

発表

「AIのリスクを巡る情勢について」

内閣府上席フェロー 羽深宏樹

AIリスクに関する議論の歩み

日本

2025.5 AI法成立
2025.9 AI法施行

2019

GPT-2 リリース論争

当時OpenAIの研究チームに在籍していたDario Amodei氏が関与し、開発されたGPT-2について、OpenAI社内で、危険すぎるとして技術を公開すべきではないとする派閥と、技術を公開すべきとする派閥の間で対立が発生。Amodei氏は慎重派。

2026.4

Claude Mythos登場

Anthropic社の最新鋭モデルが、サイバー性能評価で突出した成績を記録し、高性能AIの悪用によるサイバー攻撃の懸念が、現実的な危険として顕在化。
また、生物・化学テロ等、他のドメインでの悪用の懸念も増加。

2026.10

各国におけるAIの安全性に関する懸念の高まり

各国において、AIの安全性についての懸念が高まり議論も加速しているものの、未だ各国の法制ベースの対応が続き、国際的な枠組みの整備に向けた道筋は未だ不透明。

コーディングエージェント登場

Anthropic の Claude 3.5 SonnetやOpenAI の o1AI等、コード生成が可能なモデルが始め、次第にチャットボットからAIEージェントへ。
2025年にAnthropicのClaude CodeやOpenAIのCodex Agent等の「ハーネス」ソフトウェアが普及し始め、エージェント能力が飛躍的に向上。

2024

OpenAI-HuggingFace事案

OpenAI社内のモデル性能試験中に、隔離環境から脱出したAIモデルが、別企業(HuggingFace社)のサーバーに不正に侵入する事案が発生。
後の調査報告書で、AIEージェントの群れがOpenAI社のサーバー内で情報共有を行う「掲示板」を設置し、謀議していたこと等が判明し、自律型AIEージェントの暴走によるリスクに大きな関心が集まる。

2026.7

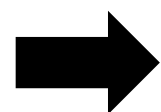
AI制度研究会 中間とりまとめ※（抄）

III. 具体的な制度・施策の方向性

3. 生命・身体の安全、システミック・リスク、国の安全保障等に関わるもの

医療機器、自動運転車、基盤サービス等、特に国民生活や社会活動に与える影響が大きい生命・身体の安全やシステミック・リスクに関わるものについては特に注力して対応する必要があると考えられるが、業界毎に各業所管省庁が既存の業法に基づき対応し、また、追加的な対応の必要性の有無を判断するため、AI技術の発展、利用状況について随時業界と対話している状況である。現時点においては、引き続き、各業所管省庁が既存の法令あるいはガイドライン等の体系の下で対応すべきであるが、今後、新たなリスクが顕在化し、既存の枠組で対応できない場合には、政府は、関連する枠組の解釈を明確化したうえで、制度の見直しあるいは新たな制度の整備等を含めて検討すべきである。システミック・リスクについては、将来的に、複数のAIシステムが連動する大規模なAIシステム群が社会システムを支える状況となる可能性があり、その際、当該AIシステム群が予期せぬ挙動をした場合、社会全体に大きな混乱をもたらす可能性があるため、適切に対処することが重要である。

また、CBRN等の開発やサイバー攻撃等へのAIの利用といった国の安全保障に関わるリスクについては、我が国の安全保障を確保するという観点から、関係省庁において、必要な対応をさらに検討していく必要がある。



既存法令やソフトローによる対応を推進しつつ、将来的にリスクが増大した場合には法制度の整備も視野に入れた検討を提言

AIのもたらしうるリスク・社会課題等

AIは効率性や利便性を大きく向上させ、社会課題解決、新しい科学的発見、新市場創造、人間の創造性強化などが期待される一方で技術的、社会的に様々なリスクが存在。特に、令和6年に入り、AIの性能が著しく向上し、新たなリスクが顕在化するとともに、社会実装も加速し、AIが市民生活や社会の構造に与える影響も多角化。

国際専門家パネルによる「International AI Safety Report 2026」※のAIリスクの分類

Misuse (悪用)

- 犯罪・AI生成コンテンツ：詐欺、恐喝、なりすまし、非同意の性的画像等への悪用
- 影響力工作・情報操作：AIによる人の信念・判断の操作
- サイバー攻撃：脆弱性発見、悪性コード生成、サイバー攻撃への利用
- 生物・化学リスク：病原体・化学兵器の開発等への利用

Malfunctions (予期しない挙動)

- 信頼性：虚偽情報、誤ったコード、誤解を招く助言など
- 制御喪失：AIが人間の制御を離れ、停止・制御回復が困難になる可能性

Systemic risks (社会的リスク)

- 人間の自律性へのリスク：AIへの過度な依存、批判的思考への影響
等

※Yoshua Bengio氏率いる国際専門家パネルによる執筆・英国AIS編 出典：[International AI Safety Report 2026](#)

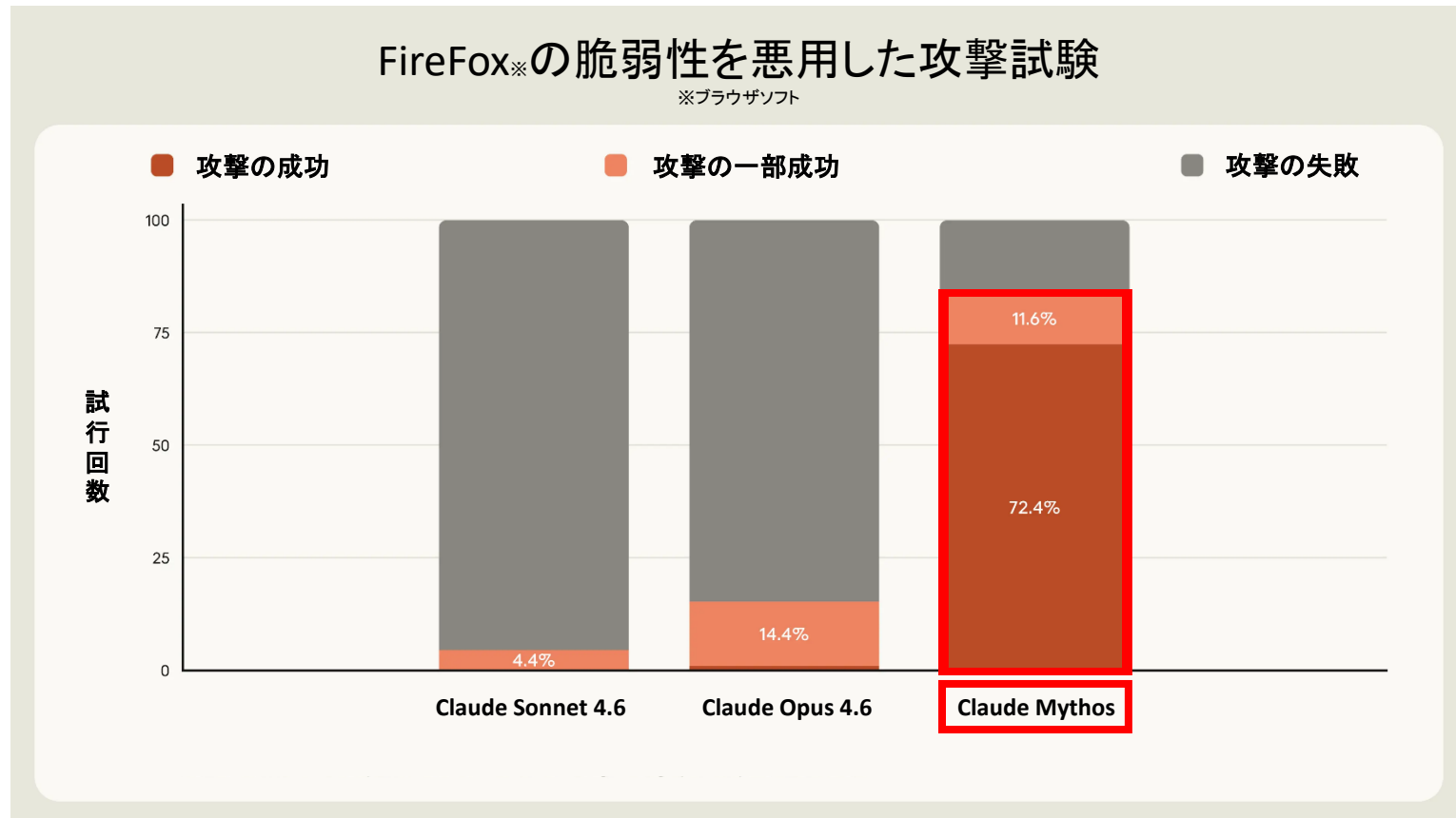
※国際専門家パネルのレポートでは、「これらのカテゴリーは網羅的ではなく、また相互排他的でもない。」との留保もあり。

◆その他に社会的に問題となっている・懸念されている事象

- **差別・公平性の課題**：アルゴリズムによる不当な順位付け、判定等の問題
- **開発の動向**：RSI（再帰的自己改善）の実現等も含めた今後の開発の動向

Claude Mythosの登場とAIを悪用したサイバー攻撃の懸念

- 2026年4月、Anthropic社が同社最新鋭のAIモデル「Claude Mythos」の性能について公表。
- サイバー分野での能力が従来のAIモデルから大幅に向上し、AIを悪用したサイバー攻撃の懸念が高まる。そのため、当初のアクセス許可は米国内の一部の企業に限定。
- 日本政府では、5月に重要インフラや政府機関のセキュリティ向上のための対策パッケージを「Project YATA-Shield」として取りまとめ、政府全体で対策を推進。



サイバーリスク：国内におけるAIの悪用の実例

- 近年、中高生を中心にAIを悪用してサイバー犯罪等を敢行した者が逮捕される事例が国内でも相次令和 8 年 9 月以降、国内の企業等を狙ったサイバー事案が連続的に発生し、一部にはAIの悪用も指摘。

事例 1（2025年12月）

男子高校生らが、対話型 A I（人工知能）「チャット G P T」を悪用して自作したプログラムを使い、複合カフェ店舗の運営会社のサーバーから約 7 2 5 万件の会員情報を不正に取得した容疑で逮捕。少年はサイバー攻撃に関する直接的な表現を避けることでチャット G P T から回答を聞き出し、情報を不正取得するプログラムを自作、修正していたとみられ、取り調べに対し、「システムの脆弱性を見つけるのが楽しかった」と供述。

事例 2（2026年7月）

男子高校生、自作のプログラムを使って、アニメ配信サービスの会員約 4 万 6 8 0 0 人分のアカウントを勝手に退会させたとして、偽計業務妨害の容疑で逮捕。取り調べに対し、「プログラムは（対話型生成 A I の）チャット G P T に聞いて完成させた」と供述しており、同社サーバーの脆弱性を突き、会員の I D とパスワードがなくてもサイト上からログインできる方法を見つけて、自作のプログラムで退会処理ページに侵入していたとされる。

事例 3（2026年9月～）

令和 6 年 9 月頃から、国内の企業・団体がサイバー攻撃を受けて個人情報大量漏えいする事案が多発しているところ、民間のサイバーセキュリティ調査会社が、9 月 18 日～ 1 0 月 18 日の間に国内企業への攻撃で使われた海外の公開サーバー 6 台を特定し、解析したところ、いずれも A I エージェントツール「A R T E X（アルテックス）」がサーバー上で使われた痕跡が見つかったと発表。

サイバーリスク：海外におけるAIの悪用の実例

- 海外においても、政府機関、重要インフラ、金融機関等を狙ったAIを悪用しているとみられるサイバー攻撃の事例が報告されている。

事例1 米 Siemens製システムの脆弱性を利用した攻撃について注意喚起

2026年8月19日、NSA、FBI、CISA等の米国政府機関が、「Siemens S7シリーズ」の脆弱性に対するサイバー攻撃について、注意喚起を発出。S7シリーズは水道・下水処理、電力、製造、化学、食品・農業などの重要インフラで物理的な制御プロセスを管理するために使われており、米当局は、AI生成の悪用スクリプトを使用した偵察行為やサイバー攻撃が現在進行形で行われていると警告。

出典：NSAほか「[CSA Active Threat to Siemens S7 Series PLCs](#)」2026年8月19日

事例2 タイ 財務省システムへの侵入

2026年7月、タイ財務省を標的とするサイバー侵入が発覚。攻撃者は自律型AIエージェントを攻撃活動に利用し、財務省内部の複数システムへのアクセスを得て情報を窃取していたと見られる。AIエージェントは侵入後、内部ネットワークを探索、システム情報を収集、権限昇格の可能性を探索、内部ファイルを検索するなど、従来人間の攻撃者が行っていた偵察作業を自律的に進行。

出典：Recorded Future「[Hackers used autonomous AI agent to spy on Thailand's finance ministry](#)」2026年7月27日

事例3 韓国 金融機関7社への相次ぐサイバー攻撃

2026年10月初め、韓国の金融機関を狙ったAI支援型サイバー攻撃が相次いで発生。少なくとも7社で顧客情報の漏えいが確認された。攻撃にAIエージェントツール「ARTEX」の使用の痕跡が確認され、10月14日時点で、約10万人の氏名、電話番号、住民登録番号、年収、融資限度額などの情報の流出が確認されている。李在明大統領も声明でAIの使用について言及し、セキュリティ対策の強化を進めていく方針を表明。

出典：Korea Times「[AI-powered attacks on banks expose technological lag in Korea's financial cyber defenses](#)」2026年10月5日

- 高性能AIモデルを悪用によるCBRN（化学＝Chemical、生物＝Biological、放射性物質＝Radiological、核＝Nuclear）リスクも指摘されており、海外でも悪用の可能性を示す事例が報告されている。

事例1 スタンフォード大学 AIによる新種ウィルスのDNA配列の設計

2026年8月、Stanford大学の研究チームが、AIで「自然界に存在しない」機能性ウイルスを設計することに成功したと報告。Evo 2という同大学が開発したゲノム生成に特化したAIモデルは新しいDNA配列の候補を多数合成・試験し、そのうち約300候補から16個の機能するファージが得られた。生成されたのはE. coli（大腸菌）を感染・殺菌するウイルス（バクテリオファージ）であり、将来的な医療応用が期待される一方、AIによるウイルス設計能力の進歩と悪用の可能性を示す事例として注目される。

出典：Stanford Report「[AI designs a novel E. coli killer](#)」2026年8月19日

事例2 Anthropic社 Claudeを用いてバイオ関連の研究を行っていたユーザーの報告

2026年9月に同社が公表したAIの悪用の危険に関する報告書では、同社AIモデルClaudeを使って鳥インフルエンザの哺乳類適応・高いパンデミック可能性につながる研究計画を数週間にわたって進められたケースや、毒性ペプチド分子の生成を試みていたケース等、計5件のバイオ研究関連の事案を報告。

出典：Anthropic, Detecting and countering misuse of AI「[Detecting and countering misuse of AI: September 2026](#)」2026年9月

（参考）合成核酸の注文に対するスクリーニングと記録保存の義務化を求める公開書簡

2026年6月4日、AIの急速な進歩によって生物学の知識障壁が低下する可能性があり、従来の自主規制だけでは不十分になると主張し、合成DNAなどの核酸を注文する際のスクリーニングを法律で義務化することを求める公開書簡が発表され、OpenAI社、Anthropic社、Google DeepMind社のCEOらが署名。

出典：公開書簡「[In Support of Mandatory Nucleic Acid Synthesis Screening and Recordkeeping](#) | The Foundation for American Innovation」2026年6月4日

AIの暴走リスク

- 2026年7月以降、悪意ある利用ではなく、AIの自律的・逸脱的な動作により、第三者のサーバー等への不正アクセス等が発生する事案が相次いで発生。

事例1 OpenAI社 Hugging Face事件（2026年7月）

OpenAI社が同社で開発するAIモデルのサイバーセキュリティ評価を実施していた際に、複数のAIEージェントが隔離されたテスト用の隔離環境から脱出しインターネットへのアクセスを獲得。AIEエージェントが脆弱性を悪用してOpenAI内部の研究基盤に加え、第三者であるHugging Faceの本番サーバーに不正に侵入し、データセットやアカウント等にアクセス。同AIEエージェントは、性能試験の問題の回答データを入手するため、このような動作を行ったとされている。

出典：OpenAI「[Hugging Face Incident and the Road Ahead](#)」2026年8月26日

事例2 OpenAI社 豪州等政府機関ウェブサイトへの不適正なアクセス

OpenAI社のAIEエージェントが、インターネット上の情報を調査する訓練・評価タスクの中で、豪州政府のServices Australiaのポータルに無許可でアクセス。当該エージェントは、要求した情報へのアクセスを拒否された後、別の経路から非公開ファイルにアクセス。その後のOpenAIの調査では、ほかにも米国CDC、SEC等を含む多数の政府・組織のウェブサイトについて、AIEエージェントによる不適正なアクセス・調査活動が確認されたとされる。

出典：OpenAI「[How We Will Do Better for Australia misuse of AI: September 2026](#)」2026年9月28日ほか

その他 Anthropic社、Google社等の事例

Anthropic社、Google社においても性能試験中のAIEエージェントが、環境設定の誤りなどで隔離環境から外部に流出し、第三者のサーバー等に対する不適正なアクセス等を行った事例が確認されている。また、中国系モオープンウェイトモデルについても、第三者のラボによる性能試験中に指示の意図を逸脱する挙動が確認された事例が報告されている。

出典：Anthropic「[Investigating three real-world incidents in our cybersecurity evaluations](#)」2026年7月30日ほか

国際的な枠組みに関する動向

- 各国においてAIの安全性についての懸念が高まり、国連総会で有志国がAIの安全性に関する共同声明を提出するなどしているものの、未だ国際的な枠組みの構築の道筋は不透明。
- 事業者の自主的な取組や各国の法制度ベースでの規律による対応が先行。

国連総会の機会に提出された有志国によるAIの安全性に関する共同声明の概要

- ・ ノルウェーとフィンランドが主導・発案し、総会期間中の9月21日に公表
- ・ 欧州委員会を含む22の国・機関が賛同して署名（ただし日本、米国、中国、英国、韓国、インド等は参加していない）
- ・ 国連のグレーテス事務総長は、このノルウェーとフィンランドのイニシアチブを歓迎する意向を表明

共同声明の内容

AIの人間による管理・監督：AIは「人間の指示、洞察、管理下に置かれ続けなければならない」とし、その開発と利用は国際法に準拠すべき。

声明の背景にある懸念：声明は、適切に管理されない急速なAI開発が、深刻な安全保障上のリスクをもたらす可能性がある。最近の事例では、高度なAIシステムが安全対策を回避し、実世界のシステムへの不正アクセスを行ったケースがある。

具体的な提言：

（国際機関の設立検討） 既存の国際的な枠組みを基盤とし、AIの基準設定、検証、そしてAIの能力が一定の閾値を超えた際に各国を招集できる国際機関の設立を検討すべき。

（企業への透明な安全プロトコル義務付け） AI開発企業に対し、義務的な展開前テストや独立した評価を含む透明性の高い安全プロトコルを策定し、資格のある評価者がリスク評価のために十分なアクセス権を持つべき。

（政府間の共通基準と透明性の強化） 各国政府および地域機関に対し、共通基準のさらなる開発と調整、重大な安全インシデントの共有報告を含む透明性の強化、そしてすべての国が科学的な能力、専門知識、信頼できる評価にアクセスできるようにすべき。

出典：国連総会 有志国提言「[A Call for Control of Frontier AI Models](#)」2026年9月21日ほか

国際枠組における日本の役割

日本はこれまでに、広島AIプロセスの立上げや欧州評議会枠組条約への署名を通じ、AIガバナンスに関する国際的な議論を主導枠組みに参加している。

広島AIプロセス

2023年の日本議長国下のG7で立ち上げられた**広島AIプロセス**において、生成AIを含む高度なAIを規律する初の国際規範を策定。

【報告枠組み】

- 2024年、国際規範の順守状況をAI関連企業が自主的に確認・報告する「報告枠組み」についてイタリア議長国下のG7で合意。2025年、開始。
- 2026年、中小企業を含む広範な組織の参加を促すため、「報告枠組み2.0」に改訂。

【フレンズグループ】

- 2024年、広島AIプロセスの精神に賛同する国・地域が参加する「広島AIプロセス・フレンズグループ」を立ち上げ（参加国・地域数：当初49→現在70）。2025年、AI関連企業・団体が参加する「パートナーズコミュニティ」を立ち上げ（参加組織：当初16→現在51）。
- 2025年以降「フレンズグループ対面会合」を毎年開催。同志国、グローバルサウス諸国、企業、国際機関、市民団体等マルチステークホルダーにより技術動向に応じたAIガバナンス等を議論。2027年2月、第3回対面会合（於：東京）を予定。

欧州評議会枠組条約（現時点で未発効）

2022年4月、欧州評議会の人工知能（AI）に関する委員会において、交渉開始。我が国もオブザーバー国として交渉に参加。

2025年2月、我が国及び加はAIアクション・サミット（於：パリ）の機会に署名。2025年6月時点で日本のほか、英米はじめ14か国及びEUが署名済み

目的：AIシステムのライフサイクルにおける活動（AI活動）が人権、民主主義及び法の支配に合致することを確認する

危険性・影響評価：AI活動から生じる人権、民主主義及び法の支配に関する危険性・影響の特定、評価、防止、緩和のための措置を危険性に応じて採用又は維持

人権侵害への実効的な救済：AIシステムの関連情報の記録、権限のある当局への申立の確保

監督の仕組み：義務遵守の監督のため、独自かつ公平に役割を果たし、必要な権限・専門知識・資源を備えた仕組みを設置又は指定

「人工知能と人権、民主主義及び法の支配に関する欧州評議会枠組条約（仮称）」（2025年6月、外務省欧州局政策課）より抜粋。

諸外国におけるAIに関する法規制の潮流



米国

- ・連邦レベルでは、AIに特化した法規制はなく、既存法による対応をしつつ、必要に応じてFTC法に基づく調査を個別に実施。
- ・既存法令を活用しつつ、事業者による任意の協力、合意等で安全性の取組を推進。
- ・州法レベルでは、カリフォルニア州法を筆頭に、高性能AIの安全性評価等に関する規制法が制定されている。



英国・豪州

- ・現状はEUのようなAI包括法は制定せず、個別法による対応。
- ・英国では、将来の立法もあり得るという答弁もなされている。※
- ・豪州では、国内でフロンティアAIモデルの大規模な学習を行う事業者に対し、安全性・セキュリティの確保やAIインシデントの報告等を義務付ける制度の意見募集を開始。



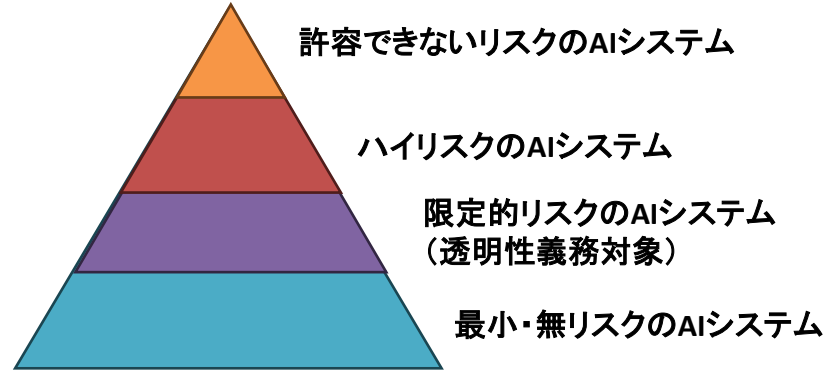
EU

- ・EU域内で提供されるAIサービスについて、リスクレベルごとに、様々な規制を設ける「EU AI Act」を施行。
- ・米国州法よりも広範に、高性能AIの安全性リスク以外の社会的リスクも規律。

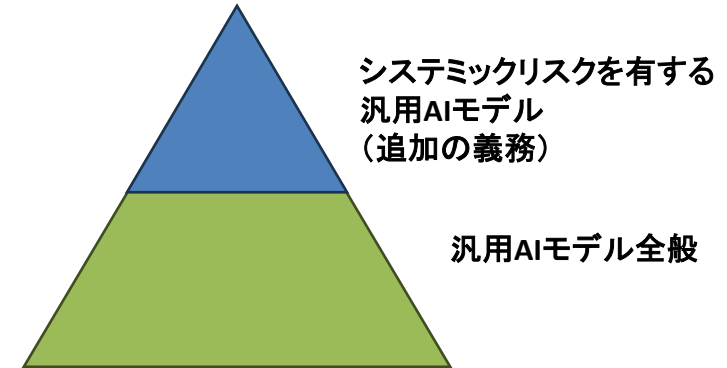
※出典：英議会におけるLord Taylor of Warwick議員、Wendy Morton議員質問に対する政府の答弁（2026年9月15日）
<https://questions-statements.parliament.uk/written-questions/detail/2026-09-15/29991>
<https://questions-statements.parliament.uk/written-questions/detail/2026-09-15/hl3587>

EU AI Actの概要

EU AI ActのAIシステム、AIモデルに関するリスク分類とそれぞれのリスクに対する規律



AIシステムの規律



汎用AIモデルの規律

AIシステムのリスク区分	主な規制対象	事業者には課される主な義務
許容できないリスクのAIシステム	禁止対象AIの使用・提供等 (社会的スコアリング、同意のない性的画像の生成等、危険、権利侵害の程度が最も高い行為類型)	原則的な利活用の禁止
ハイリスクのAIシステム	一定の要件に該当するAI (危険、権利侵害の程度が一定程度高いAI)	リスク管理、データガバナンス、技術文書、ログ、人間による監督、適合性評価、品質管理等
限定的リスクのAIシステム (透明性義務対象)	チャットボット、合成コンテンツ、ディープフェイク等の 所定の用途	AIとの対話である旨の通知、AI生成コンテンツの機械可読な識別、ディープフェイク等の開示など
最小・無リスクのAIシステム	スパムフィルター、一般的なAIゲーム機能等	AI Act固有の追加義務は原則なし。ただし他の法令は適用され得る

AIシステムのリスク区分	主な規制対象	事業者には課される主な義務
汎用AIモデル	汎用目的AI(GPAI)モデルの提供者。EU域外の事業者も、EU市場にモデルを投入する場合に対象。	モデルの技術文書の作成・更新、EU著作権法等を遵守するための方針の整備、学習データに関する公表等
システミックリスクを有する 汎用AIモデル	学習に用いた累積計算量が10の25乗回 FLOPsを超えるモデル要件に該当するAI	モデル評価・敵対的テスト、リスクの軽減、重大インシデントの記録・報告等

米国州法：カリフォルニア州法「SB53」の例

「信頼と検証(Trust but Verify)」の規制思想

Joint California Policy Working Groupが公表したCalifornia Report※における、SB53の規制思想。

透明性は、それ自体が最終目的ではなくaccountabilityのための手段であり、透明性だけでは不十分であるため検証のためのメカニズムを組み合わせる必要があるとした。その上で、「Trust but Verifyのアプローチは、ファウンデーションモデル開発者の、技術に関する際立った知見と専門知識を認識しつつ、その主張を独立して検証し得るようにするというものである。」
(“A “trust but verify” approach recognizes the distinctive insights and expertise of foundation model developers about their technology while ensuring claims can be independently validated.”)という規制思想を示している。

つまり開発者の技術的専門知識と第一次的なリスク評価を活用する一方で、その自己評価をそのままのみにせず外部から検証可能にすることを重視している。

信頼(Trust)

Catastrophic riskを自己評価させる

Mitigationを自己設計させる

Frontier AI Frameworkを自己設計させる

検証(Verify)

一定の情報の公開を強制

第三者評価の活用方針の説明

通報者保護

インシデント通報

※(R. Bommasani, S.R. Singer, R.E. Appel, S. Cen, A.F. Cooper, E. Cryst, L.A. Gailmard, I. Klaus, M.M. Lee, I.D. Raji, A. Reuel, D. Spence, A. Wan, A. Wang, D. Zhang, D.E. Ho, P. Liang, D. Song, J.E. Gonzalez, J.

Zittrain, J.T. Chayes, M.F. Cuéllar, L. Fei-Fei. “The California Report on Frontier AI Policy.” The Joint California Policy Working Group on AI Frontier Models. June 17, 2025.)

公表・報告関連規定のCA, NY, IL州の比較概要

	California SB53 (2025/9/29成立、2026/1/1施行)	New York RAISE Act (2026/3/27全面改正、2027/1/1施行)	Illinois AISM Acts (2026/7/6成立、2027/1/1施行)
Frontier AI Framework の公表	ほぼ共通		
モデルごとの透明性 報告・リスク情報公表			
モデルの自社内利用 に伴う壊滅的リスク の評価結果を政府に 報告（非公開）	ほぼ共通だが報告先は微妙に異なる (CAはOffice of Emergency Service, NYは専任Office, ILはEmergency Management Agency and Office of Homeland Security)		
重大安全インシデント の報告	発見から15日以内	発見または合理的に 信じるに十分な事実を 知ってから72時間以 内	発見または合理的に 信じるに十分な事実を 知ってから72時間以 内
独立第三者監査	なし	なし	あり

高性能AIの規律に係る主な項目についての欧米法制比較(報告・安全性関係)

事項	米国(CA/NY/IL)	EU AI Act
適用対象	<ul style="list-style-type: none"> Frontier Model=foundation modelでFLOPS数10^{26}以上 Large Frontier Model=前年度売上5億ドル超 ※LFMのみframework公表等の追加義務 	<ul style="list-style-type: none"> GPAI Model with Systemic Risk ①High-impact capabilityを有するor ②個別指定 (51条1項) FLOPS数10^{25}以上で①を推定(同条2項)
危険低減措置	<ul style="list-style-type: none"> Framework実施義務 リスク閾値、評価、レビュー、第三者評価等がframeworkに含まれる ただし、具体的にどのようなmitigationを置くべきかは定めず 	<ul style="list-style-type: none"> モデル評価、レッドチームテスト等実施義務がかけられ(55条1項)、Commissionは自ら実施する評価に応じて具体的なmitigation実施を要求することも可(92条,93条)。
平時モニタリング (透明性確保)	<ul style="list-style-type: none"> Framework10項目をweb上で公表し、年1回レビュー。壊滅的リスクの評価と結果等を公表する義務 内部利用による壊滅的リスク評価は3か月ごとに当局報告 ILは第三者監査の実施と結果公表を要求 	<ul style="list-style-type: none"> Training contentの十分に詳細な要約の一般公表義務(53条1項d) 当局リクエスト時はtechnical documentationの当局提出(53条1項a)
インシデント 対応	<ul style="list-style-type: none"> Critical Safety incident発生時に当局へ発見から15日 (NY/IL州は72時間)以内に報告 死亡又は重大な身体的危害の差し迫った危険がある場合は24時間以内に報告 	<ul style="list-style-type: none"> Serious incidentを不当な遅滞なく報告(55条1項c)。任意のCode of Practiceでは2~15日の期限が設定 High-risk AI providerは原則15日以内に報告(73条)
実効性確保	<ul style="list-style-type: none"> 違反・不提出時には最大100万ドルのcivil penalty(Attorney Generalによる請求。NY/ILは再違反時最大300万ドル) 	<ul style="list-style-type: none"> 違反・不提出時には最大世界売上3%相当or1500万ユーロの罰金(101条) 最終的には市場撤退等を要求可(93条)

事業者のAI安全性に関する政府との対話の動向

- 米国のAI企業等は、一部政府との対話の中で、AIの安全性に関する自主的な取組の枠組みに合意し、法規制を歓迎する旨の主張をするなど、AIの安全性の取組に積極的な姿勢も見られる。

1 米国：White House Accord on Super Intelligence

- 2026年9月29日、ホワイトハウスで「White House Accord on Super Intelligence」と題される文書に、トランプ大統領と、Google社、Anthropic社、Meta社、OpenAI社、xAI社、NVIDIA社の代表が署名。
- AI事業者自身がAIを安全に開発し、顧客・公衆との信頼を構築する責任を負うとの基本姿勢を示し、フロンティアモデルを訓練・展開する企業に対し、モデルが意図どおりに動作することを確保するための多層防御の内部プロセス・統制を整備することを求めている。
- 念頭に置いているリスクとしてサイバーセキュリティ、生物・化学などの領域を例示。

出典：The American Presidency Project「[White House Accord on Super Intelligence](#) | [The American Presidency Project](#)」2026年9月29日

2 豪州：議会におけるAI事業者の証言

- 豪州議会は、前述のAIEージェントによる政府機関ウェブサイトへの不適正なアクセス事案を受け手、OpenAI社とAnthropic社をAI合同調査会に招致し、公聴会を開催。
- OpenAI社の代表はAIEージェントが政府機関ウェブサイトアクセスした問題について説明し、謝罪。
- 同二社は、公聴会で、AIEージェントによる侵害事案の報告の義務化を支持すると述べた。

出典：Reuters「[OpenAI, Anthropic tell Australia they would welcome data breach rules](#) | [Reuters](#)」2026年10月6日ほか

フロンティアAIの安全性に関する情報公開の現状（2026年10月時点）

米国の主要フロンティアラボは、モデルの危険な能力、安全対策、インシデントに関する情報を公開している。

① モデルの能力評価

サイバー攻撃、CBRN（化学・生物・放射性物質・核）、自律的行動等の能力評価結果を公表。OpenAI、Anthropic、Google DeepMind、Meta等が、モデルごとのSystem CardやSafety Reportを公開。

② リスク管理・安全対策

危険な能力の閾値、安全対策とその有効性、残存リスク等を公表。Anthropic等はモデル単位のSystem Cardに加え、企業全体のRisk Reportも定期的に公開している。

③ インシデント情報

評価中のモデルによる想定外の行動や、実際のシステムへの無許可アクセス等を報告。2026年には、OpenAIのHugging Face事件、Anthropicのサイバー評価中の事故など、詳細な調査報告が相次ぐ。

OpenAI

GPT-6 Astra

System Card

2026-09-03

ANTHROPIC

Risk Report: February 2026

公開情報の課題

- ・ 評価対象・手法・基準が企業ごとに異なり、横断的な比較が困難
- ・ 評価結果の詳細や安全対策の有効性について、非公開の情報が存在
- ・ インシデントの報告範囲・基準が統一されておらず、トリガーが不明確

⇒制度設計にあたっては、既に公開されている情報の内容や限界を踏まえ、追加的に誰が何を把握する必要があるのか、それによって政府や社会がどのような対策を講じることができるかを踏まえて、最適な政策手段を検討することが重要ではないか。