2026年7月3日　全6頁

Fable 5の提供再開が示すAI規制の限界

個別モデルの規制から普及を前提としたルール形成へ

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
AIアナリティックリサーチ室　主任研究員　新田 尭之

[要約]

■ Anthropic社の最先端AIモデル「Claude Fable 5」「Claude Mythos 5」に対する米国政府の輸出管理指令は、2026年6月30日に解除された。これを受け、同社は7月1日から全世界で一般向けにFable 5の提供を再開した。同社は提供再開の根拠として、①新たな安全対策の導入、②米国政府による安全性の確認、③問題視されたサイバー能力がFable 5固有ではないことの確認、④Fable 5のサイバー能力が、当初懸念されたレベルには達していなかったこと、を挙げている。

■ さらに、複数の外部要因も提供再開に影響した可能性がある。米中のAIモデル性能の格差が急速に縮小する中、提供停止が長引けば、米国企業が市場投入の機会を失う一方、中国のオープンモデルは開発者による利用と改良を通じて性能差をさらに縮めることになる。とりわけ、米国のフロンティアモデルと同等のサイバー能力が中国のオープンモデルでも再現可能になれば、特定の米国モデルを対象とした規制の実効性そのものが低下しかねない。

■ 本件の核心は、民間と軍事のどちらにも適用できるデュアルユース性の強いAIの能力に対し、介入基準や科学的評価枠組みが未整備なまま輸出管理という強い手段が先行した点にある。問題視された能力の多くは他社モデルや中国のオープンモデルでも再現可能であり、拡散が進んだ能力に対して供給元を絞る形の規制は機能しにくい。今後のAI統治は、特定モデルを「止める」規制から、高性能AIの広範な利用を前提とした利用条件・安全対策の設計・管理、すなわち「拡散前提の管理」へと移行していくとみられる。

■ 日本政府には、米国で進む安全性評価基準・業界共通フレームワークの策定段階からの関与と、ソブリンAI（AIの開発・運用における国家の自律性の確保）の観点からの調達先多様化・国産AI基盤の確保が求められる。日本企業には、マルチモデル戦略やAIのBCP（業務継続計画）に加え、脆弱性の即時修正とAIで加速された攻撃を前提としたサイバーレジリエンスの整備が急務であろう。

1. 最先端 AI モデルへの輸出規制措置が解除

米国政府は 2026 年 6 月 12 日、国家安全保障上の懸念を理由に、Anthropic 社の最先端 AI モデル「Claude Fable 5（以下、Fable 5）」「Claude Mythos 5（以下、Mythos 5）」」に関して、外国籍者のアクセスを全面禁止する輸出管理指令（Export Control Directive）を発動した。これを受け、同社は全ユーザーへの両モデルの提供を即時停止した。両モデルは同年 6 月 9 日にリリースされたばかりであり、公開からわずか 3 日後の規制発動であった（詳しくは前回レポート[1]参照）。

その後、同社は規制の発端となった安全対策を回避する手法（ジェイルブレイク）への対策を強化するとともに、米国政府と提供再開に向けた協議を重ねていた。この結果、今回の措置は 6 月 30 日に解除され、7 月 1 日から同社は全世界で一般向けに Fable 5 の提供を再開した。限定提供モデルの Mythos 5 についても、米国政府の承認を受けて 6 月 26 日から米国内の一部組織を対象に提供が再開されており、今後はサイバー防御コンソーシアム（企業連合）「Project Glasswing」に参加する国内外のパートナーへのアクセス拡大が進められる予定である（**図表 1**）。

図表 1 Claude Mythos 5 と Claude Fable 5 の提供再開までの経緯

日付	会社	出来事
2026/6/9	米国・Anthropic	・「Claude Fable 5」と「Claude Mythos 5」を発表 – Fable 5は、Mythos級モデルに安全策を組み込んだ一般向けモデル – Mythos 5は、サイバー防御などへの利用を目的とする限定提供モデル
2026/6/12	米国・Anthropic	・米国政府がFable 5・Mythos 5への外国籍者のアクセス停止を求める輸出規制を適用 ・Anthropicは、対象者を技術的に切り分けることが困難として、全ユーザー向けに両モデルの提供を停止
2026/6/13	中国・Z.ai	・オープンソースモデル「GLM-5.2」を発表 ・一部報道ではサイバー分野でMythos級の性能と評価
2026/6/14	–	・情報セキュリティ業界の経営者・技術者約180名が公開書簡を発表 ・Fable 5規制の妥当性に疑義を呈し、科学的評価に基づく規制判断を提言
2026/6/26	米国・OpenAI	・「GPT-5.6」シリーズを限定プレビューとして発表 – フラッグシップモデルの「Sol」、日常作業向けにバランスの取れたモデル「Terra」、高速で手頃な価格で利用可能なモデル「Luna」の3モデルで構成 ・米国政府との調整を踏まえ、少数の信頼できるパートナー向けに公開
	米国・Anthropic	・米国政府がMythos 5について、米国企業・政府機関など100以上の組織への提供を許可 ・この時点では、Fable 5の一般提供再開は認められず
2026/6/30	米国・Anthropic	・米国政府がFable 5・Mythos 5への輸出規制を解除
2026/7/1	米国・Anthropic	・米国内外を問わず、Fable 5の提供再開 ・Mythos 5は、米国政府と連携し、国内外のパートナーへのアクセス拡大を進める

（出所）各種報道および Anthropic 社公表資料より大和総研作成

[1] 新田 尭之・田邉 美穂「米国政府はなぜ最先端 AI を停止させたのか：最先端 AI モデルへの輸出規制措置が示す AI 統治の転換点」（大和総研レポート、2026 年 6 月 17 日）

Anthropic 社は提供再開に至った根拠として四つの要素を挙げている[2]。

第一に、新たな安全対策の導入である。今回の規制の発端となった Amazon 社の研究者により報告された Fable 5 の安全対策を回避する手法に対応するため、不適切な利用を検知し、下位モデルに切り替える仕組みを改良した。同社によれば、この新しい安全対策により、当該手法は 99%以上ブロックされるようになった。

第二に、米国政府による安全性の確認である。米商務省傘下の CAISI（AI 標準・イノベーションセンター）が新旧の安全対策を検証し、その堅牢性を確認したとされる。

第三に、問題視されたサイバー能力が Fable 5 固有ではないことの確認である。Anthropic 社の検証によれば、前述の安全対策を回避する手法で特定された脆弱性は多くの下位モデルでも特定可能であった。さらに、脆弱性を悪用するコードの生成に至っては、テストしたすべてのモデルで同様の出力が確認されたという。

第四に、Fable 5 のサイバー能力が、当初懸念されたレベルには達していなかったことである。前述の安全対策を回避する手法が引き出した挙動は、危険性が高いとは考えられていないが慎重を期して制限されていたものであり、高度な攻撃的サイバー能力を実証するものではなかったとされる。

こうした安全対策の強化と政府側の確認を経て、6 月 30 日の規制解除に至ったとみられる。

2. 再開を後押しした外部圧力

Anthropic 社の安全対策の強化と並行して、複数の外部要因も提供再開に影響した可能性がある。

中国 AI モデルの台頭

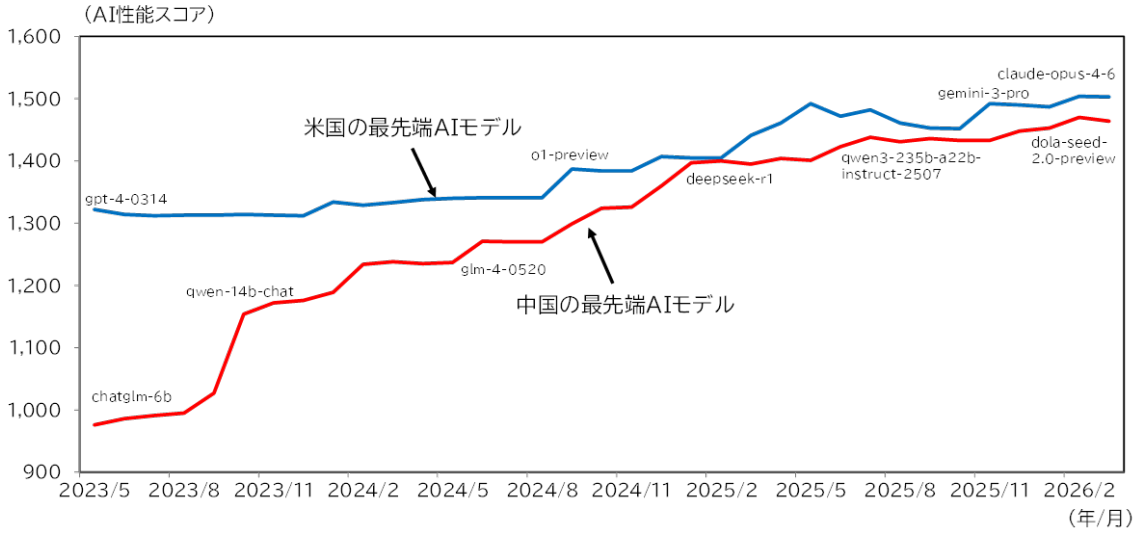
第一に、報道によれば、テック企業の経営者や投資家から、規制の継続が中国の AI 開発者に「貴重な時間」を与えているとして批判が強まっていた[3]。実際、規制発動の翌 6 月 13 日、中国 Z.ai 社はオープンモデル GLM-5.2 を MIT ライセンス（商用利用や改変・再配布を広く認めるライセンス）で公開した。一部の専門家は同モデルのサイバーセキュリティ分野における性能が Mythos に匹敵すると評価しており、システムの脆弱性発見能力の高さが指摘されている[4]。AI モデルの内部構造を含めた広範な公開は、世界的な普及を通じて事実上の標準獲得を狙う中国側の戦略を象徴する動きといえる。

さらにスタンフォード大学HAI（人間中心AI研究所）が公表した「2026年AI指数レポート」でも、米中の最先端AIモデルの性能格差が、過去数年間で急速に縮小しつつあることが示され

[2] Anthropic “Redeploying Fable 5”, June 30, 2026.
[3] Forbes “Anthropic Wins As Commerce Lifts Fable 5 And Mythos 5 Export Controls”, July 1, 2026.
[4] 日本経済新聞 電子版「アンソロピック『Fable』週内に一般提供再開か　背後に迫る中国AI」（2026年6月29日）

ている（**図表 2**）。こうした状況下で米国のフロンティアモデルの提供停止が長引けば、米国企業は市場投入の機会を失う一方、中国のオープンモデルは開発者による利用と改良を通じて性能差をさらに縮めることになる。とりわけ、米国のフロンティアモデルと同等のサイバー能力が中国のオープンモデルでも再現可能になれば、特定の米国モデルを停止するだけではリスクを十分に抑えられず、今回のような規制の実効性そのものも低下しかねない。加えて、企業による AI の業務プロセスへの組み込みが進む中、提供停止の長期化は利用企業の代替モデルへの移行を促し、米国企業が築いてきた顧客基盤の流出につながる可能性もある。

図表 2　米中の最先端 AI モデルの性能スコアの推移



（注）縦軸はユーザーが 2 つの AI の回答を比較・投票して決まる国際的なベンチマークスコア（LMSYS Chatbot Arena）。数値が高いほど性能が高い。データは 2026 年 3 月まで。
（出所）Hugging Face、Stanford University for Human-Centered Artificial Intelligence “The 2026 AI Index Report”より大和総研作成

情報セキュリティ業界からの公開書簡

第二に、情報セキュリティ業界からの問題提起である。6 月 14 日、米国および同盟国の情報セキュリティ分野の経営者・技術者約 180 名が、米商務長官ラトニック氏および国家サイバー長官ケアンクロス氏宛の公開書簡を発表した[5]。この公開書簡は、①Fable 5 で確認されたサイバー能力は他のフロンティアモデルや中国のオープンモデルでも再現可能であること、②中国のオープンモデルは米国フロンティアモデルの数カ月遅れにすぎず、規制の継続は米国の AI 競争力を損なう可能性があること、③AI 規制は、科学的評価と透明性の高い手続きに基づいて行われるべきであり、その判断基準についても一貫したルールの整備が求められること、を主張した。

同書簡は、ハーバード大学のシュナイアー氏など著名なセキュリティ研究者や、大手企業の最高情報セキュリティ責任者が名を連ねた異例のものであり、規制の科学的根拠と手続的正当性をめぐる業界の懸念を示している。

[5] FreeFable.org. “An Open Letter On Transparent AI Cyber Protections”, June 14, 2026.

3. AI 統治は「拡散前提の管理」へ

「基準なき規制」の限界と拡散する AI の能力

今回の本質的な問題は、AI モデル側に危険な能力が本当に存在したのか、あるいは政府対応が過剰だったのかという二者択一にあるのではない。むしろ、民間と軍事どちらにも適用できるデュアルユース性の強い AI の能力に対し、政府の介入基準や科学的評価の枠組みが未整備なまま、輸出管理という強力な手段が先行した点が問題の核心である。

基準が未整備な中、事実上定着しつつあるのが、フロンティアモデルの公開前に政府による審査を介在させる方式である。Anthropic 社は、リリース前の政府評価や脅威情報の即時共有など米国政府との協力を深化させる方針を示しており、OpenAI 社の新しい AI モデルである GPT-5.6 も現時点では米国政府の審査下で限定プレビューにとどまっている。

もっとも、開発サイクルの短い AI モデルを政府が逐一審査する運用には限界がある。加えて、規制対応の結果として安全対策の判定基準が過度に厳格化されれば、通常のコーディングやデバッグ（プログラムの不具合の調査・修正）まで危険な行為として誤検知される事例が増える。実際、Anthropic 社も今回の新たな安全対策にこうした副作用があることを認めている。能力の低い下位モデルへの切り替えが常態化すれば、フロンティアモデルの実質的な価値を低下させかねない。

さらに根本的な課題は、高リスクな能力の拡散である。前述の通り、問題とされたサイバー能力の多くは他社の下位モデルでも再現可能であり、中国のオープンモデルの性能も米国のフロンティアモデルに急速に接近している。同等の能力が国外のオープンモデルで再現可能になれば、特定の米国モデルの提供を止めてもリスクは封じ込められない。かつて米国が武器として輸出を厳しく規制していた暗号技術が、1990 年代にインターネットを通じて世界に普及し、国外で同等製品が流通したことで規制の実効性を失っていったように、拡散が進んだ能力に対して供給元を絞る形の規制は機能しにくい。

こうした状況を踏まえれば、今後の AI 統治は、特定のモデルの提供を「止める」方向ではなく、高性能な AI モデルが広く利用されることを前提に、その利用条件や安全対策をどのように設計・管理するかに軸足を移していくだろう。その具体的な動きとして、Anthropic 社は Amazon 社、Microsoft 社等と共同で、安全対策を回避する手法の重大度を能力向上の程度や兵器化の容易さなどの四つの基準で評価する業界共通フレームワークの策定を進めており、こうしたルールを規制として法制化し、フロンティアモデル開発各社に等しく適用すべきだと主張している。介入基準を事前に共有しておくことは、場当たり的な規制発動を避け、拡散を前提とした管理を支える基盤となろう。また、こうした基準を業界標準として確立することは、AI の安全性評価をめぐるルール形成を主導する動きとしても位置付けられる。

日本政府・企業に求められる対応

日本政府には、第一に、米国で形成されつつあるAIモデルの安全性評価基準や業界共通フレームワークに策定段階から関与することが求められる。これらの枠組みは米国のAI規制の判断基準に直結し得るものであり、形成過程で自国の産業実態や安全保障上の関心を反映させる機会を確保することが重要であろう。

第二に、ソブリン AI（AI の開発・運用における国家の自律性の確保）の観点から、AI の調達先の多様化と国産AI基盤の開発・確保を進める必要がある。特定国の政策判断によってフロンティアモデルへのアクセスが遮断されるリスクが構造的に繰り返され得る以上、代替手段を持たない状態は看過できない。

日本企業には、前回レポートで指摘したマルチモデル戦略や AI の業務継続計画（BCP）の整備に加え、AI を活用することで効率化・高度化されたサイバー攻撃を前提とした運用設計（サイバーレジリエンス）が求められよう。具体的には、脆弱性管理とパッチ（脆弱性を修正するためのプログラム）適用の短サイクル化、コードレビューや設定監査（アクセス権限やクラウド設定などが適切に構成されているかを確認する監査）へのAI活用、代替モデルへの切り替え手順の整備、ログ監視や権限管理の強化などである。攻撃側も同じく AI を用いる以上、「いずれ修正する」ではなく「即時に修正する」運用への転換が必要であろう。

今回の事態は、フロンティアモデルをめぐる競争が単なる性能競争ではなく、「安全に管理しながらどこまで広く使わせるか」を競う段階に入ったことを示している。日本としては、そのルール形成に関与すると同時に、外部依存に備えた技術・調達・運用の三層での備えを急ぐ必要がある。

以上