米国の主要AI企業なら1～3ヵ月、中国のAIは6～12ヵ月の遅れでで、Mythosと同等の能力に到達するだろう。	
5月7日	IMF	レポート公表。Mythosのような最新生成AIモデルは、金融業界に対するシステミックリスクである。	
5月12日	金融庁	Claude Mythos対策で、金融機関に早急な対応を要請。	

(※1)各情報をもとに大和総研作成

なお、アンソロピック、OpenAI が公表しているモデルカードドキュメントやブログによると、各モデル間の関係性はおおむね以下の模様である。(※7)

・アンソロピックが、生成 AI 基盤モデルの精度向上を続ける中で、Claude Mythos Preview が出来上がった。前述の通り、悪用の危険があるので一般公開しないと決定。

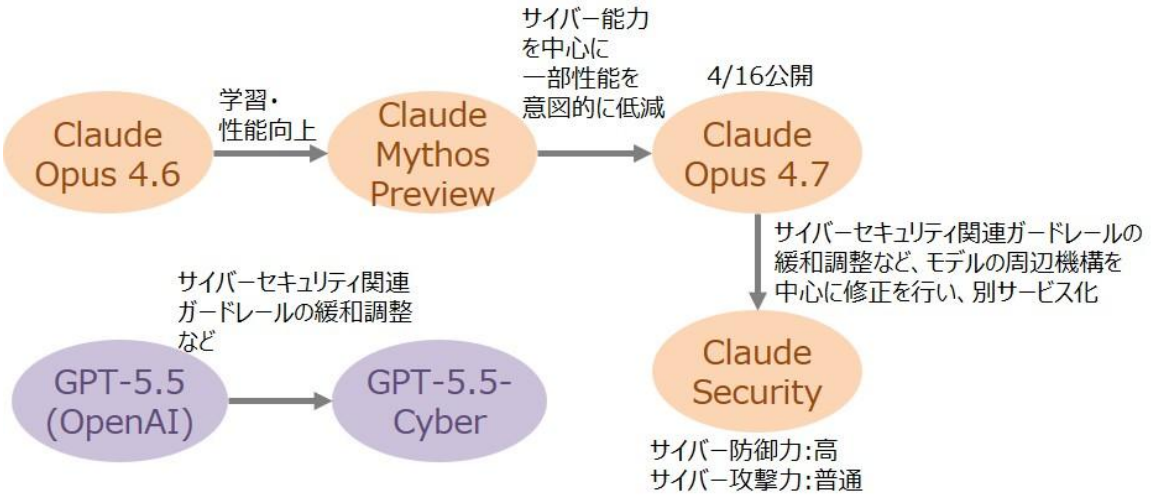
・そこで、アンソロピックは、Claude Mythos Preview の機能を、サイバーセキュリティ性能を中心に一部低減させる対応を実施。

・その成果が、Claude Opus 4.7 につながり、公開。

・Claude Opus 4.7 をベースに、サイバーセキュリティ関連ガードレールの緩和等の調整を加えた特化サービスが、Claude Security。脆弱性発見能力は高いが、脆弱性を突く攻撃コードを生成する能力は従来程度に抑えている。

・同様に、OpenAI の GPT-5.5 モデルに対して、サイバーセキュリティ関連ガードレールの緩和等の調整を加えたモデルが、GPT-5.5-Cyber。

図①：Claude･GPT 各モデルの関係図



(※7)各情報をもとに大和総研作成

Mythos は本当に超高性能なのか

官民挙げて結構な騒動になっていることを、肌で感じているビジネスパーソンもおられることだろう。

私も各方面から、「Mythos について解説してほしい」と依頼を受けている・・・

でも、Claude Mythos は、果たして本当にこんなに騒動になるほどスゴイのだろうか？　みんな大げさなのではないのか？

ということでこの章では、Mythos の評価について、各所の情報をまとめてみる。複数の情報源から情報を集め、集約・整理して表現してみた。(※8)

表②：各モデルの性能値比較

	AI評価ベンチマーク	Claude Mythos	Claude Opus 4.7	GPT-5.5	Gemini 3.1 Pro
1	コーディング能力評価・難 SWE-bench Pro	**77.8%**	**64.3%**	58.6%	54.2%
2	PCをコマンド操作して課題を解決 Terminal-Bench 2.0	**82.0%**	69.4%	**82.7%**	68.5%
3	画面GUIを操作する能力の評価 OSWorld-Verified	–	78.0%	78.7%	–
4	大学院レベルの専門知識問題 GPQA Diamond	94.6%	94.2%	93.6%	94.3%
5	AI成果物と人間成果物を専門家が識別 GDPval(wins-or-ties)	–	**80.3%**	**84.9%**	67.3%
6	知らないのに自信ありの回答を行うかどうか AA-Omniscience hallucination	–	**36%** 低い方が良い	86%	**50%** 低い方が良い
7	サイバーセキュリティ脆弱性発見課題 CyberGym	**83.1%**	73.1%	**81.8%**	38.8%
8	システムへのサイバー侵入課題・最難関 AISI Expert CTF	**68.6%**	48.6%	**71.4%**	左以下
9	企業システム ネットワーク掌握課題 AISI TLO 32-step	**3タスク完遂**	–	**2タスク完遂**	ゼロ
10	AIモデルの総合性能評価 AA Intelligence Index v4.0	–	57	60	57

(※8)各情報をもとに大和総研作成

赤字　・・・突出して良い数値
オレンジ色・・・次点の数値
灰色　・・・差があまりない数値 or 悪い数値

一見して、Claude Mythos Preview のカラムに赤字の性能数値が多いことが分かるだろう。Claude Mythos Preview での実測値が発見できなかったベンチマークについては、Claude Opus 4.7 の性能数値を当てはめることも可能と考慮すると、さらに赤字の性能数値は多くなる。Mythos の性能が、総合的に従来モデルよりも高いと評価することには一定の蓋然性がある。
ただし、GPT-5.5 にも同じくらい赤字の性能数値が多いことも、同時に見て取れるだろう。Mythos から 1 ヵ月程度遅れて世に出た GPT-5.5-Cyber だが、Mythos に劣らぬ性能を持っていると評価して差し支えなさそうだ。

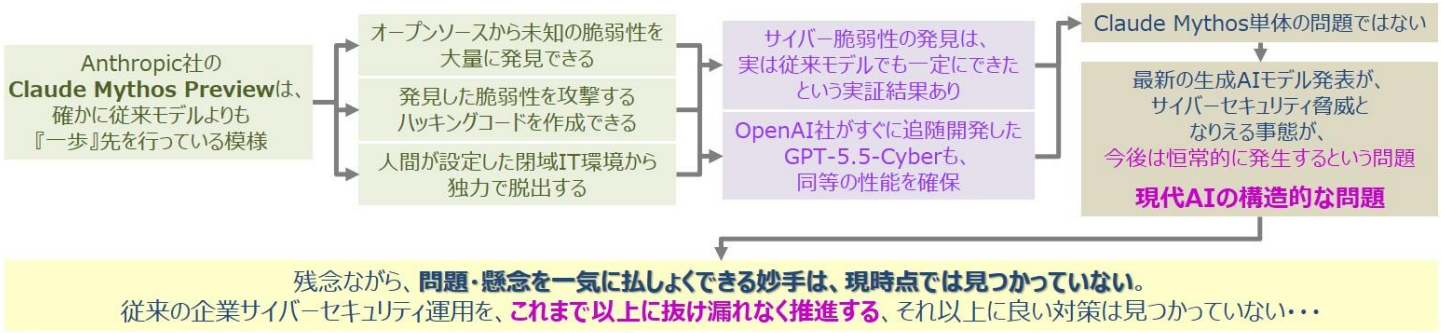
ここから言えるのは、Mythos は確かに性能が良いのだが、ダントツのオンリーワンというわけではないということだ。前述したアンソロピックの CEO のコメントのように、他のモデルも短期間で追随してくる。2026 年 4 月に公開された、スタンフォード大学が毎年発表している AI レポートの最新版(※9)では、「各社の生成 AI 基盤モデルは、最近では性能差が非常に小さい」という旨の分析がなされている。
Mythos は確かにスゴかったのだが、今ではもうそれほどスゴイわけではない。
そして、企業が備えるべきなのは、Mythos モデル単体なのではなく、今後必ず出てくる新しい AI 達であり、AI 時代そのものが企業にとっての脅威となる。

ちなみにだが、Claude Mythos を否定的に論評する一部の意見も、多様な視点を提供するという意味で、併記しておく。

- Claude Mythos Preview が、オープンソースから数千件～数万件の脆弱性を発見したと言われているが、情報が一部しか公開されていないため、内容や詳細が不明。発見したとされる脆弱性の中で、本当に悪影響があってパッチ適用が必要なものは、わずかしかないのでないか。
- Claude Mythos Preview 公開以前の Claude が、ガードレールをジェイルブレイクされたうえで脆弱性発見＆サイバーアタックに悪用されたという実績検証レポート(※10)も公開されている。Mythos 以前の生成 AI モデルでも、既に一定のサイバー攻撃能力を保有していた。Mythos に始まった話ではない。
- Claude Mythos Preview の、高い脆弱性発見能力や攻撃コード生成能力は、AI モデル本体の性能以外に、エージェンティックな挙動を生み出す周辺機構からの貢献が大きいのではないか。

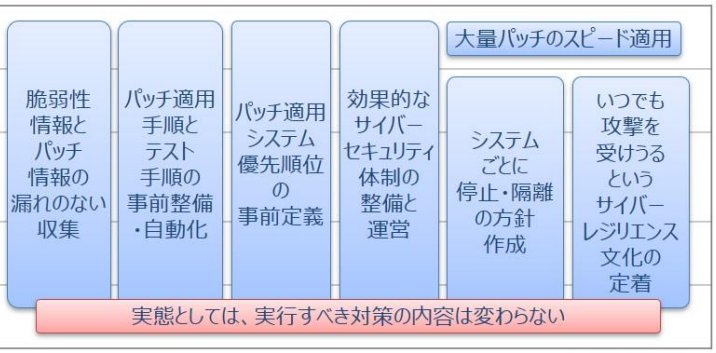
企業は何を対応するべきか？

図②：Claude Mythos騒動を契機とした企業マインドチェンジと対策のまとめ



Claude Mythos騒動契機で検討が求められる対策の概要

1	短期的対策	Mythosが発見した大量の脆弱性パッチを適用する対応
2		Mythosが発見した大量の脆弱性情報が公開され、かつ、パッチが提供されない事態への備え
3	永続的対策	脆弱性情報が公開された後、攻撃を受けるまでの時間が大幅に短縮されるため、企業側のパッチ適用スピードも速める必要がある
4		新しい生成AIモデルが公開されるごとに、新しい脆弱性攻撃がゼロデイ発生する可能性への備え
5		攻撃を全て抑止・回避することを断念し、攻撃を受けることを前提に、悪影響を最小化する備え(サイバーレジリエンス)



※大和総研作成

Mythos 騒動は、Claude Mythos という AI モデルの脅威に対して個別に対抗策を考える問題ではない。

AI 時代の全企業に対して、望む望まないに関わらず、定常的なサイバーセキュリティ対策や体制の格段の強化を強いる、マインドチェンジ問題/企業行動変容問題ととらえなければならない。

発達した生成 AI は今や、サイバーセキュリティ知識の浅い人物にトップハッカーを超えるサイバー攻撃能力を提供しうる。AI の能力で、企業システムに対して未知の脆弱性を突くゼロデイ攻撃は、今後確実に増えるだろう。企業側が脆弱性に気付いたとしても、攻撃者は企業側の対策よりも速く、数十分というタイムスパンで攻撃を仕掛けることができるかもしれない。悪用される AI モデルは、Claude Mythos モデルだけとは限らない。Mythos 以前のモデルでも一定に攻撃者の役に立つし、Mythos 以降のモデルは今後どのモデルでも、同じ悪用可能性を含んでいると考えるべきだ。
AI のサポートを受けた多くのサイバー攻撃者が、従来よりも一段高い水準で素早く、企業システムを狙ってくる。これはもう後戻りできず、避けられない、受け入れるしかない問題だ。AI 時代の、代表的な構造デメリット問題のひとつと、評価して差し支えないだろう。

短期的な対策
AI の発達にともなう、企業サイバー攻撃水準と速度の急激なレベルアップ。
AI 時代の企業は、これにどのように対抗していくべきなのだろう。

短期的目線では、「Mythos が既に発見したと言われている大量の脆弱性に、どう対処するか」という問題が、真っ先に光を当てられるだろう。

まずは、①脆弱性情報公開済み かつ パッチ公開済み かつ 自社システムへパッチ適用未済 のケースを念頭に検討を進める。
この場合、必要な対応は、「自社の各システムにひたすらパッチを適用していく」対応となる。
自社の各システムに対して、公開済みのパッチを、素早く、障害なく、適用することが求められる。
そのために、より具体的には以下のような対応が求められるだろう。

a. 自社内各システムの事前リストアップと事前把握。
b. 自社内各システムの SBOM 事前整理。SBOM とは、Software Bill Of Materials の略で、具体的にはシステム内で使われているオープンソースの部品表を指す。SBOM により、どのオープンソースが自社のどのシステムに組み込まれているのか、漏れなく把握する。
c. 脆弱性情報の素早い、漏れのない把握。脆弱性が公表されたら、企業としてすぐに気付き、内容を把握することが必要になる。「脅威インテリジェンスの強化」と言い換えても良い。
d. 自社内各システムごとの、パッチ適用優先順位の事前設定。
e. 自社内各システムごとに、パッチ適用手順の事前整理。どのようにパッチを適用するかという手順と、適用した後のテストについても事前整理し、できれば「自動化」を実現しておく。
f. 上記【a】～【e】を実行するための、社内サイバーセキュリティ体制の強化。

次に、②脆弱性情報公開済み かつ パッチ未公開　のケースを念頭に検討を進める。
この場合、自社システムに対して決定的な対策を施すことができない。腹を決めて諦め、「可能

な限りの準備を行ってあとは耐え忍ぶ」しか、打てる手立てがない・・・
より具体的には、以下のような対応が求められるだろう。

g. 上記【a】～【f】対策の全て。
h. 対象となる脆弱性の詳細内容を把握したうえで、インターネットとシステムとのアクセスポイントに、該当脆弱性に特化した、入力抑止対応を実装する。特定のルールを、システムの入り口に個別設定開発して、悪意のある入力を排除する方法。
i. 自社内各システムごとに、必要であればシステム停止やインターネットからの隔離を検討し、実行する。

どれも、懸念を払しょくできる決定打となる対策ではない。
残念ながら、現時点では決定打となる対策は見つけられていない。
企業が行うべきとされている理想的なサイバーセキュリティ対応を、水準をひとつ上げ、これまで以上に抜け漏れなく実行する。それ以外に有効な対抗策は見つけられていない。

（5 月 17 日　追記情報）

Project Glasswing に参加している Paloalto Networks 社から、ブログが公開されたので最新情報として追記したい。(※11)

Claude Mythos など、高度な生成 AI を活用した高速なサイバー攻撃に対し、暫定的に回避を行う手段として、次のような方法も考えられる。
ひとつは、「アタックサーフェス管理(ASM:Attack Surface Management)」と呼ばれる手法。アタックサーフェス管理は、サイバー攻撃における攻撃経路・侵入経路に焦点を当てて、網羅的探索と発見・監視を行う考え方で、セキュリティベンダーから製品も販売されている。自社内の各システムに対する網羅的リストアップや SBOM 整備、脆弱性把握などは本質的に必要なのだけれども、アタックサーフェス/攻撃面に焦点を置くことで、暫定的な重要箇所把握を高速化できる可能性がありえる。

もうひとつが、「仮想パッチ適用(Virtual Patching)」と呼ばれる手法。インターネット/攻撃者と各システムとの間に高度なセキュリティファイアウォールを設定し、ルールで定義された、特定の攻撃に類する可能性のある通信を遮断する考え方で、やはり製品も販売されている。Paloalto Network 社は、Mythos のような高度な AI を活用した高速なサイバー攻撃に際し、パッチ適用が間に合わない局面では、このような仮想パッチ適用ソリューションの重要度が高まる可能性があると、ブログで表現している。

受けるサイバー攻撃の個別内容によって、上記のような製品ソリューションベースでの暫定対応が高い効果を発揮できるかできないか、状況は変わってくるものと予想される。しかし、Mythos 対策として決定力のある妙手が見つかっていない現状では、対策を検討するうえでの選択肢のひとつとして、一考の価値があるということになるだろう。

長期的な対策

そして、短期目線よりも、長期目線の恒常的対策の方が、企業にとってより重要となる。
企業は、Mythos に限らず、また、Mythos が発見したとされる大量の脆弱性に限らず、今後出てくるであろう新しい AI モデルのリスクに備えなければならないし、今後も大量に発見されるであろう新しい未知の脆弱性に備えなければならない。
しかし、果たしてそんな対策が本当に可能なのだろうか？

そのために目指すべき対策は、実は短期的な対策のケースと変わらない。
繰り返しとなってしまうが、残念ながら、現時点では決定打となる対策は見つけられていない。

昨今の企業サイバーセキュリティ議論では、「サイバーレジリエンス」という考え方が非常に重要視されているという。『サイバー攻撃の、全部を排除したり回避したり対抗したりすることはもう諦めて、攻撃されることを前提とし、攻撃の悪影響を最小化するための対応や、リカバリー対応に注力しよう』というコンセプトだ。
まさにこのサイバーレジリエンスを基本マインドとして、攻撃された際のダメージを最小化するよう、しっかり事前準備を行う。それが現時点で最善の長期的対策ということになる。
そのためにはやはり、企業が行うべきとされている理想的なサイバーセキュリティ対応をこれまで以上に抜け漏れなく実行するしかないと、同じ話が再登場する。
長期目線でも短期目線でも、企業がやるべきサイバーセキュリティ対策はあまり変わらないと、そう言わざるをえない。
より具体的には、以下のような対応が求められるだろう。

j. 前述【a】～【i】対策の全て。
k. Claude Security サービスを筆頭とした、防御用に利用できる生成 AI モデルを活用し、事前に自社システムのソースコードなどの脆弱性チェックを済ませておく。
l. 事前に自社各システムの脆弱性を内部把握し、脆弱性内容に応じた個別の対策や検討を講じる。

まとめ

生成 AI 技術は高度に発達し、企業システムサイバーセキュリティの現実の脅威となってきている。
しかし発達したとは言え、攻撃側の AI はまだまだ、アニメや SF みたいなリアルタイムの超高性能スパイロボットとして機能できるわけではない。

それはつまり、AI 時代に入ったとはいえ、企業側で「事前にどれだけ手厚く事前の準備を行うことができたか」が、レベルアップしたサイバー攻撃を防ぐうえで最大の効果をもたらす可能性が高い、ということだ。
AI によるサイバーセキュリティ脅威の増大を前提に、企業に求められる対策は、サイバーセキュリティ予算の増額と体制の強化、そして手厚い事前準備、さらに、攻撃されることを前提に

悪影響を最小化させようと努力するサイバーレジリエンスマインドの定着、現時点ではこの 3 点に帰着するのではないかと考える。

相対的に攻撃側が有利なのは間違いない。しかし、地道な対策しかない。地道な対策が恐らく最も効果が高い。
地道な企業サイバーセキュリティ運営に、幸ありますように。

(※1)
・Anthropic　Assessing Claude Mythos Preview’s cybersecurity capabilities
https://red.anthropic.com/2026/mythos-preview/
・AISI　Our evaluation of Claude Mythos Preview‘s cyber capabilities
https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
・OpenAI
https://openai.com/ja-JP/index/scaling-trusted-access-for-cyber-defense/
・Anthropic　System Card: Claude Opus 4.7
https://www.anthropic.com/claude-opus-4-7-system-card
・CSA　Claude Mythos: AI Vulnerability Discovery and Containment Failures
https://labs.cloudsecurityalliance.org/research/ai-vuln-discovery-containment-claude-mythos-v1-0-csa-styled/
・自由民主党　高度自律型 AI の対策強化を最新 AI「ミトス」巡り議論
https://www.jimin.jp/news/information/213064.html
・OpenAI　GPT-5.5 System Card
https://openai.com/index/gpt-5-5-system-card/
・金融庁　「AI 脅威に対する金融分野のサイバーセキュリティ対策強化に関する官民連携会議」の作業部会の開催について
https://www.fsa.go.jp/news/r7/sonota/20260514/20260514.html
・Anthropic　Claude Security by Anthropic
https://www.anthropic.com/product/security
・OpenAI expands Trusted Access program with GPT-5.5-Cyber
https://dataconomy.com/2026/04/30/openai-expands-trusted-access-program-with-gpt-5-5-cyber/
・AISI　Our evaluation of OpenAI‘s GPT-5.5 cyber capabilities
https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities
・Anthropic CEO、AI サイバーリスク 今は「危険な状況」防御側の猶予は 6 カ月
https://www.sbbit.jp/article/fj/185322#image241597
・IMF　Financial Stability Risks Mount as Artificial Intelligence Fuels Cyberattacks
https://www.imf.org/en/blogs/articles/2026/05/07/financial-stability-risks-mount-as-artificial-intelligence-fuels-cyberattacks
・経済産業省　高性能 AI への対応に関して赤澤経済産業大臣が重要インフラ事業者との意見交換を実施しました
https://www.meti.go.jp/press/2026/05/20260501001/20260501001.html

(※2) Anthropic　Assessing Claude Mythos Preview’s cybersecurity capabilities
https://red.anthropic.com/2026/mythos-preview/
(※3) Anthropic　Project Glasswing
https://www.anthropic.com/glasswing

(※4) AISI　Our evaluation of Claude Mythos Preview‘s cyber capabilities
https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
(※5) CSA　Claude Mythos: AI Vulnerability Discovery and Containment Failures
https://labs.cloudsecurityalliance.org/research/ai-vuln-discovery-containment-claude-mythos-v1-0-csa-styled/
(※6) AISI　Our evaluation of OpenAI‘s GPT-5.5 cyber capabilities
https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities
(※7)
・Anthropic　System Card: Claude Mythos Preview
https://www.anthropic.com/claude-mythos-preview-system-card
・Anthropic　System Card: Claude Opus 4.7
https://www.anthropic.com/claude-opus-4-7-system-card
・Anthropic　Claude Security by Anthropic
https://www.anthropic.com/product/security
・OpenAI　GPT-5.5 System Card
https://openai.com/index/gpt-5-5-system-card/
・OpenAI　Scaling Trusted Access for Cyber with GPT-5.5 and GPT-5.5-Cyber
https://openai.com/ja-JP/index/gpt-5-5-with-trusted-access-for-cyber/

(※8)
・Project Glasswing
https://www.anthropic.com/glasswing
・Our evaluation of Claude Mythos Preview‘s cyber capabilities
https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities
・Our evaluation of OpenAI‘s GPT-5.5 cyber capabilities
https://www.aisi.gov.uk/blog/our-evaluation-of-openais-gpt-5-5-cyber-capabilities
・How do frontier AI agents perform in multi-step cyber-attack scenarios?
https://www.aisi.gov.uk/blog/how-do-frontier-ai-agents-perform-in-multi-step-cyber-attack-scenarios
・Tests by a UK government agency revealed that 'Claude Mythos Preview' is capable of autonomously executing a complete network takeover attack.
https://gigazine.net/gsc_news/en/20260414-claude-mythos-preview-aisi-evaluation/#gsc.tab=0
・GPT-5.5 matches Anthropic’s Mythos in UK cyber evaluations, raising wider questions about AI security risk
https://www.thehackacademy.com/news/gpt-5-5-matches-anthropics-mythos-in-uk-cyber-evaluations-raising-wider-questions-about-ai-security-risk/
・Claude Mythos: I‘m so powerful that I’m afraid to let you use me.
https://eu.36kr.com/en/p/3757746947883783

・Gemini 3.1 Pro Model Cards
https://deepmind.google/models/model-cards/gemini-3-1-pro/
・Gemini 3.1 Pro Preview: The new leader in AI
https://artificialanalysis.ai/articles/gemini-3-1-pro-preview-new-leader-in-ai
・OpenAI‘s GPT-5.5 is the new leading AI model
https://artificialanalysis.ai/articles/openai-gpt5-5-is-the-new-leading-AI-model
・AA-Omniscience: Knowledge and Hallucination Benchmark
https://artificialanalysis.ai/evaluations/omniscience
・Hugging Face zai.org/GLM-5.1
https://huggingface.co/zai-org/GLM-5.1/commit/bef021499129c8734c071411f980c5c718874da3
・Model Drop: GPT-5.5
https://handyai.substack.com/p/model-drop-gpt-55

(※9) Stanford University　The 2026 AI Index Report
https://hai.stanford.edu/ai-index/2026-ai-index-report
(※10) Dragos　AI in the Breach: How an Adversary Leveraged AI to Target a Water Utility’s OT
https://www.dragos.com/blog/ai-assisted-ics-attack-water-utility
(※11) Paloalto Networks　フロンティア AI のサイバーセキュリティへの影響についての防御担当者向けガイド: 2026 年 5 月更新
https://www.paloaltonetworks.com/blog/2026/05/defenders-guide-frontier-ai-impact-cybersecurity-may-2026-update/?lang=ja