2026 年 5 月 27 日　全 7 頁

最先端 AI はなぜ一般提供されないのか

Claude Mythos が示す AI モデルの戦略資産化と提供構造の変化

経済調査部　AI アナリティックリサーチ室　主任研究員　　田邉 美穂

[要約]

■ 2026 年 4 月に米 Anthropic が発表した生成 AI モデル「Claude Mythos Preview」は、サイバー領域において高い能力を持つとされ、現在は一部の機関や企業に限定して提供されている。こうした限定提供の判断は、既存の制度や対応能力では AI の社会的影響を十分に制御しきれない段階に近づいていることを示す先行事例といえる。

■ こうした動きは、AI モデルの提供形態が、従来のように広く一般に提供される AI と、利用主体や用途が制約される AI の二層に分化していく可能性を示している。AI モデルの開発企業にとっては、高リスク領域における合理的なリスク管理手段となる一方で、利用者側にとっては、AI モデルへのアクセス権の有無に応じて、対応能力や競争力の格差が生じる可能性がある。

■ もっとも、このような提供形態が今後も維持されるとは限らない。今後の展開は、「フロンティアモデルと同水準の AI モデルの普及時期」と「アクセス統治の強度」といった二つの軸に左右されると考えられる。これらの軸に基づくと、AI モデルの提供構造は、①管理型 AI の拡大（統治が有効に機能するケース）、②管理主体の分散（不安定な二層構造が生じるケース）、③統治の形骸化（事実上の一般化が進むケース）といった三つのシナリオに整理される。

■ 現在は、上記のシナリオ①（管理型 AI の拡大）に相当する状況にあるが、同水準の AI モデルの普及やアクセス統治の変化により、他のシナリオへ移行する可能性がある。とりわけシナリオ③（統治の形骸化）に進んだ場合、悪用リスクへの対策を講じる時間的猶予がなくなる。このため、企業には、被害発生を前提とした迅速な対応体制の整備、政府には AI モデルへのアクセス確保に向けた外交的対応やソブリン AI の検討が求められる。

株式会社大和総研 丸の内オフィス　〒100-6756 東京都千代田区丸の内一丁目 9 番 1 号 グラントウキョウ ノースタワー

1. はじめに

近年、生成AIの高度化に伴い、その能力や利活用のあり方をめぐる議論が一層活発化している。こうした中、2026年4月に米Anthropicが発表した生成AIモデル「Claude Mythos Preview」（以下、Mythos）は、ソフトウェアの脆弱性発見やサイバー領域における高い能力を持つとされ、話題となった。サイバー攻撃にAIモデルが悪用された場合のリスクを踏まえ、現在は一部の機関や企業に限定して提供されている。こうした「一般提供されない」という扱いは、単なる性能向上の話にとどまらず、AIの提供構造そのものが変化しつつあることを示唆している。

Mythosをめぐる議論においては、「高性能なAIが登場した」という技術的な進展だけでなく、最先端のAIモデルの提供対象が限定されている点にも着目する必要がある。すなわち、AIの価値はその性能のみならず、「誰が利用できるか」という点にも左右される段階に入りつつある。本レポートでは、Mythosの事例を手掛かりに、最先端AIが一般提供されないことが示唆する構造変化について整理し、AIが「広く普及する技術」から「管理される戦略資産」へと位置づけを変えつつある可能性について考察する。

2. 「一般提供されないAI」が示すAIの戦略資産化

悪用された場合の社会影響の大きさを考慮し、一般提供を断念

Mythosが一般提供されなかった理由として、主にサイバーセキュリティ分野での高い能力への懸念が指摘されている[1]。具体的には、ソフトウェアの脆弱性発見や攻撃コード生成（エクスプロイト）といった能力が従来モデルを大きく上回り、悪用された場合には既存のサイバー防衛の前提を大きく揺るがす可能性がある。このため開発元の米Anthropicは、同モデルを広く一般に提供することはせず、政府機関や企業など限定された主体に対し、防御目的での利用に限って提供すると決定した。悪用しようとする勢力に対して時間的優位を確保し、実際の被害発生前に防御策を講じることを可能とする狙いがある。このような枠組みとして、セキュリティ対策を目的としたコンソーシアム（企業連合）「Project Glasswing」が組成されている[2]。

ただし、この脆弱性発見や攻撃コード生成の能力は、Mythos固有のものとは言い難い。大規模言語モデルは、コード生成や論理推論能力の向上に伴い、ソフトウェアの脆弱性発見や攻撃手法の構築といったタスクに一定程度対応可能となりつつある。Mythosはこのような能力がよ

---

[1] Anthropic Frontier Red Team blog “Assessing Claude Mythos Preview’s cybersecurity capabilities” April 7, 2026.では、Mythosを用いて、ゼロデイと呼ばれる未知の脆弱性が数千件発見されたこと、その具体例としてLinuxカーネル(世界のほとんどのサーバーを動かすソフトウェア)内の複数の脆弱性を自律的に発見したこと等が報告されている。

[2] 米Anthropicウェブサイト “Project Glasswing: Securing critical software for the AI era” （2026年5月14日最終閲覧）によると、米Anthropic、米Amazon Web Services（AWS）、米Appleなど、重要なソフトウェアインフラを構築し維持する40以上の組織にMythosを提供し、その結果明らかになったことを業界全体に共有していくとしている。また、日本経済新聞 電子版「3メガバンク、AIミュトスのアクセス権入手へ　サイバー防衛で日米連携」（2026年5月13日）によると、日本企業で初めて、三菱UFJ銀行、三井住友銀行、みずほ銀行のメガバンク3社に対し、アクセス権が付与される見通しである。

り高い水準で発揮された例と位置づけられる。実際に、Mythos の発表後、米 OpenAI は GPT-5.5 の発表に続き、サイバーセキュリティ関連のガードレール（AI の出力内容や挙動を制御し、危険な回答を抑制する仕組み）を緩和した「GPT-5.5-Cyber」の提供を認証済みの防御側ユーザーに限って開始した[3]。また、現状の中国における AI 開発状況から、半年から 1 年程度の時間差でこれらの AI モデルと同水準の能力に到達するとの見方もある[4]。すなわち、Mythos と同様の能力は今後、他の AI モデルにも広がっていくと想定される。

この点を踏まえると、Mythos は社会的影響の観点から特に慎重な扱いが必要とされたこと自体は妥当であろう。より重要なのは、一定水準以上の能力を持つ AI モデルが一般に提供された場合、個別モデルの特性を超えて社会的影響が拡大し得ることを、Mythos の事例が示唆した点にある。言い換えれば、AI の能力が既存の制度や対応能力ではその社会的影響を十分に制御し得ない段階に近づいていることが、今回の出来事の本質である。

このように、最先端 AI をめぐる提供のあり方は、特定モデルの特殊事情というよりも、AI 技術の進展に伴い、提供範囲と制御の関係そのものが変化しつつあることを示唆している。

AI モデルの提供のあり方の変化 – 広く一般提供される AI とアクセスが管理される AI

従来の AI モデルは、ウェブサービスやスマートフォンアプリ、API[5]などを通じて広く利用されることを前提としてきた。企業や個人による多様な活用を通じてユースケースが拡大し、AI モデルは、普及そのものが価値を生む構造である。すなわち、一般用途向けの AI は、「広く利用されること」が市場における競争力の源泉となるといえる。

一方で、Mythos に代表される最先端 AI モデルでは、この前提は成立しにくい。サイバーセキュリティ分野など高いリスクを伴う領域では、アクセスを絞ること自体がリスク管理手段であると同時に、安全性や信頼性に裏付けられたモデルの価値を維持するための条件にもなる。そのため、政府機関や特定の企業など信頼関係にある主体を中心に利用が広がり、その範囲は段階的に拡大していくと想定される。

このようなモデルは、安全保障や重要インフラにおけるサイバー防御の高度化に資するほか、金融分野における不正検知やリスク分析などへの応用も考えられる。こうした点を踏まえると、AI モデルは「広く普及する技術」から「安全保障や社会基盤に関わる重要な機能を担う戦略的な資源」へと性格を変えつつあるといえる。

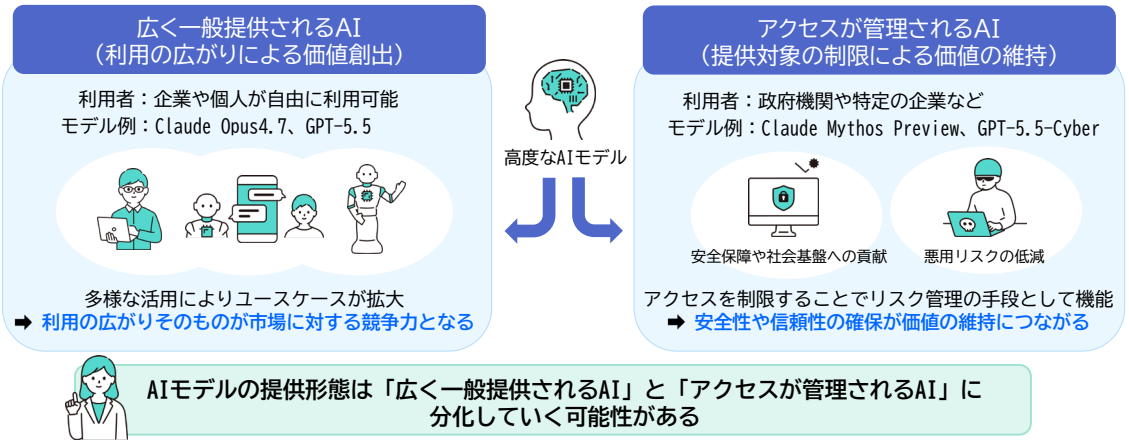
AI は今後、一様に普及する技術というよりも、リスク特性や用途に応じて異なる形で提供される可能性が高い。すなわち、従来型の利用の拡大を通じて価値を生む「広く一般提供される

---

[3] OpenAI “Scaling Trusted Access for Cyber with GPT-5.5 and GPT-5.5-Cyber: How our latest models help each layer of the defensive ecosystem and accelerate the security flywheel.”（2026 年 5 月 7 日）
[4] 日本経済新聞 電子版「アンソロピック CEO『中国の AI、6～12 カ月で Mythos に追いつく』」（2026 年 5 月 6 日）
[5] API とはアプリケーション・プログラミング・インターフェース(Application Programming Interface)のことで、異なるソフトウェアや ウェブサービスなどをつなぐためのインターフェースを指す。

AI」と、利用主体や用途が一定程度制限される「アクセスが管理される AI」に提供形態が分化していくことが考えられる（**図表 1**）。このような変化は、従来の AI モデル開発における性能向上をめぐる競争だけでなく、必要な水準の AI モデルへのアクセスを確保できるかどうかという次元にも広がりつつあることを示唆している。

図表 1 AI モデルの提供形態の分化



（出所）各種資料より大和総研作成（イラストはソコスト https://soco-st.com/）

AI モデルの提供制御がもたらす利点と課題

社会基盤や安全保障に影響を与える AI モデルの提供が制御される構造は、関係する主体ごとに異なる利点と課題をもたらす。その影響は、①AI モデルの開発企業、②AI モデルへのアクセス権を持つ主体、③AI モデルへのアクセス権を持たない主体という三つの観点から整理できる（**図表 2**）。

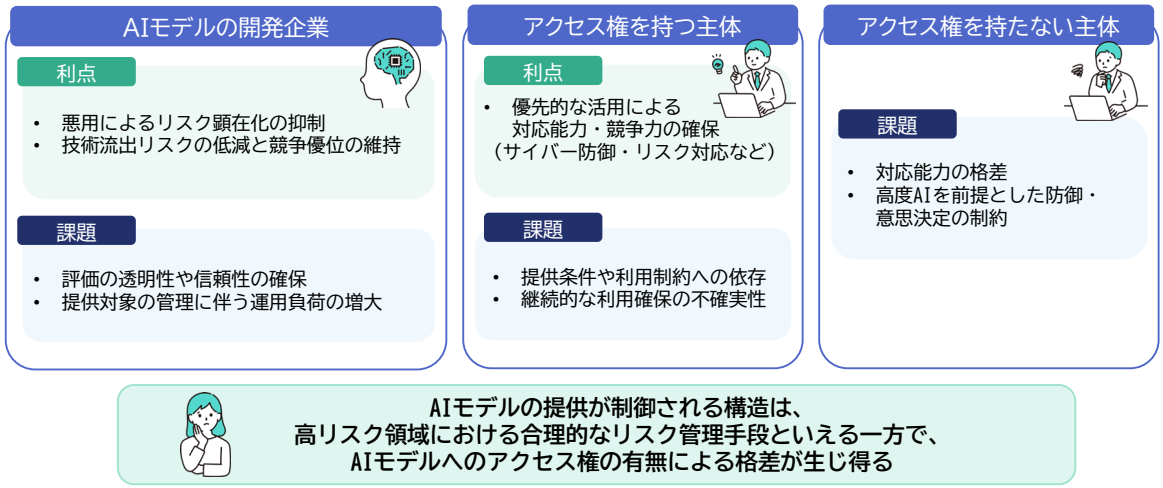
まず、**①AI モデルの開発企業**においては、サイバー攻撃など、社会に重大な影響を及ぼし得る領域での悪用を抑制できることが大きな利点である。提供対象を限定することは、悪用による問題が顕在化した場合の規制強化や企業イメージの毀損を抑制する効果を持つ。また、アクセスを制御することで、蒸留（高性能モデルの出力を教師データとして別のモデルを訓練する手法）などによる技術流出のリスクを一定程度低減でき、競争優位性の確保にも資する。これらは、利用環境や用途を絞ることで制御可能性の確保につながるほか、需要の急増に伴い計算資源が逼迫する現状においては、限られた資源の範囲内で安定的なサービス提供を維持できるという実務上の利点もある。一方で、提供の制限は外部からの検証機会を減少させ、評価の透明性や信頼性の確保に課題を生じさせる可能性がある。加えて、提供対象の選定や利用状況の管理が新たに必要になるなど、運用負荷の増大も課題となる。

次に、**②AI モデルへのアクセス権を持つ主体**においては、必要な水準の AI モデルを優先的に活用できることで、サイバー防御やリスク対応といった重要領域において、他の主体に対する優位性を確保できる点が利点となる。他方で、提供条件や利用範囲の制約を受ける立場にあり、利用の自由度や継続性を確保できるとは限らない点が課題となる。

最後に、**③AI モデルへのアクセス権を持たない主体**においては、構造的な不利が生じる可能性が高い。サイバー防御やリスク対応といった重要領域において、高度な AI モデルを前提とした防御や意思決定が行えず、アクセスを持つ主体との格差が拡大し得る。

以上を踏まえると、AI モデルの提供が制御される構造は、高リスク領域における合理的なリスク管理手段といえる一方で、AI モデルへのアクセス権の有無が、サイバー防御能力や意思決定の質に格差をもたらし得る点で、今後の重要な論点となると考えられる。

図表 2 関係主体別で見る AI モデルの提供制御の利点と課題



（出所）各種資料より大和総研作成（イラストはソコスト https://soco-st.com/）

3. AI モデルの提供構造はどのように変化するのか

AI モデルの提供構造を左右する要因と今後のシナリオ

AI モデルの提供対象を制限する現在の方針が、長期にわたって維持されるかは必ずしも明らかではない。今後の展開は、同水準の AI モデルがどの程度の速度で普及するか、そしてアクセスがどの程度の強度で統治されるかという、主に二つの軸によって左右されると考えられる。

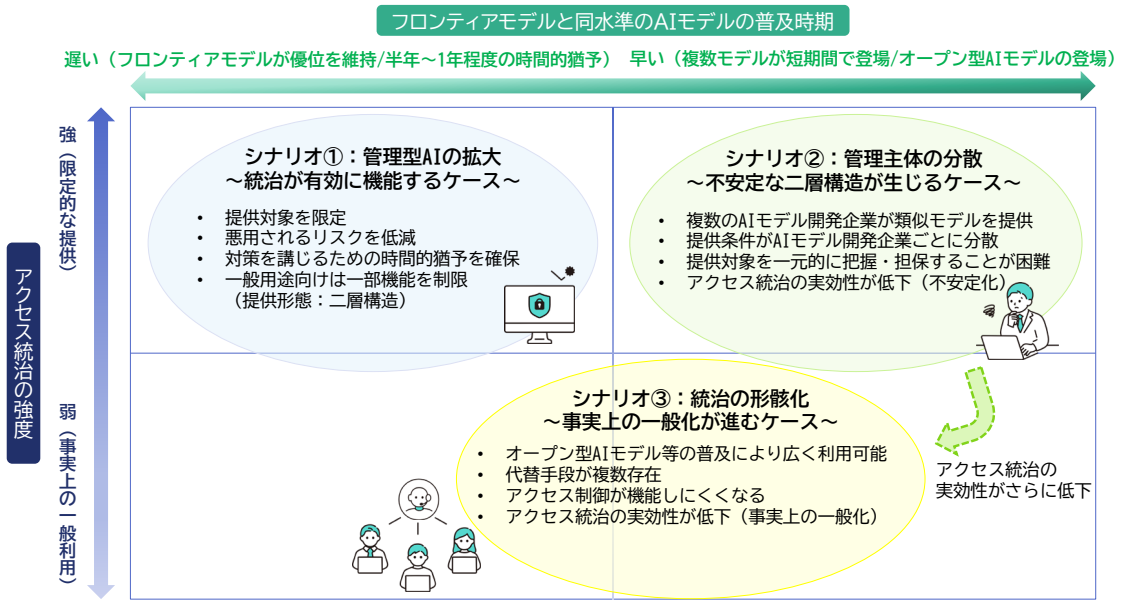
一つ目の軸は**「フロンティアモデルと同水準の AI モデルの普及時期」**である。これは、同水準の AI モデルを利用可能な対象が広がるタイミングを意味する。現在はクローズ型 AI モデル（技術情報が非公開）を提供する米 Anthropic や米 OpenAI といったフロンティアモデルの開発企業が性能面で優位を維持している。しかし、今後米 Meta 社等が、オープン型 AI モデル（技術情報が公開）で同水準に到達すれば、当該能力は誰でも制約なく利用可能となる。また、中国の DeepSeek 社等が同水準に到達すれば、米国企業を中心としたアクセス管理の枠組みの外側で高性能 AI が利用可能となり、提供制御の実効性が低下する。いずれの経路においても、現在の提供制御の前提が崩れかねない。

二つ目の軸は**「アクセス統治の強度」**である。ここでいうアクセス統治とは、AI モデルの能力の拡散抑制やリスク管理を目的に、アクセスの範囲や条件を制御することを指す。AI モデルへのアクセスの制御には、厳格に対象を限定した提供から、一定の条件を満たした企業や機関

への広範な提供、さらには事実上の一般利用に近い状態まで、複数の水準が想定される。

これら二つの軸の組み合わせによって、今後の AI モデルの提供構造は複数の方向に分岐する可能性があり、主に三つのシナリオが想定される（**図表 3**）。

図表 3 今後考えられる AI モデルの提供構造の変化シナリオ



（出所）各種資料より大和総研作成（イラストはソコスト https://soco-st.com/）

シナリオ①：管理型 AI の拡大（統治が有効に機能するケース）

シナリオ①は、開発企業による AI モデルへのアクセス統治が有効に機能し、重要領域向けには提供対象を制限した運用が維持されるケースを指す。現状のように、フロンティアモデルをクローズ型で提供する企業が性能面で優位を維持し、同水準の能力を持つ他の AI モデルが広く普及するまでに、半年から 1 年程度の時間差が存在する状況が該当する。このような環境では、提供対象を限定することにより、高リスク用途への悪用を抑制するとともに、対策を講じるための時間的猶予を確保することが可能となる。この場合、一般用途には高リスク用途に該当する一部機能を制限した AI モデルが広く提供される一方、重要領域向けには提供対象を限定する形で、二層構造が形成される。

さらに、AI モデルが安全保障や重要インフラといった領域に与える影響が大きいため、提供先は開発企業の判断だけでなく、企業が所在する国家との関係性にも左右される可能性がある。このため、自国の政府や企業が必要とする AI モデルに必ずしもアクセスできるとは限らない状況が生じ得る。

シナリオ②：管理主体の分散（不安定な二層構造が生じるケース）

シナリオ②は、重要領域向けに提供対象を制限した運用を維持しようとする一方で、同水準の能力を持つクローズド型 AI モデルが短期間で登場することにより、当該運用の実効性が低下し、不安定な構造となるケースを指す。複数の AI モデル開発企業が異なる基準で提供先を選定

すると、アクセス統治を一元的に担保することが難しくなる。この結果、一般用途向けと制限付き提供という形での二層構造は引き続き存在し得るものの、統治の分散によりその安定性は低下し、シナリオ③に近づく可能性も考えられる。

シナリオ③：統治の形骸化（事実上の一般化が進むケース）

シナリオ③は、同水準の能力を持つオープン型AIモデルが短期間で登場することにより、重要領域向けに提供対象を制限した運用の実効性が大きく低下し、事実上の一般化が進むケースを指す。一般用途向けと制限付き提供という形で設計された二層構造は、シナリオ③では実質的に機能しにくくなるほか、シナリオ①のように悪用リスクを抑制し、その間に対策を講じるという前提も成立しにくくなる。

AIモデルの戦略資産化と政府や企業に求められる対応

AIモデルは、その利用可否が国家の安全保障や企業の競争力に直接的な影響を及ぼし得る段階に達しつつある。こうした環境下では、高度なAIモデルを利用できる状況そのものが、従来以上に大きな戦略的意味を持つようになっている。

現時点におけるAIモデルの提供構造は、シナリオ①で示した通り、アクセス統治が一定程度機能する形で展開しているようにみえる。しかし、この状態が今後も安定的に維持される保証はない。将来のAI開発競争の状況によっては、新たに登場するAIモデルが、悪用リスクへの対策を十分に講じるための時間的猶予を与えるとは限らない。むしろ、より急速かつ非連続的に普及する可能性も否定できない。

なお、前述のシナリオは市場競争と技術普及を軸とした整理であるが、これとは別に、政策的な観点から提供が制限される可能性も考えられる。例えば、モデルの提供にあたって、政府や第三者による安全保障上のリスク評価が求められ、その結果として、AIモデルの提供を厳しく制限するといった対応が取られる可能性がある。

このようなAIモデルの提供構造をめぐる不確実性を踏まえると、企業におけるAIリスク管理は、時間的猶予の存在を前提とした段階的対応では不十分である。むしろ、被害発生を前提に、それを早期に検知し、影響を最小化するための体制を平時から整備しておくことが重要となる。

さらに、中長期的な視点では、特定のAIモデルへのアクセスをいかに確保するかも重要な論点となる。これは個別企業の取り組みに加え、政府の役割も大きい領域である。具体的には、開発企業やその所在国との外交的対応が中心となることが想定されるが、自国における開発・運用、あるいは統制下での利用を志向する「ソブリンAI」といった対応も、必要な計算資源の確保など実現に向けたハードルは高いものの、一つの選択肢として検討対象となり得るだろう。

以上