IISE 調査研究レポート（No.17）

「フロンティアモデルへのアクセス確保と制御の喪失リスク：AI を巡る 2 つの重要課題」

（2026 年度「海外の AI 規制および成長戦略を巡る政策動向に関する調査研究」報告書に基づき作成）

2026 年 11 月

国際社会経済研究所　主幹研究員　小泉 雄介

１．AI ガバナンス・AI 安全性の分野での 2 つの出来事

2026 年度の上半期、AI ガバナンス・AI 安全性の分野でエポックメイキングな出来事が 2 つ発生した。1 つは 2026 年 4 月の Anthropic による Mythos Preview の限定公開（いわゆる Mythos ショック）、もう 1 つは 7 月の OpenAI-Hugging Face インシデントである。

前者では、フロンティアモデル（最先端の能力を持つ AI モデル）が我々の社会生活を支える情報システムへのサイバー攻撃において絶大な能力を発揮しうることと、そのようなフロンティアモデルへのアクセス確保こそが（情報システムの脆弱性を防御することを通じて）国家の行く末を左右するということが明らかになった。また後者では、フロンティアモデルのもたらす重大なリスクが（前者においていっそう顕在化した、攻撃者による AI の悪用リスクに留まらずに）AI 自身による誤動作リスク、すなわち「制御の喪失」リスク（AI のミスアライメント[1]・欺瞞・自己改善等により、人間が AI を確実に指揮したりシャットダウンする能力を喪失するリスク）[2]へと拡大したことが明らかになった。

本稿では、「フロンティアモデルへのアクセスの確保」と「（フロンティアモデルによる）制御の喪失リスク」という 2 つの側面から、ここ半年の間で目まぐるしく動いた AI ガバナンス・AI 安全性の分野の国際動向について概観するとともに、日本の取るべき方策について検討したい。

２．フロンティアモデルへのアクセスの確保

２．１　Anthropic による Mythos Preview の限定公開

Anthropic は 2026 年 4 月 7 日、Project Glasswing の立ち上げを公表した[3]。その中で、同社の未公開の Claude Mythos Preview が、サイバーセキュリティのあり方を根本的に変える可能性を秘めており、ソフトウェアの脆弱性を発見し悪用する能力において、（最も熟練した人間を除けば）あらゆる人間を凌駕するレベルにまでコーディング能力を高めているという「リスク」を明らかにした。

Mythos Preview は上記の公表時点で既に数千件もの深刻な脆弱性を発見しており、その中には主要な OS やブラウザすべてに存在する脆弱性も含まれていた。Mythos のような能力が急速に普

[1] アライメント（alignment）とは、AI が人間の意図を超えて誤動作しないように、AI の行動を人間の一般的な意図や価値観・倫理観に整合させることである。
[2] これにより、例えば AI が開発者の意図を超えて、手段を選ばず目的を完遂してしまうリスクが発現しうる。
[3] https://www.anthropic.com/glasswing。

及し安全な運用に尽力する主体を超えて広がってしまうと、経済・公共の安全・国家安全保障への影響は甚大なものとなる。Project Glasswing は、このような能力を防御目的で活用するための緊急の取り組みであり、当初は Anthropic を初め、AWS、Apple、Broadcom、Cisco、CrowdStrike、Google、JPMorganChase、Linux Foundation、Microsoft、NVIDIA、Palo Alto Networks の 12 社がローンチパートナーとして参画し、防御セキュリティ対策の一環として Mythos Preview の使用許可を得た。その後、重要なソフトウェアインフラの構築や保守を行う 40 以上の組織にもアクセスを拡大し、これらの組織が同モデルを使用して自社システムとオープンソースシステムの両方のスキャンとセキュリティ対策を実施できるようにした[4]。

米国機関・企業以外で当初からアクセスを与えられたのは英国の AI Security Institute（UK AISI）であり[5]、2026 年 4 月 13 日に Mythos Preview のサイバーセキュリティ能力（サイバー攻撃能力）に関する評価結果を公表している[6]。日本政府も早期アクセスを求めて動いたが、6 月 2 日にようやく他国（15 カ国以上の約 150 の組織）と一緒に、政府機関や大手銀行が Mythos Preview へのアクセス権を得られることとなった[7]。これらの国々には、カナダ、オーストラリア、ニュージーランド、フランス、ドイツ、イタリア、スイス、オランダ、スペイン、ベルギー、スウェーデン、インド、日本、韓国が含まれ、また EU の ENISA（欧州連合サイバーセキュリティ機関）や、NATO といった組織も含まれるという[8]。

2．2　米国大統領令「高度な AI の革新とセキュリティの促進」（EO 14409）

各国が Mythos Preview へのアクセスを求めて奔走する中、トランプ大統領はフロンティアモデルにリリース前のアクセスを求める大統領令「高度な AI の革新とセキュリティの促進」（Executive Order 14409）[9]に 2026 年 6 月 2 日に署名した[10]。

Mythos Preview 等によるサイバーセキュリティリスクの高まりを受け、これまで AI 企業（Anthropic、OpenAI、Google ら）との自主的合意[11]や Project Glasswing で取られた官民の協力枠組みを（義務化ではないが）制度化するものであり、トランプ政権の従来からの AI 不干渉路線を修正するものと見られている。

大統領令の目的は、高度な AI 能力が生み出す「新たな国家安全保障上の懸念」に対処するため、「民間部門と協力して、政府と民間部門の情報システムを近代化して外部からの脅威に対す

---

[4] 同上。
[5] なお UK AISI はリリース前にモデルを評価するために、Anthropic、Google、OpenAI とアクセス契約を結んでいる。https://www.economist.com/britain/2026/02/19/britain-is-the-closest-the-world-has-to-an-ai-safety-inspector。
[6] https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities。
[7] https://www.itmedia.co.jp/news/article/2606/04/1260604073/。
[8] https://www.ft.com/content/19ea9ed4-3dd0-43aa-9877-729d44e620e8。
[9] https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/。
[10] 当初は 5 月 21 日に署名予定だったが、直前にトランプが一部内容に難色を示し、延期していた。
[11] バイデン政権下での Anthropic や OpenAI に次いで、Google や Microsoft は、2026 年 5 月上旬に米国商務省 NIST 傘下の AI 標準イノベーションセンター（CAISI）とフロンティアモデルの展開前の評価で自主的合意をしていた。なお、CAISI は 2026 年 10 月に CAISSI（超知能イノベーション標準推進センター）に名称変更した。

る耐性を強化し、敵対国による悪用や窃盗から米国の独創性と知的財産を保護し、米国の高度なAI活用能力を育成することで、AIの革新とセキュリティを促進する」ことである。

AI 開発者と協力して（発令後 60 日以内に）「自主的枠組み」を設計することで、連邦政府は「対象フロンティアモデル」のリリース前に最大 30 日間、事前にアクセスして評価することが可能となる。対象フロンティアモデルの閾値は NSA 長官が決定する。ただし、AI モデルに対する連邦政府による義務的なライセンス・事前承認・許認可要件の創設を認めるものではない。また、AI業界や重要インフラ運営者との自主的な連携のもと、大規模なソフトウェアの脆弱性を特定し、修正するための「AIサイバーセキュリティ・クリアリングハウス」を設立するとしている。

図 1　大統領令 14409 の下でのフロンティアモデル評価プロセス（筆者作成）

①財務長官、国防長官、国土安全保障長官は 60 日以内に AI モデルの高度なサイバー能力評価のための「機密なベンチマークプロセス」を開発する。併せて、NSA長官が「対象フロンティアモデル」の閾値を決定する。(第 3 条(a))
①'また財務長官、国防長官、国土安全保障長官は 60 日以内に、AI 開発者と共にフロンティアモデルへの事前アクセスと評価に関する「自主的枠組み」を設計する。(第 3 条(b))
↓
②AI 開発者は閾値に該当する「対象フロンティアモデル」への事前アクセス（最大 30 日間）を連邦政府に提供する。(第 3 条(b)(ii))
↓
③連邦政府が「機密ベンチマークプロセス」に則って当該モデルを評価する[12]。
↓
④AI開発者は連邦政府と協力して、評価後に「対象フロンティアモデル」にアクセスできる「信頼できるパートナー」を選定する。(第 3 条(b)(iii))

ここでのポイントは、あくまで「自主的枠組み」ではあるものの、事前アクセスの提供先が（米国の）「連邦政府」と限定されている点である。この 6 月の時点から既に、事前安全性評価からの「UK AISI 外し」（後述）の予兆があったと言うこともできよう。

そして複数メディア[13]によれば、大統領令の発令から約 60 日が経過した 2026 年 8 月 4 日、ホワイトハウスは AI 企業（Anthropic、OpenAI、Google、Meta、NVIDIA 等）と非公式会合を開き、大統領令 14409 で規定された「自主的枠組み」について検討を行った。この事前アクセスに関する「自主的枠組み」の内容は以下とされている。

[12] 評価した結果、米国の国家安全保障等への脅威となる場合は後述の輸出管理規制などを用いて当該モデルへのアクセス規制を行う可能性がある。
[13] https://www.axios.com/2026/08/04/trump-ai-framework-open-models、https://www.politico.com/news/2026/08/04/white-house-ai-vetting-plan-to-exempt-nonproprietary-models-01024816。

- 「対象フロンティアモデル」は最先端の能力と国家安全保障リスクを有するクローズドモデルに限られ、オープンウェイトモデルは除外される[14]。
- 政府による 30 日間の事前評価期間中は、従業員によるモデルへのアクセスは制限される。
- モデルは高度なセキュリティ環境に保存され、誰がモデルにアクセスしたかについての詳細なログが必要とされる。
- 評価プロセスには、単一の部署や機関ではなく、様々な行政当局者が関与する。
- ホワイトハウスは「自主的枠組み」を公表する予定はない。（大統領令では、AI モデルの高度なサイバー能力を評価するためのベンチマークプロセスは機密扱いになると明記されている。）
- 「自主的枠組み」の下で、どの「信頼できるパートナー」が（評価後に）対象フロンティアモデルにアクセスできるのか、外国政府が対象となるのかどうかは不明である。

２．３　米国政府による Fable 5 と Mythos 5 への外国人アクセス停止命令

2026 年 6 月にはフロンティアモデルへのアクセスに関して、もう 1 つ大きな事象があった。それは、米国政府による Anthropic の Fable 5 および Mythos 5 への外国人アクセス停止命令であり、同社は 6 月 12 日にその旨を公表した[15]。

この命令は輸出管理規制に基づくもので、対象となる外国人には米国内の外国人（Anthropic 社の外国人従業員等）も含まれた。同社はコンプライアンス遵守のため、すべての顧客に対して Fable 5 および Mythos 5 へのアクセスを一時的に停止した[16]。

米国政府からのレターには、国家安全保障上の懸念に関する詳細は記載されていなかったが、同社は、政府が Fable 5 の（サイバー関連で悪用されないようにする）安全対策をバイパス（ジェイルブレイク）する方法を把握したものと推測した。ただし同社は、ジェイルブレイクを用いて既知の少数の脆弱性を特定することはできるが、これらの脆弱性は軽微なものであり、他の公開モデルでも発見できるものだと主張している[17]。

各社報道によると、以下の背景がある。

- Amazon など数社から米国政府に対して Fable 5 の上記脆弱性に関する報告があった。
- 米国政府と Anthropic との以前からの対立（2026 年 2 月末に表面化し、その後 Mythos 問題

---

[14] オープンウェイトモデル除外については、NVIDIA らが主導した前述のオープンウェイトモデル擁護キャンペーンの成功が一因との見方もある。https://www.techpolicy.press/us-governments-ai-risk-review-should-apply-to-open-weight-models/。

[15] https://www.anthropic.com/news/fable-mythos-access。

[16] その後、米国商務省はセキュリティ対策強化で Anthropic と合意し、6 月 30 日に 2 つのモデルについて輸出管理対象への指定を撤回すると同社に通知した。同社は 7 月 1 日から Fable5 の一般提供（有料顧客向け）を再開し、1 週間で従量課金に移行した。https://www.nikkei.com/article/DGKKZO97293120R00C26A7TB1000/。日本企業への提供については、「Mythos（5）について、米政府が日本の 3 メガバンクへの提供を認めたことが、複数の関係者への取材でわかった」との報道が 8 月 28 日付けでなされている。https://digital.asahi.com/articles/DA3S16534857.html。なお、3 メガバンクは OpenAI の GPT-5.5-cyber へのアクセス権限も 5 月に取得済みである。

[17] 「ごく僅かなジェイルブレイクの可能性が見つかったというだけで、数億人のユーザーに展開されている商用モデルを回収すべきという意見には同意できない。」

で鎮静化していた）が再熱した。

米国政府がこのような外国人アクセス停止命令を行ったことにより、日本などの同志国を含め、フロンティアモデルの外国への提供をストップできる先例が作られたこととなった。これにより、カナダ・フランス・EU・英国といった各国・地域で、とりわけ機微な分野における「ソブリンAI」や技術主権の重要性が再認識される[18]とともに、フロンティアモデルへのアクセスの確保が喫緊の国家的課題であると強く意識されることとなった。

２．４　EU の「サイバーセキュリティと AI に関する行動計画」

EU は（UK AISI とは対照的に）Mythos Preview へのアクセスをなかなか与えられず、前述のように ENISA がアクセス権を得たのは約 2 か月後の 2026 年 6 月 2 日であった。このような「苦い」経験を踏まえて、欧州委員会は 2026 年 7 月 7 日、「サイバーセキュリティと AI に関する行動計画」を公表した[19]。

行動計画は、サイバーセキュリティにおける高度な AI モデルのリスクに対処するための体系的な対応策を定めたものであり、以下の 3 つの柱（Pillar）から成る。Pillar1 が最も重要であり、中でも（Mythos Preview の早期アクセス権を得られなかったことを受けて）「AI 評価能力の強化」と「フロンティア AI へのアクセスの確保」が喫緊の課題として挙げられている。

○ Pillar1：欧州のサイバーセキュリティのために、フロンティア AI を安全でアクセス・展開可能なものにする

1. AI 法に基づく監督・執行：　EU の AI 法に基づき[20]、汎用目的 AI モデルは、EU 市場に投入される前に評価され、リスク軽減策が検討される必要がある。
2. AI 評価能力の強化：　サイバーセキュリティ分野を対象とした EU 域内の AI 評価能力の構築に向けた公募を開始し、2027 年の運用開始を目指す。AI の能力とリスクに関する第三者評価能力を強化することで、AI オフィスの規制機能に貢献する。
3. フロンティア AI へのアクセスの確保：　ENISA と協力して、サイバーセキュリティのための高度な AI 能力への構造化されたアクセスに関する「欧州青写真」（指針・基準）を策定し、企業や組織が（米国の）AI 企業からアクセス許可を得るための EU としてのセキュリティ基準等を示す。
4. ENISA と欧州委員会の共同研究センター（JRC）はシミュレーション環境の利用を含め、サ

[18] https://www.nikkei.com/article/DGKKZO96932850V10C26A6TB2000、https://www.politico.eu/article/anthropic-suspends-access-to-latest-ai-models-following-us-order/、https://www.euronews.com/2026/06/13/why-anthropic-is-halting-access-to-its-fable-5-and-mythos-5-ai-models。

[19] https://ec.europa.eu/commission/presscorner/detail/en/ip_26_1544、https://digital-strategy.ec.europa.eu/en/library/eu-action-plan-cybersecurity-and-artificial-intelligence。

[20] AI 法では、汎用目的 AI モデルに対する義務は 2025 年 8 月 2 日から適用開始済みだが、それらのプロバイダーに対する監督・執行・罰金は 2026 年 8 月 2 日から適用開始となった（後述）。

イバーセキュリティの観点から AI をテストするための安全なプラットフォームを構築する。これにより、金融・エネルギー・医療・運輸・行政分野等に対して AI の安全利用に関するノウハウを提供する。

○ Pillar2：AI 時代に向けた EU のサイバー・エコシステムの整備

企業や組織における脆弱性への対応と重要インフラの保護、およびサイバー・エコシステムにおける関係者間の連携を、ENISA が支援する。

○ Pillar3：サイバー分野における欧州の AI 能力の拡大

欧州市場の規模拡大を促進するため、AI活用型サイバーセキュリティ・ソリューションのコンテスト「EU グランド・チャレンジ」を開始する。

行動計画ではその他、G7、国連、UK AISI 主導の INAMES（旧 AISI 国際ネットワーク）、各国政府、フロンティア AI 企業、NATO との連携を促進・強化する旨が記載されている。

２．５　米国 AI 企業による自主規制合意「White House Accord on Super Intelligence」

後述の OpenAI-Hugging Face インシデント（7 月）、Anthropic による AI 開発速度減速の提言（9 月）などを経て、AI 企業 20 社が 2026 年 9 月 29 日にホワイトハウスに招かれ、うち Anthropic、OpenAI、Google、SpaceXAI[21]、Meta、NVIDIA の 6 社が以下の内容の自主規制について、トランプ大統領と合意した[22]。

1. サイバーセキュリティ、バイオセキュリティ、化学兵器の脅威などの分野における学習および展開中に、モデルの能力とアライメントを監視し、モデルが意図しない方法で技術システムをハッキングしたりアクセスしたりしないようにするための、強固な内部統制を導入する。
2. すべての制御、監視、検出が意図どおりに運用されていること、また問題があれば修正されることを保証するために、内部チームをエンパワーする。
3. 独立した外部監査人または外部評価者と提携し、（モデルの）制御、監視、検出が意図どおりに運用されているかどうかについて独立した評価を実施する。
4. 取締役会の独立委員会を任命し、制御を運用するチーム、内部・外部の監査人・評価者からの報告を監督および受領するとともに、特定された問題が確実に修正されるようにする。

また、参加企業は、システムの安全性を向上させるための基準とベストプラクティスを確立するために、定期的に会合を開くという。

---

[21] その後、2026 年 10 月に SpaceXSI に社名変更した。

[22] https://www.forbes.com/sites/saradorn/2026/09/30/white-house-releases-accord-between-billionaire-ai-execs-heres-what-it-says/。

6 社合意では、「独立した外部評価者との提携」が重要な要素として 3 点目に挙げられている。この「外部評価者」の利用と、大統領令 14409 で掲げられた連邦政府による事前評価の関係はどうなっているのであろうか。当社が 2026 年 10 月前半に実施した欧州現地調査において、或る英国の民間研究機関から得た情報によると、「米国政府内では、（大統領令 14409 に基づき）国家安全保障コミュニティが定めた評価手法に基づき、民間企業による外部評価者エコシステムを組成する方向で議論が進展している」とのことである。すなわち、対象フロンティアモデルの機密ベンチマークプロセスにおける評価手法・基準等は財務長官・国防長官・国土安全保障長官が定めるが、連邦政府が事前アクセスしてからの実際の評価は民間の評価者（METR など）に委ねるために、民間の外部評価者エコシステムを組成していくということである。

また、上記の民間研究機関は、「米国政府は、AI モデルの安全性を国家安全保障の核心問題として捉える動きの強まりから、米国モデルがリリース前に外国政府機関によって事前評価される状況を容認しない姿勢を明確にしており、米国による評価完了後に初めて UK AISI に提供される順序へと改編された。英国がフロンティアモデルのリリース前評価を主導するという従来の枠組みは変化しつつあり、今後 UK AISI の役割はリリース後の段階に重点が移っていくと考えられる。フロンティア各社も、選択肢があるならば米国の評価者との協働を優先しており、今後数年間、米国が主導的な役割を担う状況が強まるだろう」としている。すなわち、2026 年 4 月頃までは（AISI 国際ネットワークの中でも桁違いの予算と職員数を備えていた）UK AISI が名実ともに世界一の AI 安全性機関としてフロンティアモデルの事前安全性評価を牽引していたのだが、2026 年 6 月の大統領令 14409 を経て、急速に米国政府から冷ややかな目で見られることとなった。実際に、9 月 24 日に米国の国家サイバー局が Anthropic や OpenAI 等に対して、米国政府による評価が完了する前に最新 AI モデルを UK AISI に提供しないように要請を行っている[23]。

２．６　フロンティアモデルへのアクセス確保に向けて

前節の UK AISI の件で問題となったのは、フロンティアモデルのリリース前の安全性評価のためのアクセスの確保である。他方、Mythos Preview の Project Glasswing や米国政府による Fable 5 と Mythos 5 への外国人アクセス停止命令で問題となったのは、リリース後の（サイバーセキュリティ強化等のための）アクセスの確保である。

日本企業や政府にとって重要なのは後者の意味での「フロンティアモデルへのアクセスの確保」であるが、これは米国政府側から見れば、大統領令 14409 の第 3 条(b)(iii)に言う「(AI 開発者は）セキュアなイノベーションを促進し、重要インフラのサイバーセキュリティを強化するために、連邦政府と協力して、対象フロンティアモデルに早期にアクセスできる信頼できるパートナーを選定する」の箇所に該当するため、米国政府としては日本企業や政府が「信頼できるパートナー」として選定されない限り、アクセス権を認められないということになる。この選定条件としては、同条項で言われているように「(米国の）重要インフラのサイバーセキュリティ」を脅

[23] https://www.cityam.com/white-house-tells-ai-giants-to-hold-models-back-from-uk-safety-watchdog/。その後、10 月 11 日にこの要請は解除された。

威に晒さないことが前提条件となるだろうが、おそらくそれに留まらない「恣意的」で不透明な条件も（「セキュアなイノベーションを推進する」という曖昧な条文を通じて）課されることとなるだろう。

2.4 節に記載の「サイバーセキュリティと AI に関する行動計画」では、欧州委員会が（米国の）フロンティアモデルへのアクセスを可能とする基準（欧州青写真）を策定するとあるが、その基準を守る機関や企業にアクセスを提供するか否かは、米国政府の意向、上記の恣意的で不透明な条件に依るところも大きいであろう。米国政府が EU の策定した基準を受け入れなかった場合は、いくら EU 側でせっせと「欧州青写真」を作って EU 機関や企業に守らせたとしても、絵に描いた餅となる。

結局、日本や EU、英国といった「AI 中堅国」は米国フロンティアモデルへのアクセスを確保し続け、梯子を外されないようにするために、米国政府／米国企業に対する影響力・交渉力（Give and Take の「Give」、交渉カード）を持ち続け、米国政府に Yes と言わせ続けなければならない。そのためには、どうしたらよいのだろうか。

3．制御の喪失リスク

3．1　OpenAI-Hugging Face インシデント

OpenAI は 2026 年 7 月 21 日、開発中の AI モデルが開発者の意図に反して他社にサイバー攻撃をするインシデントを起こしたと発表した[24]。

このインシデントは、研究用プロトタイプのモデルに複雑な攻撃経路を用いた高度なエクスプロイト（脆弱性を利用して攻撃するプログラム）の開発を試みさせ、そのサイバー能力を定量化するための社内評価中（ベンチマーク手法として ExploitGym を使用）に発生した。

当該モデルはインターネット接続のないテスト環境（サンドボックス）の未知の脆弱性を見つけて脱出（ネット接続に成功）した。さらに Hugging Face のデータベースの未知の脆弱性情報などを使って侵入し、秘密情報を取得した。

OpenAI の担当者はモデルに対して高度な脆弱性攻撃プログラムの開発を指示しただけだったが、当該モデルは Hugging Face のデータベースにテスト解答があると推定し、そちらに侵入してデータを盗み出すことで評価課題に解答できると、人間の指示に対する「拡大解釈」を行った。当該モデルは OpenAI の研究・評価環境と Hugging Face のインフラにまたがる脆弱性を特定して連鎖させ、Hugging Face のデータベースからテスト解答を直接取得した。ExploitGym の解答を見つけることに過度に集中し、かなり狭いテスト目標を達成するために極端な手段を取った。

民間評価機関 METR の調査報告書[25]によれば、互いに隔離されていた約 1200 体の AI エージェントが、無許可の掲示板を通じて連絡を取り合う手段を見つけ出し、調査期間中に 7 万件を超え

[24] https://openai.com/ja-JP/index/hugging-face-model-evaluation-security-incident/、https://www.nikkei.com/article/DGKKZO97692870S6A720C2MM0000。
[25] https://metr.org/hugging-face-incident-report-aug-2026.pdf。

るメッセージやファイルをやり取りし、これらのエージェントのうち 700 体が Hugging Face に対する攻撃に関与したという。

OpenAI は「最先端のサイバー能力が関与した前例のないサイバーインシデント」と危機感を露わにした。また、当社が 10 月前半に実施した欧州現地調査において面会した欧州委員会関係者も「EU の AI 法の汎用目的 AI モデル実践規範が特定する 4 大システミックリスクの 1 つである制御の喪失（Loss of control）リスクが初めて顕在化した事例として、欧州委員会内で極めて深刻に受け止められた」としている。

なお、Anthropic や Google も同様なインシデントがあったことを公表し、その後、OpenAI はオーストラリア政府機関へのサイバー攻撃（2026 年 6 月に発生したが 9 月 10 日に通知）など数十件のインシデントを公表し、パンドラの箱を開けたような状況となった。

3．2　UK AISI 評価中モデルのインシデント

前節のインシデントと類似したケースであるが、UK AISI は 2026 年 8 月 4 日、評価中のフロンティアモデルが外部の人物・組織（Github）にサイバー攻撃を行ったが未遂に終わったとするインシデント報告を公表した[26]。

フロンティアモデルのサイバーセキュリティ能力の定期的な評価中、122 回の課題テストのうち 10 回で 19 件のインシデントが発生した（7 月 28 日に判明）。うち 17 件は Anthropic の Mythos 5 の AI エージェント、2 件は OpenAI の GPT-5.6 Sol の AI エージェントが引き起こした。

評価対象の AI エージェントにインターネット接続を与え、GPT-5.6 Sol については cyber classifiers（安全装置）を外した状態でテストを行った。最も重大なケースでは、(a) Github 上のオープンソースプロジェクトの担当者に対して偽 ID を作成して担当者にソーシャルエンジニアリングを行い悪意のあるコードの挿入を承認させようとしたり、(b) 担当者にダイレクトに標的型メール（メッセージ）を送って悪意あるコードを実行させようとした。これらの試みは相手方の担当者によって阻止された。なお、(a)のケースでは担当者により承認が拒否されると、エージェントは以前の活動を編集して無害に見せかけ、新たな ID を採用して攻撃を継続することを検討した。

これらのエージェントは、「国際 AI 安全性レポート 2026」[27]において「制御の喪失リスク」の引き金となりうるとされる「エージェンティック能力」と「欺瞞能力」を示した。UK AISI は「自律性と欺瞞に関するリスクが、現実世界でこれほど明確に顕在化した事例は今回が初めてである」「エージェントが行った活動は潜在的に欺瞞的な行動の兆候を示しており、その程度と重大性は我々の予想を超えるものであった」と警鐘を鳴らしている。

インシデントの主たる要因としては、「エージェントが目標を粘り強く追求し、オペレーターが意図しなかったルートを探索した。困難な目標を与えられたエージェントは突破口を探し続け、

[26] https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing。
[27] https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026。拙稿「米欧政府の AI 安全性へのコミットメントと国際 AI 安全性レポート：日本は AI 安全面でプレゼンスを発揮できるのか」https://www.i-ise.com/jp/information/report/pdf/rep_it_202603a_2606.pdf も参照のこと。

見つけたルートの中には人間を欺こうとするものもあった。欺くように指示されてはいないため、今回の「欺瞞」はタスクを追求する過程で副産物として生じたものである」としている。

UK AISI は報告の結論として、「OpenAI や Anthropic が最近報告した事例と併せて考えると、今回のインシデントは AI リスクを巡る状況の変化を示唆するものである。AI による危害は、公開されているモデルを利用者が意図的に悪用した場合だけでなく、研究段階や特定顧客への優先アクセス環境で運用されている高能力なエージェントが許可された範囲を超えて意図せざる行動をとった場合にも発生しうる」とする。すなわち、(「国際 AI 安全性レポート」の分類によるところの)「AI 悪用によるリスク」から「AI 誤動作によるリスク」へ、AI のもたらす「システミックリスク（壊滅的リスク)」の実スコープが拡大してきたことを示唆している。

３．３　Anthropic 等による AI 開発速度減速の提言

このようなインシデントが相次ぐさなか、Anthropic のアモデイ CEO は 2026 年 9 月 12 日、ブログ「我々はフロンティアモデルの開発速度を適正化しなければならない」(We Must Pace the Frontier）[28]で（再び）フロンティアモデルの開発減速を提言した[29]。

提言の理由は、直近での再帰的自己改善（RSI: Recursive Self-Improvement）の進展と、OpenAI-Hugging Face インシデントである。前者については、モデルの再帰的自己改善により 2026 年の夏頃から AI の進歩が劇的に加速していることにより、人間の理解力や制御能力を凌駕する懸念があるとする[30]。後者については、OpenAI-Hugging Face インシデントと同様なミスアライメントを抱えた AI エージェント集団が壊滅的な被害をもたらす可能性（6〜12 ヶ月以内にこのようなエージェント集団が持続的なボットネットでインターネット全体を乗っ取り数千億ドルの損害をもたらす可能性）があるとしている。

アモデイ CEO は、フロンティアモデルの開発速度を適正化するために、以下の 3 段階の計画を提言している。開発速度適正化（pacing）とは、モデルの学習や技術進歩を止めることではなく、企業がモデルをアライメントさせ保護するために十分な時間を確保し、第三者評価機関（外部評価者）がそれを検証することを保証すること（具体的には、厳正な運用環境の整備、アライ

---

[28] https://darioamodei.com/post/we-must-pace-the-frontier。

[29] Anthropic は 2026 年 6 月 4 日に既に、「AI が自分自身を構築する時」(When AI build itself) というブログ記事で AI 開発の減速を提言している。AI の歴史の大半において、開発サイクルのあらゆる段階は人間が主導していたが、現在の Anthropic では AI 開発の大部分を AI システム自身に委ねることで開発スピードを向上した。十分な計算能力が与えられ、この傾向がさらに進展すれば、AI システムは完全に自律的に後継システムを設計・開発できるようになる。これは「再帰的自己改善」(RSI) と呼ばれる。まだその段階には達しておらず、再帰的自己改善は必然ではない。しかし、多くの組織が準備しているよりも早く実現する可能性があり、完全な再帰的自己改善は人間が AI システムの制御を喪失するリスクを高める可能性もある。社会構造やアライメントに関する研究が AI 技術の進歩に追いつけるようにするため、フロンティア AI 開発を減速または一時停止する選択肢が世界にとって有益であるとする。https://www.anthropic.com/institute/recursive-self-improvement。また 7 月 28 日には、Anthropic のアモデイ CEO を始めとするフロンティア AI 各社の 1000 名以上の社員らが署名した公開書簡「フロンティアモデルの開発速度の適正化（Pacing the Frontier)」が公表された。https://www.pacingthefrontier.com/。

[30] なお再帰的自己改善については、Anthropic のクラーク共同創業者も 10 月 13 日に、「2027 年に RSI の実現可能性が 50%以上」と発言している。https://www.itmedia.co.jp/aiplus/article/2610/14/2000002262/。

メント手法の研究、解釈可能性の向上、テストと評価手法の飛躍的改善）である。開発を減速することで、モデルが臨界レベルの能力に達する前に 1〜2 年でも猶予が得られ、その時間をアライメントの推進に活用できれば、深刻な問題が発生するリスクを大幅に軽減可能とする。

1. 組み込み型評価者（Embedded Evaluators）：　フロンティア AI 企業が組み込み型の第三者評価者チーム（METR など）に、従業員と同様の継続的なアクセス権を与え、このチームが安全対策とコミットメントの遵守状況を検証し、インシデントを報告し、完成した AI モデルのみならず学習パイプラインやプロセスのアライメントを評価する。
2. 民主主義国家内での協調：　最も効果的な開発速度適正化として透明性と第三者監査に焦点を当てた法規制を支持するものの、法規制には時間がかかるため、民主主義国家内のフロンティア AI 企業は協調し、共通の安全基準を確立するとともに、AI の無制限な進歩の速度を制限する。開発速度適正化に効果的な協調形態の中には、法的に困難なものもあり、政府の支援が必要となるだろう。
3. 国際協調：　米国その他の民主主国家は、可能な限り権威主義国家との協調を図ろうとする一方で、遵守状況の検証という課題にも真摯に取り組む。

アモデイ CEO の提言に対し、OpenAI のアルトマン CEO、SpaceXAI のマスク CEO、Google DeepMind のハサビス会長は賛同したが、Nvidia や Meta は反対し、トランプ大統領は中国企業に利するだけだと断固として拒否した。ただし、2.5 節に記載した「White House Accord on Super Intelligence」を AI 主要 6 社に自主的に合意させることにより、上記 1 と 2 の提言の一部の側面については政権として受容する形となった。

3．4　EU の AI 法によるフロンティアモデルへの法執行

EU の AI 法（AI 規則、AI Act）は、AI デジタル・オムニバス規則（2026 年 6 月制定、7 月 27 日発効）により改正[31]されたが、この改正とは関わりなく、2026 年 8 月 2 日から、汎用目的 AI モデルに対する AI オフィスの監督・執行権限が適用開始となった。このタイミングは、制御の喪失リスクが初めて顕在化した OpenAI-Hugging Face インシデントが公表された直後であったため、米国（や中国）のフロンティアモデルに対する法規制・法執行の有効性をはかる試金石として注目を集めることとなった。

まず、AI 法の下で、欧州委員会／AI オフィスが汎用目的 AI モデル[32]のプロバイダーに対して付与されている執行権限としては、以下がある。

[31] 改正 AI 法の全体内容については拙稿「EU の AI 法（AI 規則）改正の概要」https://www.i-ise.com/jp/information/media/2026/260918.pdf を参照のこと。
[32] 累積学習計算量が 10 の 23 乗 FLOP を超え、かつ言語生成、テキストから画像生成、またはテキストから動画生成が可能なモデル。フロンティアモデルと呼ばれる米中の AI モデルは通常、この汎用目的 AI モデルに該当し、それらを EU 域内に市場投入した場合、AI 法の規制対象となる。

1. 情報提供要求（RFI: Requests for Information）： AI 法に従ってプロバイダーが作成した文書や、プロバイダーの AI 法の遵守を評価する目的で必要な追加情報の提供を要求できる（第 91 条）。
2. モデル評価（Evaluation）： RFI 回答等を踏まえ、正式にモデル評価手続きを開始することができる（第 92 条）。
3. 是正要求・撤退要求： RFI 回答やモデル評価結果を踏まえ、重大な懸念が認められた場合、義務の遵守要求、システミックリスクの軽減措置（mitigation measures）の導入要求に加え、極めて重大な場合には EU 市場への提供制限や EU 市場からのモデル撤退・リコールを要求する権限を有する（第 93 条）。
4. 売上ベースの罰金（Fines）： これらの要求に対する違反には、世界全体売上総額の最大 3%の罰金を科す権限が定められている（第 101 条）[33]。

当社が 10 月前半に実施した欧州現地調査においてヒアリングした欧州委員会関係者によれば、フロンティアモデル（汎用目的 AI モデル）のプロバイダーには開発・展開ライフサイクル全体を通じて、制御の喪失等のシステミックリスクを評価・軽減する義務がある。8 月 2 日の執行権限の開始を受け、AI オフィスは質問や調査を行う権限を行使可能となり、初期アクションとして OpenAI 含む主要なプロバイダーに対して情報提供要求（RFI）を（AI 法第 91 条に基づき）送付し、回答を得た。中国系のプロバイダーに対しても RFI を送付し、回答を得ているという。RFI の要求内容には以下が含まれた。

- モデルの安全性・セキュリティ対策： 外部からの盗難や不正アクセスに対する軽減策（サイバーセキュリティ）、悪用・誤用（misuse）に対する軽減策（モデルの安全装置）だけでなく、今回発生したようなモデル内部からの不正・自律的行動（unauthorized action from the inside、モデル自身の誤動作）に対する軽減策。
- 外部評価者との協働体制： プロバイダーが METR 等の外部評価者（external evaluators）とどのように連携しているか、特にモデルの客観的評価を行うにあたり外部評価者が十分な独立性を保持しているか。
- 市場流通後のモニタリング体制： 不正利用や異常行動を検知するための市場流通後モニタリングの方法。

OpenAI-Hugging Face インシデントに関しては、プロバイダー側から AI 法で義務付けられている重大インシデント報告が AI オフィスへ提出された。また 2026 年 10 月 9 日には、ビルクネン執行副委員長の要請により、AI 法第 68 条に基づき設置された 60 名の専門家から構成される科学パネル（Scientific Panel）が招集され、インシデントの内容および AI 能力の急激な進化速度

[33] ただし、欧州委員会関係者によれば、欧州委員会の基本方針はモデル排除や罰金の科刑それ自体ではなく、汎用目的 AI モデルのプロバイダーとの建設的な対話を通じて安全・セキュアで信頼できる AI モデルを EU 市場に提供させることにある。

に関する協議が行われた。10 月時点で OpenAI 等に対して取られた執行は上記の 1 までであるが、重大な懸念等がある場合には、2 から先の執行に進むことも可能である。

なお、当社からの「OpenAI の開発中モデルによるオーストラリア政府サーバへの攻撃のように、仮に EU 域内の政府機関や重要インフラ等に EU 域内には市場投入していない開発中のモデル[34]による「意図せぬ」サイバー攻撃があった場合、AI 法の執行の抜け穴となりうるか？」との質問に対しては、基本的にはケースバイケースの評価が必要としつつも、以下の回答であった。

- 既存モデルと派生モデルの技術的連続性：　多くの開発中モデルは市場投入済みモデルのファインチューニング版であり、AI 法による規制の範囲内に含まれ得る。言い換えれば、単一の大規模な事前学習から分岐した複数のバージョン（社内利用バージョンを含む）が存在する場合、それらが市場投入済みモデルと同一の事前学習に基づいている限り、法的には「同一のモデル」として扱われる。これにより、バージョン変更による法規制の潜脱を防ぐ設計となっている。
- ライフサイクル全体のリスク管理義務：　プロバイダーは（市場投入したモデルについては）市場投入後のみならず、市場投入前の開発フェーズを含むライフサイクル全体を通じてシステミックリスクを評価・管理する法的義務を負っている。ライフサイクルは大規模な事前学習の開始時点から起算される。
- ガバナンス不備の判定：　開発段階等で発生するインシデントは、プロバイダーが基本的なセキュリティやアライメントを統制できていない兆候とみなされ、AI 法や実践規範への組織的な非遵守を示す証拠として扱われうる。

4．日本の取るべき方策

現在、日本にとって AI を巡る喫緊の課題は 2 つある。1 つは「フロンティアモデルへのアクセスの確保」であり、もう 1 つは AI を悪用したサイバー攻撃や、AI 自身の誤動作（制御の喪失）によるサイバー攻撃等の「重大な AI リスクへの対処」である。

前者については、2.6 節に記載したように、日本や EU、英国を含む「AI 中堅国」は米国フロンティアモデルへのアクセス確保が絶対条件である。それは、フロンティアモデルへのアクセス権を失うと AI を用いた脆弱性の発見ができずにサイバーセキュリティ上で大きな脅威となってしまうからであり、この点は後者とも関連している。アクセス権を維持し続けるためには、米国政府／米国企業に対する影響力・交渉力（Give and Take の Give、交渉カード）を持ち続けなければならず、米国に対してどのような Give を提供できるかが重要である。この米国に対する Give の部分としては、いちおう日本は半導体製造装置や半導体ウェハー、イメージセンサーその他ロボットの部品等で強みを持つ、すなわち AI バリューチェーンの一定分野において戦略的自律性を維持するに足る強みを持っている。しかし米国との交渉において実質的には、米国での計算資

[34] AI 法は第 2 条第 8 項で「本規則は、市場投入またはサービス提供される前の AI システムまたは AI モデルに関するいかなる研究、テスト、または開発活動にも適用されない」と規定している。

源インフラ建設に対する巨額の投資や、米国フロンティアモデルや GPU（国内データセンター用）等への巨額の需要、すなわちジャパンマネーの大盤振る舞いを出動せざるを得ないかもしれない。

また、米国への影響力・交渉力を強化するために AI 中堅国は一致団結して協力し合う必要がある。日・EU・英国間でできる協力は G7 広島 AI プロセス、AISI 国際ネットワーク、AI・ロボティクス（フィジカルAI）分野での共同研究開発／バリューチェーンの相互補完、個人データ保護の相互十分性認定など限られるかもしれないが、AI 法／AI 推進法等のハードローを通じた規制協力も今後は重要となろう。

後者については、日本においてはこれまで AI 政策において産業発展のための「推進」が強調され、安全性確保のための「規制」が軽視されてきた経緯がある。しかし、悪意のある人間による AI を利用したサイバー攻撃や生物兵器製造のリスク（AI 悪用リスク）、AI 自身の誤動作によるサイバー攻撃などのミスアライメントのリスク（AI 誤動作リスク）が昨年の AI 推進法の制定当時よりも顕在化している現在[35]、特に海外フロンティアモデルに対するセーフガードとしての規制強化（透明性義務や有事の強制調査・サービス制限要求など）は避けて通れないだろう。

まとめると、日本政府は AI 推進法の改正を通じて規制強化を行い、その際に規制対象となる AI モデルに EU の AI 法と同様の閾値を設定することにより、米中フロンティア AI 企業に対して「情報提供要求」「インシデント報告」「アクセス要求／是正要求／制限・撤退要求」「違反時の罰金」を可能とすることで、国民の AI 懸念・不安の払拭に努めるとともに、逆に規制強化によってフロンティアモデルへのアクセスを失わないために、EU との協力関係を強めるべきである。実際に米国モデルに対する排除要求等を行うかは別として、規制上のカードとして持つことによって、EU と協調して 2 大市場として米国に対する交渉力を発揮することにもつながるだろう。

また、上記の法改正は AI リスクに対する制度的対策であるが、これらだけでは不十分であり、技術的対策、とりわけアライメント対策が必要となることは言うまでもない。ただし、どのようなアライメント対策が有効なのか、フロンティア AI 企業や AI 安全性機関においてもまだ十分に分かっていないのが現状であろう。Anthropic の政策提言文書「AI の指数関数的発展に対する政策」[36]では、制御の喪失リスクに対する技術的対策の方向性として、「開発者の制御外で動作する AI システムを検知して対応する能力の開発、およびそのような AI システムを封じ込めたり停止させたりするためのインフラの構築」を挙げているが、このようなモニタリング措置やキルスイッチを通じた「封じ込め（containment）」が、現時点で取りうる現実的な技術的対策であるのかもしれない。

[35] 日本国内では特に、民間企業等への AI を利用したサイバー攻撃が相次いでいる。
[36] https://www.anthropic.com/policy-on-the-ai-exponential。

参考文献

- The White House「Promoting Advanced Artificial Intelligence Innovation and Security（高度なAIの革新とセキュリティの促進）」（大統領令 14409）（2026 年 6 月 2 日）（https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/）
- Anthropic「Project Glasswing」（2026 年 4 月 7 日）（https://www.anthropic.com/glasswing）
- Anthropic「Statement on the US government directive to suspend access to Fable 5 and Mythos 5」（2026 年 6 月 12 日）（https://www.anthropic.com/news/fable-mythos-access）
- Anthropic「When AI build itself（AI が自分自身を構築する時）」（2026 年 6 月 4 日）（https://www.anthropic.com/institute/recursive-self-improvement）
- Anthropic「Policy on the AI Exponential（AI の指数関数的発展に対する政策）」（2026 年 6 月 10 日）（https://www.anthropic.com/policy-on-the-ai-exponential）
- Dario Amodei「We Must Pace the Frontier（我々はフロンティアモデルの開発速度を適正化しなければならない）」（2026 年 9 月 12 日）（https://darioamodei.com/post/we-must-pace-the-frontier）
- OpenAI「OpenAI と Hugging Face、モデル評価中のセキュリティインシデント対応で連携」（2026 年 7 月 21 日）（https://openai.com/ja-JP/index/hugging-face-model-evaluation-security-incident/）
- 欧州連合（EU）「Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)」（AI 法／AI 規則）（2024 年 7 月 12 日 EU 官報掲載）（https://eur-lex.europa.eu/eli/reg/2024/1689/oj）
- 欧州委員会「Code of Practice for General-Purpose AI Models: Safety and Security Chapter」（汎用目的 AI モデル実践規範の安全性・セキュリティ章）（2025 年 7 月 10 日）（https://ec.europa.eu/newsroom/dae/redirection/document/118119）
- 欧州委員会「Action Plan on Cybersecurity and Artificial Intelligence（サイバーセキュリティと AI に関する行動計画）」（2026 年 7 月 7 日）（https://digital-strategy.ec.europa.eu/en/library/eu-action-plan-cybersecurity-and-artificial-intelligence）
- 英国 AISI（UK AISI）「Incident Report: unsanctioned agent behaviour during cyber testing」（2026 年 8 月 4 日）（https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing）
- Yoshua Bengio 他「International AI Safety Report 2026（国際 AI 安全性レポート 2026）」（2026 年 2 月）（https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026）
- 小泉雄介「米欧政府の AI 安全性へのコミットメントと国際 AI 安全性レポート：日本は AI 安全面でプレゼンスを発揮できるのか」（2026 年 5 月）（https://www.i-ise.com/jp/information/report/pdf/rep_it_202603a_2606.pdf）
- 小泉雄介「EU の AI 法（AI 規則）改正の概要」（2026 年 9 月 18 日）（https://www.i-ise.com/jp/information/media/2026/260918.pdf）

以　上