新型 AI “Claude Mythos”の脅威とは何か

2026 年 5 月 27 日
7 月 24 日改定
㈱双日総合研究所
グローバル渉外部　加藤和平

- 米国新興 AI 企業の Anthropic は、自社の新型 AI モデル「Claude Mythos」について、**「サイバーセキュリティ能力が極めて高いため、一般ユーザーへの提供を見送る」**という異例の発表を行った。実際、各種ベンチマークにおいて、Mythos のハッキング能力は同業他社の生成 AI を大きく上回っている。
- その後、安全装置を講じた一般向けモデル「Fable 5」が公開されたものの、**米国政府による輸出管理指令により公開からわずか 3 日で全世界での提供停止**を強いられ、約 3 週間後に再開されるという混乱も生じた。
- Anthropic が提供見送りの判断を下せたのは、「Public Benefit Corporation（公益目的株式会社）」という、株主利益と社会公益の双方の追求を認められた特殊な会社形態のためである。
- 米国政府は当初、Anthropic と AI の利用制限を巡る対立により、同社を政府調達から排除した。ところが Mythos の登場後、自らは独占的な使用を求める一方、民間への拡散を制限する姿勢へと転じた。今回の一連の騒動は、国家が安全保障を理由に、一民間企業の高性能 AI サービスへ直接介入するという先例を残した。
- Mythos 級の高性能 AI が、もはや研究開発段階ではなく、現実に「使える」ところまで到達した以上、**同等の性能を持つ AI が誰の手にも渡る時代**は、遠からず到来する。
- 高性能 AI が悪意ある攻撃者の手に渡れば、サイバー攻撃を仕掛ける側のコストが劇的に下がる。これまで割に合わず放置されてきた無数の脆弱性が、一転して安価な攻撃対象へと変わりかねない。
- 企業の AI 投資は、**自社サービスを強化する「攻め」から、AI による攻撃から事業を守る「守り」へ**と、その比重を移さざるを得ない。サイバーセキュリティは、もはや「受動的な IT コスト」ではなく、「経営戦略」として位置づけるべきものとなる。
- あらゆる脆弱性を塞ぎ、完璧に守り切ることは構造的に不可能である。ゆえにこれから求められるのは、**守るべき資産を能動的に絞り込み、限られた資源をそこへ集中させる**という経営判断である。

はじめに

2026 年 4 月 7 日、米国新興 AI 企業の Anthropic は、自社の新型 AI モデル「Claude Mythos Preview（以下、Mythos）」について、**「サイバーセキュリティ能力が極めて高いため、一般ユーザーへの提供を見送る」**という異例の発表を行った。その後、6 月 9 日に安全装置を組み込んだ一般向けの後継モデルを投入したが、**その 3 日後に米国政府の輸出管理指令によって全世界での提供停止を強いられ**、約 3 週間後にようやく再開を許されるなど、更なる混乱も発生した。

これまで AI については、工学的アプローチから「AI をどうしたら賢く、より高い能力にできるか」という技術開発の追求と、社会科学・哲学的アプローチから「AI がシンギュラリティ（技術的特異点）[1]に達したとき、社会システムはどう変容するか」を思索する議論がそれぞれ展開されてきた。しかし、今回の Mythos 一般提供見送りや、公表停止という国家の介入は、これまでそれぞれの領域で展開されてきた「技術」と「社会」という二つの議論が交差した瞬間であると言えるのではないか。今回の Anthropic の高性能 AI と米国政府による介入は、**高度なテクノロジーの進化が人間のコントロールを超え、いよいよ現実の国家秩序や法秩序を揺るがす瞬間にまで到達する歴史的な転換点**となりえるのかもしれないのである。

本レポートは、この「一企業が作った AI モデルの公表見送り」と「国家による AI 開発企業のサービスへの介入」という事例を起点に、最新 AI の性能と、それが実社会やビジネスに与える影響について記載するものである。

第 1 章で生成 AI「Claude」の設計思想やモデルの製品群を整理し、第 2 章で一般提供が見送られた「Mythos」の突出した性能を、他の AI と比較しながら詳述する。第 3 章では、今回の提供見送りという特異な経営判断に至った Anthropic の企業形態について解説する。続く第 4 章では、この高性能 AI に対する米国政府の介入を時系列順に整理する。第 5 章では、今後の AI 投資の変化を、「攻め」と「守り」という観点から整理する。そして第 6 章では、企業経営としてこの局面にどう対応すべきかを考察する。

本レポートは 2026 年 5 月 27 日時点の情報で執筆したレポートを基礎としている。その後、2026 年 6 月 30 に双日総合研究所で執筆したレポート「Claude Mythos 狂騒曲：何が問題の本質なのか？」で扱った論点および 7 月 22 日時点の情報・報道を踏まえ、2026 年 7 月 24 日に一部内容の追記・改訂を行ったものとなっている。

なお、AI の性能と、それを取り巻く環境は現在も驚異的なスピードで変化している。数ヶ月後には情勢がさらに変化し、記載した情報が陳腐化している可能性がありうるところは予め申し添えたい。

目次

[1] 人工知能が自己改善を繰り返し、人間の知能を超える瞬間を指す概念

第 1 章：生成 AI「Claude」とその進化

1.1 Claude シリーズとその急成長

Claude は Anthropic が開発した大規模言語モデル（LLM）の対話型 AI サービスである。タスクの難易度や計算資源の消費量に合わせ、軽量かつ高速に処理を行うエントリー（初級）モデルの「Haiku（俳句）」、速度と高度な処理能力を両立させた実務向けの中級モデル「Sonnet（詩）」、複雑な論理推論やコード解析を行える最高峰モデル「Opus（作品）」の 3 つのモデルで構成されている。2026 年 6 月には更なる上位推論モデルで、Mythos に安全装置を組み込んだ「Fable（寓話）」が加わった。

さらに、これらの「問いかけに答える対話型 AI」から、指示に従って設定した目標に対して AI 自身がツールやプログラムを使いこなし、実行まで自律的に行う「自律型エージェント（Agentic AI）」のサービスも提供している。同社が 2026 年 4 月に正式リリースした「Claude Cowork」は、ユーザーのローカル環境と直接連携し、指示事項をマルチステップに分解して、AI 自身で自律的に目標を完遂する。ユーザーのパソコン内に常駐して主体的に実務をこなす「自律した同僚」を生み出す同サービスは、大きな人気を獲得している。

実際、Anthropic の売上は大幅に伸びている。非上場企業のため詳細な財務データは推計に基づくが、同社の成長は過去に類を見ないスピードとなっている。年換算ランレート（ARR）[2]は、2024 年末時点の約 10 億ドルだったのが、2025 年末には約 90 億ドルと 1 年で 9 倍になった。そして、2026 年第 1 四半期（1〜3 月）で ARR が約 300 億ドルを達成、5 月 28 日に発表された総額 650 億ドルの資金調達（企業価値評価額 9,650 億ドル）の時点で、ARR が約 470 億ドルに達する見込みであると発表されている。設立から数年の企業がこの規模で短期間に売上を伸ばした事例は IT 業界でも珍しく、同社のサービスへの市場の満足度と期待の高さが反映されている。

さらに、同社は 6 月 2 日、米証券取引委員会（SEC）へ上場申請書類を非公開で提出したと報じられており、10 月の NASDAQ 上場を目指しているとされている。

1.2 倫理観を支える「憲法 AI」

Claude は安全性と実用性の両立を重視している特徴がある。その信頼の源泉と言われているのが、同社独自の安全設計「Constitutional AI（以下、憲法 AI）」である。

従来の AI 開発では、不適切な表現や攻撃的な応答、不正行為等の悪意ある AI の利用に対して、人間がその都度不適切な出力であると判定して AI へフィードバックし、そのようなアウトプットを禁止するトレーニング手法（Reinforcement Learning from Human Feedback：RLHF）が主流だった。しかし、RLHF は人間の作業負荷が大きいだけでなく、個人の主観や倫理観によって「何を有害とするか」が異なるため、AI へのフィードバック基準が曖昧になるという課題を抱えていた。

これに対し、Anthropic は世界人権宣言等の過去の人類が築いてきた倫理規範を参考にした憲法 AI を明文化した。そして、AI の出力をその憲法に照らし合わせ、自律的に安全な表現へと修正させるトレーニングを行ったのである。このアプローチにより、Claude は人間の曖昧な主観や倫理観ではなく、**明文化された規範に基づいてアウトプットを安定**させられるようになり、悪意あるユーザーによる不正利用に対して極めて高い倫理観を維持できるに至ったのだ。

[2] ある時点の売上を年間売上に引き直した指標で、サブスク型のソフトウェアサービス売上を算定する際に使用される

そして同社は、憲法 AI を誰もが自由に閲覧・利用できるパブリックドメインとして全世界に公開している。Claude は性能の高さだけでなく、開発企業の公益を重視するスタンスからも支持者が多いのである。

1.3 製品ラインナップにおける Mythos の位置づけ

そして、この Claude の現行サービス最高性能である Opus に続く新たな先端 AI モデルが Mythos である。世界のサイバーセキュリティの現場に衝撃を与えた Mythos だが、本モデルはハッキング専用に開発されたものではなく、**Claude がこれまでローンチしてきた製品の延長線上にある最新モデルであるはず**だったのだ。

次章では、この Mythos が他の AI モデルと比べてどこが突出して高性能であったのか、その実態を詳述する。

第 2 章：Mythos の性能 -高すぎる能力ゆえの自主規制-

2.1 高すぎるハッキング性能

前章で述べたとおり、Mythos は既存 AI の延長線上にある発展的新モデルのはずだった。ところが、この Mythos は Anthropic が行った各種テストの結果、驚異的な性能を持っていることを、Anthropic 自らが明らかにしている。

Mythos において最も懸念されているのは、その**ハッキング能力の高さ**である。

Anthropic が Mythos の一般提供見送りとともに発表したレポート、「Assessing Claude Mythos Preview's cybersecurity capabilities（Claude Mythos Preview のサイバーセキュリティ機能の評価）」に記載されている検証結果によれば、Mythos は主要な OS や Web ブラウザを対象に、人間の専門家ですら見落としていた、**数千件規模のゼロデイ脆弱性[3]を自律的に発見する能力を持っている**ことが判明したのである。

さらに、単にバグや脆弱性を見つけるだけではなく、複数の小さな脆弱性を分析し、人間の指示や介在無しに、対象システムに対する最も適切な攻撃を行うコードを自動生成・実行・修正する能力を示した。

つまり、**コンピューターサイエンスの素人であっても、大まかな指示を Mythos へ入力することで、特定のシステムの脆弱性を瞬時に発見し、その脆弱性を突く最適な攻撃方法まで自動で提案・実行する、ということが現実に可能になるツール**なのである。

2.2 既存 AI との性能比較

実際、Mythos の能力は既存の AI サービスを凌駕している。

例えば、AI のコーディング性能を示すベンチマーク[4]である SWE-bench Verified[5]において、Mythos は 93.9%を記録した。これは、同じ Claude の Opus4.6（当時最新版）の 80.3%を大幅に上回り、レポート執筆の現時点で他のどの AI よりも高い。今まで、このベンチマークは新バージョンのたびに数%程度伸びてきたが、一気に 2 桁%伸びるという驚異的な性能向上を記録した。

また、サイバーセキュリティ性能を評価するベンチマークである CyberGym[6]で、Mythos は 83.1%を記録した。同ベンチマークの Opus4.6 は 66.6%にとどまっている。さらに、別のセキュ

[3] 開発者自身も把握していない未知の脆弱性
[4] 元々は PC 等のハードウェアの能力を測定・比較するための指標を指していた用語だが、現在では AI の推論や回答生成能力の比較のために使う指標もこう呼称している
[5] AI がどの程度ソフトウェアのバグを自律的に解決できる能力があるかを評価・測定するベンチマーク
[6] 複雑なサイバーセキュリティ攻撃やそれに対する防御のタスクをどの程度実行できるか評価・測定するためのベンチマーク

リティベンチマークである Firefox 147[7]では、Mythos が 181 件の exploit 生成[8]に対して Opus4.6 で 2 件の生成と、既存最新版と比べて 90 倍にも上る異次元の生成能力を発揮した。

さらに、AI 性能を測る最先端ベンチマークである Humanity's Last Exam（HLE）[9]では、Mythos は no tools[10]で 56.8%、with tools[11]で 64.7%を記録した。この指標では同業他社のモデルである OpenAI の ChatGPT 5.4 が no tools で 39.8%、with tools で 52.1%、Google の Gemini 3.1 Pro が no tools で 44.4%、with tools で 51.4%となっており、いずれの指標でも二桁ポイントと大幅に上回っている。

		Claude		OpenAI	Google
		Mythos	Opus4.6	ChatGPT 5.4	Gemini 3.1 Pro
SWE-bench Verified		93.9%	80.3%	-	-
CyberGym		83.1%	66.6%	-	-
Firefox 147		181 件	2 件	-	-
HLE	no tools	56.8%	-	39.8%	44.4%
	with tools	64.7%		52.1%	51.4%

（注）いずれも Mythos Preview が公表された 2026 年 4 月時点の数値であり、比較対象も当時の各社最新モデルである。

つまり、Mythos は**研究者でも解くのが困難な問題を、各種ツールやプログラミングを駆使しつつ解答できる高い能力を有している**のである。これらの指標を見れば、Mythos が既存システムの脆弱性を素早く発見し、各種ツールを組み合わせその脆弱性を攻撃するプログラムを生み出すことが容易にできうることは想像に難くない。

さらに、Mythos の性能は一般社会で運用されているシステムに対しても遺憾なく発揮できている。レポートでは、最も堅牢と評される「OpenBSD[12]」で **27 年間ずっと発見されてこなかった脆弱性を Mythos が短時間で検出した**、と報告されている。

ただし、これらの Mythos の性能の真実性に対しては、反論もなされていることは留意すべきである。例えば、世界中で広く使われる通信ソフト cURL の開発者は、自身の保有するコード群を Mythos で検証した範囲では従来のツールを超える発見はなく、そのサイバー攻撃能力は「（開発者の）誇大宣伝だ」と評している。

2.3 一般提供の見送りと「Project Glasswing」の発足

Anthropic は Mythos について、自身の公式サイトでレポート「Assessing Claude Mythos Preview's cybersecurity capabilities」とともに「Mythos の一般提供見送り」を発表した。

そして、この一般提供見送りと同時に、業界横断プロジェクトとして**「Project Glasswing（ガラスの翼プロジェクト）」をスタート**した。Project Glasswing は、Anthropic を中心に、Amazon Web Services, Apple, Microsoft, NVIDIA, Google, JPMorgan Chase といった米国を代表する約 50 の企業・団体により構成されている。本プロジェクトでは、 参加企業に「Mythos Preview」を共

[7] 実際の商用ソフトウェアを模したテスト環境に対して、実践的なサイバー攻撃を何回成功したかをテストするベンチマーク
[8] 数百回テストを行ったとき、何回攻撃を成功させられたか
[9] 2025 年に開発された生成 AI 専用のベンチマークで、大学修士～博士課程レベルの未解決に近い超難問の解読能力を比較するために使用される
[10] 外部との接続を断たれたオフライン環境で、AI モデルが持つパラメータ内の知識と推論のみで回答させる
[11] AI に「Web 検索」「プログラミングツール」「データ抽出ツール」などの外部ツールを使用・実行する権限を与えた状態で回答させる
[12] セキュリティ・ハッキング対策を重視して設計された堅牢度を重視した OS で、主に軍事・政府機関の防衛システムや、金融機関サーバーのシステムに採用されている

有し、ソフトウェアが公に出る前に Mythos にチェックさせ、その安全性を検証し脆弱性を共有することなどを目的としている。Anthropic は本取り組みに対し、**金額換算で最大 1 億ドル相当の Mythos Preview 利用クレジットを割り当て**ており、自らが発明した製品が社会一般へ危害を加えることを防ぐため、多大なコストを支払っている。

そして、参加組織は段階的に拡大され、6 月上旬までに 15 か国・200 の組織にまでアクセスが認められており、国際的な枠組みへと広がりつつある。

2.4 なぜ Anthropic は提供見送りができたのか

ここで 1 つの疑問が生じる。**Anthropic はなぜ 4 月に「最新モデルの一般提供見送り」という判断をできたのか**、である。

資本主義の原則に基づけば、より性能の高い製品を開発して市場に投入し、同業他社に先んじて利益を上げるのが営利企業の存在意義となる。実際、これまで Anthropic は、OpenAI や Google、xAI などの同業他社と競い、より便利な AI サービスを開発・投入してきた。Mythos の性能は同業他社のサービスを性能面で凌駕しており、市場に投入すれば大きなシェアを獲得できる革新的なサービスであったはずである。何より、Anthropic も AI 開発のために多額の出資を受けているにも関わらず、**「会社の成長よりも社会的害悪性を危惧し、製品の一般提供を見送る」と判断することは出資者への背信行為**となりかねない。

次章では、この判断をすることができた Anthropic という企業形態について解説する。

第 3 章：PBC -公益を追求する株式会社-

3.1 公益目的株式会社という新形態

Anthropic は、日本の会社法体系上まだ設立の枠組みがない**「Public Benefit Corporation（公益目的株式会社、以下 PBC）」**という特殊な法人格を採用している。

PBC とは、「**社会貢献や公益の達成を第一に考えながら、同時にビジネスとしての利益も追求する**」という思想に立脚した新しい企業形態である。

従前からある株式会社の目的とは、つまるところ「**株主利益の最大化**」であり、企業が行う活動は全て「株主利益のため」にならねばならない。このため、会社の事業が多大な利益を生むものの社会に大きな損害を出す可能性がある状況において、経営者が出資した株主への利益還元よりも社会貢献・社会公益を優先してしまうと、**株主から背任行為として訴訟を起こされる**リスクがある構造が存在した。

一方、社会的意義を重視して活動する NPO（非営利組織）には、「得られた利益を出資者や利害関係者へ分配してはいけない」という厳格なルールが定められている。そのため、莫大なインフラ投資や研究開発費が必要な事業のために民間からの巨額のエクイティを受け入れることが極めて難しくなり、短期間に事業を大きくすることが難しいという欠点があった。

これら「株主利益」と「社会公益」のジレンマを解決すべく誕生したのが PBC である。PBC は、**経営者が定款に定めた「公的利益の追求」と「出資者への利益分配」の双方を同時に追い求めることが法的に認められている**。また、社会安全性の追求といった公益を優先して短期的な利益創出を行わなかった場合でも、「取締役の義務を適正に果たしている」として、訴訟の場で経営者が裁判所から保護される立場となる。これにより、経営陣は企業・投資家からの大規模出資を受け入れつつも、出資者からの訴訟リスクを軽減でき、会社利益と公益の双方を両立しうる経営判断を下すことが可能となる。

3.2 PBC の法的位置づけ

米国では、企業や経営者の義務を規定する共通の「会社法」が存在せず、会社設立は各州で定める州法の管轄となる。

このため、PBC 設立の根拠や義務の定義は各州の法律によって規定されており、州によって PBC が設立できる地域と、そうでない地域に分かれている。例えば、法人設立に最も適しているデラウェア州では PBC の設立が可能である一方、ノースカロライナ州やミシシッピ州などではレポート執筆時点で同様の法律が導入されておらず、これらの州で PBC を起業することは不可能である。[13]

なお、この新しい企業の枠組みは、Anthropic を始めとする AI 開発企業のほか、環境配慮を理念に掲げるアウトドア衣料品大手の「パタゴニア」や、環境・水処理・廃棄物処理の世界最大手であるフランスのヴェオリアグループの米国法人「ヴェオリア・ノースアメリカ」などでも採用されており、持続可能な資本主義の新たな選択肢として広がりを見せている。

3.3 公益志向である AI 開発企業

Anthropic が通常の株式会社ではなく、PBC という形態を選択した理由は、この企業の「出自」にある。

Anthropic は、ChatGPT の開発企業である OpenAI 社から退職した社員らによって、2021 年に誕生した企業である。実は、OpenAI は元々、**「人類に安全な AI を届ける」ことを究極の目的にした非営利団体（NPO）としてスタート**した。しかし、AI 開発に必要な計算資源獲得に巨額投資が必要である一方、資金獲得に難航したことから出資を受けるために資金の受け皿となる子会社を設立。「非営利団体が営利子会社を傘下に収める」という歪な組織構造を取るようになった。また、AI 開発が巨額の収益を生む事業であることが自明になっていく中で、商業化を急ぐ一部経営陣と、当初の設立理念である「安全性」を重視すべきだと主張する研究者チームとの間で、路線対立が勃発した。

このとき、「安全性を軽視し、株主利益追求へ傾斜していく組織のあり方」に強い危機感を抱き、当時 OpenAI の開発チームに所属していた主要メンバーが離脱。このうち、ダリオ・アモディ、ダニエラ・アモディ兄妹を中心として、Anthropic を創業するに至ったのである。

次章では、Mythos の登場及び一般公開を巡って、米国政府の動向を中心にまとめる。

第 4 章：国家による AI 企業への介入

4.1 対立激しい米国政府と Anthropic

第 3 章で述べたとおり、Anthropic は高い倫理観を掲げ、実際の経営判断にまで反映させてきた企業である。だが、この高い倫理観は米国政府との深刻な対立を引き起こした。

遡ること 2026 年 2 月、米国防総省（DoD）と Anthropic の間で 2 億ドル規模の AI 利用契約交渉が進められていた。だが、自社 AI が「人間の恒常的な監視」や「自律型致死兵器（LAWS）」に転用されることを懸念した Anthropic が DoD へ利用制限を求めたため、広範な利用を求める DoD との交渉は決裂した。その後 DoD は、意にそぐわない Anthropic を **AI 企業では初となる「サプライチェーンリスク」認定**を行い、同社製品を政府調達から事実上排除、トランプ大統領も連動して全連邦機関に Anthropic 製品の利用停止を指示した。この判断は訴訟にまで発展し、現在も政府と Anthropic は係争中である。

[13] そのため、多くの PBC は、各種企業法務が最も整備されているデラウェア州法に基づき「デラウェア州法人」として登記するケースが見られる

4.2 米国政府の方針転換 -独占要求と拡大制限-

ところが Mythos の登場後、米国政府は「安全保障上の脅威」と断じた企業の製品に対し、今度は**「政府には優先的に使わせろ、ただし民間への拡散は制限する」**と要求し始めたのである。

2026 年 4 月 17 日の報道によれば、ホワイトハウス行政管理予算局の連邦最高情報責任者（CIO）が、各政府機関に Mythos を利用させるよう Anthropic へ要求していたことが判明している。さらに 4 月 30 日には、Mythos へのアクセスを認める組織数の拡大計画を保留するよう同社へ求めていたことも明らかになった。加えて、Anthropic 排除の震源地である国防総省で、5 月 12 日に Anthropic 製 AI を排除する移行作業と並行して **Mythos を政府全体のソフトウェアの脆弱性発見・修復のために導入している**ことが明らかになった。国防次官は、Anthropic の政府調達からの除外は継続するとしつつ、Mythos の扱いは「別個の問題だ」と言い切っている。

つまり、米国政府は「Anthropic は安全保障上の脅威なので政府機関での製品使用を禁止にする。だが、その会社が作った Mythos は、我々が使うし他社にはなるべく使わせない」というダブルスタンダード的立場を、公然と取ったのである。本来、Mythos は Claude のシリーズの 1 つに過ぎないのだが、「これは別物だから」という強引な論理で囲い込みを行ったことは、Mythos の持つ能力の高さと危険性が一定水準を超えた領域に達した証拠となっている。

4.3 一般公開 3 日後の全面停止措置

2026 年 6 月 9 日、Anthropic は Mythos Preview の後継モデルとして承認組織にのみ提供される「Mythos 5」と、高リスク領域への応答を遮断する安全装置を備え一般利用者へ提供される「Fable 5」を公表した。かねてより同社が表明していた「悪用を防ぐ安全装置を確立した時点で Mythos 級の能力を一般提供する」という方針が、ついに実行に移された。

しかし、この一般提供措置のわずか 3 日後の 6 月 12 日、米商務省のラトニック長官は Anthropic に対して**あらゆる外国籍者による Mythos 5 および Fable 5 へのアクセスを禁じる輸出管理指令を発令**した。Anthropic は利用者の国籍に応じてサービスを切り替える手段を持たないため、この命令を遵守するために全世界の全利用者に対して両モデルを停止せざるを得なかった。

そして、停止から約 2 週間後の 6 月 26 日に Mythos 5 が重要インフラ関連の米国組織へ提供、その後は停止の根拠となっていた輸出管理指令が解除されたことにより、7 月 1 日には Fable 5 の提供を再開した。現在は「従量課金制」として、一般ユーザーも使用可能となっている。

4.4 AI 利用への国家介入が意味するもの

輸出管理指令が 6 月末に解除されたことにより、Fable 5 と Mythos 5 をめぐる一連の騒動は、いったん収束となった。

だが、この騒動は「国家が安全保障を理由に、一民間企業の生み出した高性能 AI サービスへ直接介入し、その提供を止めた」という先例となった。このため、最先端 AI の利用が国家の判断にも左右される時代に入ったと言えよう。

なにより、Mythos 級の高性能 AI は、もはや研究開発段階ではなく、一般に『使える』までになった。国家が危険であると認識するレベルにまで、AI の性能が達したのである。今後、Anthropic の AI サービスに限らず、競合の OpenAI や中国製の AI も同等の性能に到達するのは時間の問題であり、高性能な AI が誰の手にも容易に渡る時代が到来する。この新たな環境において、企業はこの現実にどう向き合うべきなのか。次章から、その考察に入りたい。

第 5 章：AI 投資の転換 -「攻め」から「守り」へ-

5.1 AI による「攻撃コスト」の低下

Mythos を初めとする超高性能な AI が一般に出回る時代を迎えたことで、企業にとっての AI 及び自社ビジネスの IT 部門への向き合い方は根本から覆されようとしている。

従来、サイバー攻撃を伴うハッキングは仕掛ける側にとって決して割の良い行為ではなく、コスト・リスクの割に成功率が低い、いわば「油田開発」に近い性質のものだった。

ハッキングを通じて不当な利得を得るには、以下の 3 つの高いハードルを越える必要がある。

① **探索コスト**：攻撃すれば金銭的リターンが期待でき、かつ防御が手薄なシステム（＝掘りやすい大規模油田）を見つけ出す

② **開発コスト**：対象システムを綿密に分析し、独自の脆弱性に対する有効な攻撃プログラム（＝採掘インフラ）を多大な時間とリソースをかけて構築する

③ **摘発リスク**：相手の防御策によって失敗に終わる可能性もある（＝本当に油が出る保証はない）うえに、常に逮捕・摘発（＝政府による開発中止）という社会的高リスクがつきまとう

これらが障壁となって攻撃に見合う合理性が見込めないことから、割に合わないがゆえにハッキングの対象とならずに済んだ、という事例も存在した。

ところが、超高性能 AI の登場により、この構図は一変する。AI の高い分析力及びコーディング力が①探索コストと②開発コストを劇的に押し下げることで、**サイバー攻撃への ROI（投資対効果）が劇的に向上**するのである。

これは、かつては採算の合わなかった油田が、AI という新たな採掘技術によって、採算ラインへと転換することを意味する。つまり、これまで**攻撃側のコストの高さゆえに放置しても実質的な実害が出にくかった無数のバグや脆弱性が、すべて有効かつ安価な攻撃ルートへと変貌**しかねないのである。

5.2 「守り」の AI 投資とその限界

攻撃側がこれほどまで有利になる場合、企業の AI 投資への向き合い方も変えざるを得ない。

これまで、企業の AI 投資の主眼は「攻め」にあった。自社のサービスにどう AI を組み込み、既存事業の強みと掛け合わせて付加価値を高めるかが求められてきたのである。しかし、Mythos 級の AI が登場したことで、**今後の AI 投資には「守り」の視点が否応なく求められる**。AI による攻撃からいかに事業を守るかという「守り」の発想が、これからの経営に欠かせないものとなっていくのである。

もっとも、この「守り」の投資は、「攻め」の投資とは異なり、際限がないという難しさを抱えている。「攻め」であれば費用対効果に基づく段階的・合理的な投資判断が可能だが、サイバー防衛などの「守り」はリスクの見積もり次第で必要額がどんどん膨らんでいく。また、防御への投資自体が直接的な利益を生むわけではなく、顧客に対する付加価値の創出にも繋がりにくい。それでも、AI による攻撃が現実の脅威となる以上、これを経営上の必要不可欠なコストとして引き受けざるを得ない時代が、すでに到来しつつあるのである。

5.3 完璧な守りが存在しえない中で求められる経営判断

また、多大なコストをかけて「守り」を固めようにも、あらゆる脆弱性を塞ぎ、システムを完璧に守り切ることは、現実的には難しい。

まず、現場で運用されているシステム・ソフトウェアにおいて、**バグが無い製品は存在しない**、と断言してよい。バグ修正にもコストがかかる以上、そのバグがシステムに致命的な影響を与えた

り、実害に直結したりしない限り、ベンダーがコストや安定運用を犠牲にしてまでそれを全て修正することはないのが IT 業界の現実である。いかなるシステムも、修正されないまま放置された弱点を、構造的に抱え込み続けている。だが、**攻撃側のコストが劇的に下がるとなると、これらの弱点はもはや放置したままではいられなくなる**のだ。

では、攻撃側と同じく、防御側も AI で武装すれば守り切れるのか。確かに、高性能 AI による脆弱性発見の自動化は、防御策としてすでに取り組みが行われている。第 4 章で紹介した国防総省の Mythos を活用した取り組みも、その一例である。

だが、**発見された脆弱性を確認し修正する人間側の作業が追い付かず、発見数と修正完了数の差は開く一方**となっている。実際、Mythos による探索により多くの脆弱性候補が大量に報告されるようになったものの、それらを人手で確認し、修正作業を行うことにまで手が回らず、脆弱性のみが積みあがっていくという状況はすでに起こっている。AI により IT を巡る環境が変化する中、この問題を企業の IT 部門のみで対処することは不可能となってきている。

そのため、これから求められるのは、**守る対象を能動的に絞り込み、限られた資源をそこへ集中させるという経営判断**となる。次章では、そのために経営判断として何をすべきか、より実務的な視点から整理したい。

第 6 章：AI 時代に何をどう守り抜くか

6.1 基本的対策の徹底

高度な最先端 AI が攻撃に用いられる時代を前にすると、我々はつい、同様の最新システムを活用した防御策の導入検討に思考が寄ってしまう。しかし、システム攻撃の足がかりの多くは、依然として**初歩的な対策の不備**を突いたものである。

このため、**基本的な対策の徹底は未だに有効**である。多要素認証（複数の方法で本人確認を行う仕組み）、適切なアクセス権限の管理、ログ（操作や通信の記録）の取得と監視といった基礎的な保護策を確実に運用する。高度な AI が攻撃に使われる環境においても、侵入の足がかりの多くは依然として基本の不徹底にある。**新しい製品やソリューションを次々に追加するよりも、すでに持っているものを限界まで使いこなすことが先決**である。

6.2 自社の IT 環境の把握

次に取り組むべきは、**自社のシステム・IT 環境の正確な把握**である。自社がどのようなシステムとデータを、どこに、どれだけの価値で、どのような形で抱えているのかの把握無くして、守るべき資産の精査は不可能である。

また、この把握作業は、継続的に行う必要がある。情報資産はクラウドや SaaS（インターネット経由で利用するソフトウェア）、国内外の拠点、子会社や買収先、現場が個別に導入したシステムへと散在し、把握した端から増減していく。長年改修を重ねたシステムの内部は、もはや誰も全体像を持っておらず、何がどこに繋がり、どのデータがどこを流れるのかは、実地で調べ直さなければ分からない。このため、**把握のために実態を継続的に調べ続ける組織体制を整える**必要がある。

6.3 守るべき情報資産の再定義

そのうえで、把握した情報資産のうち、自社の事業にとってそこまで影響しないと判断しうるものに対して、**重要度を「下げる」という経営判断**が必要となる。

2005 年に個人情報保護法が全面施行されて以降、情報資産に対しては、「守る」「残す」「漏らさない」という対応がほぼ一貫してきた。関連法制の改正のたびに対象範囲の拡大と管理体制の強化が繰り返される一方で、「この情報資産は守るに値しない」と保護の対象から外す判断が、組織的に促されることはほとんどなかった。結果として、守るべき情報資産は積み上がるが、捨てる判断には責任が伴うために誰もやりたがらず、加えて、本質的に重要でない情報資産に対する管理コストを看過・残置することの問題は表出しにくいことも手伝って、最終的に「残す」という選択がなされ続けてきた。これにより、情報資産は単調に増え続け、すでに多くのセキュリティ実務の現場はこの増加する対象を抱えきれずに飽和している。

ここへ高性能 AI による脆弱性の早期発見や安価な攻撃能力が現実化すれば、膨大な管理対象を現場が抱えきれず、いずれ立ち行かなくなるのは時間の問題である。だからこそ、経営判断として実務の負担を減らし、管理コストを適正なレベルに維持すべく、対象を絞るということが求められる。

6.4 重要情報資産へのリソース集中

そして、**絶対的に守るべき資産を選んだら、リソースを集中させ、それに対する徹底的に防御策を講じる**。AI を用いることでハッカーの攻撃能力が向上しうる以上、絶対に守らなければならない資産には、これまで以上に手厚いケアが求められる。

また、「守る対象から明示的に外す」と決めた資産へも丁寧な対応を行う。事業継続に影響しないことを事前に確認し、バックアップ（複製の保存）とリカバリ訓練（復旧手順の演習）を繰り返して万一の破損や漏洩が起きても業務が止まらない状態を保ち、法令上の報告義務は当然に果たす。仮にその資産が攻撃を受けても「これは組織として守る対象から外すと事前に合意した資産であり、事業継続に影響しない」と説明できる共通認識を平時のうちに形成しておくこと、すなわち、**守らないという判断を、攻撃を受けてから慌てて正当化するのではなく、受ける前に組織の合意として固めておく意思決定**が必要となる。

おわりに

本レポートでは、Claude Mythos の公開を巡る一連の騒動に端を発し、最終的には今後の企業の経営戦略における IT の位置づけについて論じた。

こうして見てくると、**AI 時代の「守り」の経営とは、情報システム部門内のみの技術論に留まらない**ことが分かる。Mythos 級の攻撃能力はやがて競合や悪意ある攻撃者にも普及し、その結果として脆弱性発見が自動化・安価化するだろう。あらゆる軽微な脆弱性が突かれうる時代にあっては、**完璧を求める前提そのものを見直し、情報漏洩やセキュリティインシデントに対する対処法を、新たな脅威の現実に見合ったものへ調整していく**必要がある。

米国政府による介入事例のとおり、最先端 AI へのアクセス可否はもはや我々の制御が及ばない領域へと進んでしまった。外から与えられる能力を当てにできない以上、自前の優先順位付けと基礎の徹底こそが、これからの時代においてもっとも確実な備えとなるのである。

参考文献・web サイト

【危険すぎて Anthropic が封印「Claude Mythos」爆誕】今井翔太「人類は一線を超えた」／“エイプリルフール”を疑う高性能／メタは新型モデルで AI 競争に本格参戦【AI QUEST】
Project Glasswing: How Claude Mythos Finds Zero-Day Vulnerabilities Autonomously (2026) | NxCode
Claude Mythos Preview 入門 — SWE-bench 93.9%・Project Glasswing の全貌 #AI - Qiita
Claude Mythos Preview: Benchmarks, Pricing & Project Glasswing
Anthropic 30B 調達｜評価額 57 兆円の全貌 ｜ 株式会社 Uravation
Claude Mythos 狂騒曲：何が問題の本質なのか？（総研レポート）

Anthropic 公式発表
Assessing Claude Mythos Preview’s cybersecurity capabilities
Project Glasswing: Securing critical software for the AI era
Claude’s Constitution AI Anthropic
Statement on the US government directive to suspend access to Fable 5 and Mythos 5 Anthropic
Redeploying Claude Fable 5 Anthropic

引用した報道
アンソロピック、米ＩＰＯを非公開で申請　オープンＡＩに先行 | ロイター
Anthropic says it hit a $30 billion revenue run rate after 'crazy' 80x growth | VentureBeat
Anthropic set to hit $10.9 billion in revenue in Q2, source says
アンソロピック、ミュトス利用を日本など 15 カ国 200 社に拡大 - 日本経済新聞
White House to give US agencies Anthropic Mythos access, Bloomberg News reports | Reuters
White House Opposes Anthropic’s Plan to Expand Access to Mythos Model - WSJ
米政権、アンソロピック「ミトス」を政府機関に提供へ＝報道 | ロイター
米国防総省、アンソロピックの出禁継続も「Mythos は別問題」 - 日本経済新聞
アンソロピックのミュトス級ＡＩ、米が外国人アクセス停止命令　安保懸念 | ロイター
アンソロピックのＡＩ「フェイブル」「ミュトス」、米が輸出規制解除 | ロイター