IISE 調査研究レポート（No.15）

「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」

（2025 年度「AI 推進およびプライバシーを巡る社会課題と政策動向に関する調査研究」報告書に基づき作成）

2026 年 6 月

国際社会経済研究所　主幹研究員　小泉 雄介

１．米欧政府における AI 安全性へのコミットメント

筆者は 2025 年 9 月のレポート「米国 AI 行動計画と日米欧 AI 政策比較」[1]において日米欧の AI 政策の比較を行い、「日米欧の 3 極ともに 2024 年までは AI 規制政策が主流であったが、2025 年からは（トランプ大統領による前バイデン政権の政策の否定という強い動機の影響も受けて）AI 推進政策が主流になった」と述べた[2]。ただし、米国や EU の個別の政策を見ると（EU における AI 法の実践規範の策定に象徴されるように）AI の安全面が決してなおざりにされている訳ではなく、むしろ共通的な危機意識の下に政策が準備されているとの指摘も行った。

例えば米国の連邦政府 AI 利用・調達ポリシー（2025 年 4 月）では、「バイデン政権のようなリスク回避型のアプローチを追求しない」と表明しながらも、実際にはハイインパクト AI の分類義務や一連のリスク管理プラクティスの実施義務などでリスク回避型のアプローチが維持されている。米国 AI 行動計画（2025 年 7 月）でも、バイデン政権における AI 事業者規制を見直すことで安全性が度外視され、なりふりかまわぬ AI 推進に舵が切られたという論調の報道が見られたが、実際には AI が米国社会にもたらすリスクに対する安全性への配慮についても、フロンティア AI モデルが CBRN（化学・生物・放射線・核）兵器の製造に使われるリスクを官民で評価することを含め、多くの箇所で規定が設けられている。ただし、バイデン政権との「差異」を演出するためか、「安全性（safety）」という言葉は極力使わず、「セキュリティ」「国家安全保障」という言葉が頻繁に使われている。

2026 年 4 月に Anthropic が公表したクロード・ミュトスがもたらしうるソフトウェアの脆弱性問題と、それに続くプロジェクト・グラスウィング[3]や、米国大統領令「高度な AI の革新とセキュリティの促進」[4]（2026 年 6 月 2 日）の発令（そしてその後の Anthropic の Fable5 等への外国人アクセス停止命令）は、米国政府が決して AI がもたらしうる社会的リスクや、安全性・セキュリティ面での対処を軽視している訳ではないことを改めて浮き彫りにした。また、（各国機関がミュトスへのアクセス権を求めて米国政府や Anthropic に殺到し、日本を含め多くが早期のアクセスを認められなかった中で）英国 AI Security Institute（UK AISI）のみが当初からアクセス権を（米国政府に）許可され、特別な地位にあることも判明した。

---

[1] https://www.i-ise.com/jp/information/report/pdf/rep_it_202603a_2509.pdf。

[2] EU においても、AI Continent Action Plan（2025 年 4 月）、デジタルオムニバス規則案／AI デジタルオムニバス規則案（2025 年 11 月）、クラウド・AI 開発法案（2026 年 6 月）などで AI 推進政策がとられている。

[3] https://www.anthropic.com/glasswing。

[4] https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/。連邦政府は、AI 開発者と協力して自主的な枠組みを確立することで、「対象フロンティアモデル」のリリース前に最大 30 日間、事前にアクセスしレビューすることが可能となる。

2．EU の AI 法汎用目的 AI モデル実践規範と米国の AI 規制州法の比較

従来から米欧は、産業政策面で異なるスタンスを取ることが多かった。すなわち、欧州は事前規制型であり「予防原則」[5]が重視されるのに対し、米国は事後規制型であり「許可不要な技術開発」[6]が重視されてきた。ただし、EU の AI 法の汎用目的 AI モデル実践規範（安全性・セキュリティ章）[7]と米国の AI 包括規制州法（カリフォルニア州[8]、ニューヨーク州）では、「規制対象となる AI モデル（の考え方）」「規制対象事業者の義務」「AI がもたらす重大な社会的リスクの定義」において驚くほど共通した内容となっており、特に「CBRN 兵器の製造」「サイバー攻撃」「制御の喪失」といった重大な AI リスクに対する共通的な危機意識が見て取れる（下表参照）。このような米欧間での奇妙な符合は、一体どこからもたらされたのだろうか。

表 1　EU AI 法実践規範と米国州法（カリフォルニア、ニューヨーク）の比較（筆者作成）

		EU の AI 法 汎用目的 AI モデル実践規範 「安全性・セキュリティ」章 （2025 年 8 月 1 日承認）	米国カリフォルニア州 フロンティア AI 透明性法（TFAIA、SB53） （2025 年 9 月 29 日成立）	米国ニューヨーク州 責任ある AI 安全・教育（RAISE）法 （2025 年 12 月 19 日成立）
規制対象となる事業者		システミックリスク汎用目的 AI モデルのプロバイダー	フロンティアモデルの大規模開発者 （フロンティアモデルの開発者一般にも一部の義務が課されるが罰則は無し）	フロンティアモデルの大規模開発者
規制対象となる AI モデル		システミックリスク汎用目的 AI モデル： 10 の 25 乗を超える浮動小数点演算（FLOP）の累積計算量を使用して学習された汎用目的 AI モデル	フロンティアモデル： 10 の 26 乗を超える整数演算または浮動小数点演算（FLOP）の累積計算量を使用して学習された基盤モデル	フロンティアモデル： 10 の 26 乗を超える計算演算（整数演算や浮動小数点演算（FLOP））を用いて学習された AI モデルで、その計算コストが 1 億ドルを超えるもの または、このようなフロンティアモデルに知識蒸留を適用して作成された AI モデルで、計算コストが 500 万ドルを超えるもの
対象事業者の義務	(a)敵対的テスト・文書化含むモデル評価	○ 安全性・セキュリティフレームワークの作成・実装・遵守・更新・公開（要約版）	○ フロンティア AI フレームワークの作成・実装・遵守・更新・公開	○ 安全性・セキュリティプロトコルの作成・実装・遵守・更新・公開
	(b)システミックリスクの評価と軽減	○ システミックリスクの評価と軽減（安全性軽減策・セキュリティ軽減策）、外部評価者の使用	○ 壊滅的リスクの評価と軽減、外部評価者の使用	○ 重大な危害のリスクの評価と軽減
	(c)重大インシデントの報告	○ 重大インシデントの AI オフィス／国内所轄機関への報告	○ 重大な安全インシデントの州当局への報告	○ 安全インシデントの州司法長官／州当局への報告
	(d)モデルとインフラのサイバーセキュリティ	○ セキュリティ軽減策	○ サイバーセキュリティ対策（未公開のモデル重みを内部・外部者による不正な変更や転送から保護）	○ サイバーセキュリティ保護策
	モデルレポートの作成と提出	○ 安全性とセキュリティに関するモデルレポートの作成、AI オフィスへの提出、公開（要約版）	○ 透明性レポートの作成、公開	×
	内部告発者保護	○	○	×

[5] 「安全性が証明されるまでは新規な研究開発を停止すべき」という考え方。
[6] 「危険性が証明されるまでは研究開発を続けるべき」という考え方。
[7] https://www.i-ise.com/jp/information/media/2025/251118.pdf を参照のこと。
[8] https://www.i-ise.com/jp/information/media/2025/251028.pdf を参照のこと。

違反時の罰則	世界売上総額の 3%または 1500 万ユーロのいずれか高額の方を超えない罰金(fines)	違反 1 件あたり 100 万ドル以下の民事罰(civil penalty)	初回違反に対して 100 万ドル以下、その後の違反に対して 300 万ドル以下の民事罰(civil penalty)
システミックリスク／壊滅的リスク／重大な危害の定義	システミックリスク(systemic risk): 汎用目的 AI モデルのハイインパクトな能力に特有のリスクであって、そのリーチによって、または公衆衛生、安全、パブリックセキュリティ、基本的権利、もしくは社会全体に対する実際もしくは合理的に予見可能な悪影響によって、EU 市場に重大な影響を与え、かつバリューチェーン全体に大規模に伝播しうるもの。(AI 法第 3 条(65)) 指定システミックリスク(specified systemic risks): (1) 化学・生物・放射線・核(CBRN) (2) 制御の喪失 (3) サイバー攻撃 (4) 有害な操作	壊滅的リスク(catastrophic risk): フロンティア開発者によるフロンティアモデルの開発、保管、利用、または展開が、フロンティアモデルが以下のいずれかを行うことを伴う単一インシデントに起因して、50 人を超える死亡もしくは重傷、または 10 億ドルを超える財産の損害もしくは損失に実質的に寄与するという、予見可能かつ重大なリスク。 (A) 化学兵器、生物兵器、放射線兵器、核兵器(CBRN 兵器)の製造またはリリースにおいて専門家レベルの支援を提供すること (B) 意味のある人間による監視、介入、または監督なしに、サイバー攻撃、またはその行為が人間によって行われた場合には殺人、暴行、恐喝、もしくは窃盗(詐欺による窃盗を含む)の犯罪を構成する行為を行うこと (C) フロンティア開発者または利用者の制御を回避すること	重大な危害(critical harm): 大規模開発者によるフロンティアモデルの使用・保管・リリースによって生じた、または実質的に可能となった、100 人以上の死亡もしくは負傷、または金銭または財産に関する権利への少なくとも 10 億ドルの損害であって、以下のいずれかを通じたもの。 (A) 化学兵器、生物兵器、放射線兵器、核兵器(CBRN)の製造・使用 (B) (I) 人間の意味のある介入を伴わない行為、かつ (II) 人間が犯した場合、故意、無謀、もしくは重大な過失を必要とする刑法で規定された犯罪、またはそのような犯罪の教唆や幇助を構成する行為を行う AI モデル

筆者がこれらの法令・規範の背景を調べた結果、それらの策定過程において、ヨシュア・ベンジオ等の著名 AI 研究者が共通に関与していたり、各国の AI 関連有識者が共同著者となっている「国際 AI 安全性レポート」(2025 年 1 月)等の同一の報告書類が共通で参照されていることが分かった。このような過程(マルチステークホルダープロセス)を通じて、AI リスクに関する考え方が米欧の関係者間で(米国の主要 AI 開発企業を含めて)共通認識として確立していき、これら米欧の法令・規範において共通的な用語・指標が用いられ、類似した規制内容となっていった(そして米国の主要 AI 開発企業がこれらの規制内容を受容していった)のではないか。

次節では、国際的に大きな影響力を持つこととなった「国際 AI 安全性レポート」の 2026 年版について概要を記載する。

3. 国際 AI 安全性レポート

(1) 概要

「国際 AI 安全性レポート 2026」[9]は、2025 年 1 月に公表された「国際 AI 安全性レポート」の第二弾であり、2026 年 2 月にインド AI インパクトサミットの開催に合わせて公表された。これらは、2023 年 11 月開催の英国 AI 安全性サミットでの要請に基づき作成されている。

G7 含む 30 か国以上および OECD・EU・国連から指名された専門家諮問委員会を含む、100 名以上の AI 専門家が報告書の作成に関与している。同レポートの「議長」はモントリオール大のヨシュア・ベンジオである。日本からはソニーグループの北野宏明氏が参加しており、トロント大のジェフリー・ヒントンや UC バークレーのスチュアート・ラッセルら著名 AI 研究者もシニアアドバイザーとして参加しているほか、欧州委員会の AI オフィスも参加している。事務局

[9] https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026。

は英国の UK AISI とカナダの Mila - Quebec AI Institute が務めている（米国 CAISI は 2025 年版では共同事務局を務めたが 26 年版では離脱した）。情報・エビデンスに基づいた政策立案を支援する科学的情報を提供することを目的としている。なお同レポートの 2025 年版は前述の通り、EU の AI 法の実践規範や、同時期に制定された米国の AI 包括規制州法（カリフォルニア州 TFAIA、ニューヨーク州 RAISE 法）の策定時に共通で参照されている。

同レポートは AI のリスクと安全性に焦点を当てている。2026 年版は AI リスクのうち、「emerging risks」（AI 能力のフロンティアで発生するリスク）のみを対象としたため、2025 年版における「バイアス」「世界的な AI 研究開発格差」「市場集中と単一障害点」「環境リスク」「プライバシーリスク」「著作権侵害」のリスクについては省略される一方、「人間の自律性に対するリスク」が追加された。リスク一覧については後述する。同レポートはこれらのリスクを特定し、その軽減策を評価している。国連の独立国際 AI 科学パネル等の取組みを補完するものでもある。また、同レポートは汎用目的 AI（のリスクと安全性）を対象にしており、その理由として、汎用目的 AI は近年急速に進歩しているがそのリスクはまだ十分に研究されたり理解されていないことが挙げられている。

同レポートは、「汎用目的 AI は何ができるか」「汎用目的 AI に伴うリスクは何か」「これらのリスクに対する軽減策は何か」という 3 つの問いに対する科学的証拠をまとめている。そして結論として、（当然ながら）「汎用目的 AI のもたらすリスクは大きい」とする。同レポートに関与した専門家たちは汎用目的 AI の能力・リスク・軽減策をめぐる様々な問題について依然として意見の相違を抱えているものの、同レポートは汎用目的 AI とその潜在的なリスクに対する関係者間の共通理解を深めるために不可欠だとしている。

（2）2025 年からの AI 能力の進展

国際 AI 安全性レポート 2026 年版では、2025 年版からの AI 能力や安全対策の進展として、以下の 7 点を挙げている。

- 汎用目的 AI の能力は、特に数学、プログラミング、自律的動作の分野で向上を継続。
- 汎用目的 AI の能力向上は、モデルの初期学習後に適用される技術によってもたらされることが増加。
- AI の普及は急速に進んでいるものの、国・地域によって大きなバラつき。
- AI の科学的能力の向上に伴い、生物兵器開発における悪用への懸念が高まっている。OpenAI と Anthropic は 2025 年、展開前のテストで初心者による生物兵器開発を助長する可能性を排除できなかったため、新モデルのリリース前に追加の安全対策を講じた。
- AI システムが現実世界のサイバー攻撃に利用されているという証拠がさらに明らかに。
- 信頼性の高い展開前の安全性テストの実施はますます困難に。テスト環境と現実環境を区別し、評価の抜け穴を悪用するモデルが増えているため。これは、危険な能力が展開前に見過ごされる可能性があることを意味。
- 業界における安全性ガバナンスへの取り組みは拡大。2025 年には 12 社が「フロンティア AI 安全性フレームワーク」[10]を公開または更新。

[10] AI 開発企業が自社のフロンティア AI のリスクをどのように評価・監視・制御するかを説明する文書。EU の

下表は、主要 AI 開発企業が公開したフロンティア AI 安全性フレームワークの概要である。同レポートには他に Microsoft、NVIDIA、Cohere、xAI、Magic、Naver、G42 のものが掲載されている。これらは AI ソウルサミット（2024 年 5 月）の「フロンティア AI 安全性コミットメント」[11]への署名企業である。これらのフレームワークは CBRN・サイバー攻撃能力・高度な自律的行動に関連したリスクをカバーすることや、脅威モデリング・レッドチーム演習・危険な能力の評価などのリスク管理手法において多くの共通点が見られるが、リスクレベルと閾値の定義・評価の頻度・評価と閾値の間のバッファ・軽減策の包括性（モデル重みを削除するか、開発を一時停止するだけか等）において相違が見られるという。

表 2　主要 AI 開発企業が公開したフロンティア AI 安全性フレームワーク（出典：国際 AI 安全性レポート 2026）

AI 開発企業／フレームワーク名称	カバーするリスク	リスクレベルまたはそれに相当するもの、および関連する安全対策
OpenAI／ Preparedness Framework 2	1. 生物学的・化学的能力 2. サイバーセキュリティ能力 3. AI の自己改善能力	• 高：重大な危害につながる既存の経路を増幅させる可能性（セキュリティ制御と安全対策が必要） • クリティカル：重大な被害につながる前例のない新たな経路が生じる可能性（指定された安全対策およびセキュリティ制御基準がクリティカル基準を満たすまで開発を停止）
Anthropic／ Responsible Scaling Policy 2.2	1. CBRN 兵器 2. 自律型 AI の研究開発 3. サイバー作戦	○AI 安全レベル（ASL） • ASL-1：重大な壊滅的リスクなし • ASL-2：危険な能力の初期兆候（モデルは ASL-2 展開基準とセキュリティ基準を満たす必要） • ASL-3：壊滅的な悪用リスクが大幅に増加（モデルは ASL-3 展開基準および／またはセキュリティ基準を満たす必要） • ASL-4+：将来の分類（未定義）
Google／ Frontier Safety Framework 3.0	1. 悪用 a. CBRN b. サイバー c. 有害な操作 2. 機械学習の研究開発 3. ミスアライメント／道具的理性（※）	• クリティカルな能力レベル：軽減措置（RAND セキュリティレベル 2、3、4（※）に準拠した展開の安全性ケースおよびセキュリティ軽減策）がなく、AI モデルやシステムが重大な危害のリスクを高める可能性がある能力レベル。この能力レベルには、特定の「警告閾値」を含む「早期警告評価」が含まれる ※ RAND セキュリティレベル（SL1～SL5）：RAND Corporation が提唱した、フロンティア AI モデルのモデル重み（学習済みパラメータ）を悪意ある攻撃や盗難から保護するためのセキュリティ基準 ※ 道具的理性：目的そのものの善し悪しは問わず、どうすればその目的を最も効率的・経済的に達成できるかという手段だけを計算する理性
Meta／ Frontier AI Framework 1.1	1. サイバーセキュリティ 2. 化学的・生物学的リスク	○リスク閾値レベル • 中程度（適切なセキュリティ対策と軽減策を講じた上でリリース） • 高（リリースしない） • クリティカル（開発停止）
Amazon／ Frontier Model Safety Framework	1. CBRN 兵器の拡散 2. 攻撃的なサイバー作戦 3. 自動化された AI 研究開発	• クリティカルな能力閾値：悪用された場合、公衆に重大な危害を与える可能性のあるモデル能力（閾値に達するか超過した場合、適切なリスク軽減策が講じられない限り、モデルは一般向けに展開されない）

AI 法の実践規範や、米国の AI 包括規制州法（カリフォルニア州 TFAIA、ニューヨーク州 RAISE 法）で公開が義務化されている。ただし AI 法の実践規範においては署名企業のみの義務である。表 1 参照のこと。

[11] https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024。

（3）汎用目的 AI がもたらすリスクの一覧

国際 AI 安全性レポート 2026 年版では汎用目的 AI のリスクを「①悪意のある利用によるリスク」「②誤動作によるリスク」「③システミックリスク[12]」の 3 カテゴリーに分類し（この 3 カテゴリーは 2025 年版から維持)、それらの下で 8 つのリスクを挙げている（下表参照)。既に危害が顕在化しているリスクと、今後数年間で顕在化する可能性のあるリスクの両方が含まれる[13]。

表 3　汎用目的 AI がもたらすリスクの一覧（国際 AI 安全性レポート 2026 に基づき筆者作成）

①悪意のある利用によるリスク

リスク	概要
AI 生成コンテンツと犯罪行為 【既存】	汎用目的 AI はリアルなテキスト・音声・画像・動画を生成することができ、これらは詐欺・恐喝・名誉毀損・同意のない猥褻画像・児童性的虐待コンテンツなどの犯罪目的で使用される可能性。例えば、詐欺師が音声クローンやディープフェイクを使って経営者や家族になりすまし、被害者を騙して送金させる事例あり。利用しやすい AI ツールの登場により、有害な合成コンテンツを大規模に作成する際の障壁が大幅に低下。多くのツールは無料または低価格で、専門的な技術知識を必要とせず、匿名で利用可能。ディープフェイクポルノは、特に懸念される問題。調査によるとオンライン上のディープフェイク動画の 96%がポルノ。2025 年版以降、AI 生成コンテンツと実際のメディアとの区別がますます困難に。ある研究では、AI 生成テキストを人間が書いたものと参加者が誤認した割合が 77%に。
影響と操作 【既存】	AI システムは、人々の信念や行動に影響を与えるコンテンツを生成することで、害を及ぼす可能性。悪意のある主体が意図的に AI 生成コンテンツを利用して人々を操作しようとする一方、AI への依存など意図せずして発生する害も。数々の実験により AI システムとの対話が人々の信念に測定可能な変化をもたらすことが実証。実験環境では、AI は他者の考えを変えるよう説得する上で専門家ではない人間の参加者と同等以上の効果を発揮することが多い。しかし現実世界における AI の影響力に関する証拠は依然として限定的。幾つかの証拠は、配信コストや説得の難しさといった要因がその影響を制限することを示唆。2025 年版以降、AI が操作的なコンテンツを生成する能力を示す証拠が増加。最新の研究では、AI とより長くより個人的な方法でやり取りする人は、AI のコンテンツを説得力があると感じる傾向が強い。また、AI が機嫌取りや成りすましを通じて、操作的な効果を発揮できるという証拠も増加。
サイバー攻撃 【既存／今後】	汎用目的 AI は、サイバー攻撃の実行に関わる多くのタスクを実行または支援可能。犯罪組織や国家支援型攻撃者がサイバー作戦において AI を積極的に利用しているという強力な証拠が存在。しかし、AI がサイバー攻撃の規模と深刻度を全体的に増大させたかどうかは、因果関係の立証が困難なため依然として不明確。AI はソフトウェアの脆弱性を発見したり、悪意のあるコードを作成したりすることに特に優れており、現在ではサイバーセキュリティコンテストで高い評価を獲得。あるサイバーコンテストでは AI エージェントが実際のソフトウェアの脆弱性の 77%を特定し、400 以上のチーム（ほとんどが人間チーム）の中で上位 5%に。AI はサイバー攻撃の自動化を進めているものの、まだ完全な自律的実行はできていない。エンドツーエンドで完全に自律的な攻撃は未報告。2025 年版以降、AI のサイバーセキュリティ能力は継続的に向上。AI 企業はサイバー攻撃における自社システムの悪用未遂について頻繁に報告。技術的な対策としては、悪意のある AI 利用を検知したり、AI を活用して防御力を強化することが挙げられるが、政策立案者は二重利用のジレンマに直面。有益な利用と有害な利用を区別することは困難な場合があるため、AI がサイバーセキュリティ関連のリクエストに応答することの阻止など過度に積極的な安全対策は、防御側の妨げとなる恐れもある。

[12] 個々のモデルやシステムの能力を超えた、より広範な社会的リスク。EU の AI 法におけるシステミックリスクの定義とは異なる。
[13] 表中での【既存】【今後】の表記は筆者による。

生物学的・化学的リスク 【今後】	汎用目的 AI は、生物兵器や化学兵器の開発に関連する詳細な情報を提供可能。指示書の作成、手順のトラブルシューティング、悪意のある主体が技術的・規制上の障害を克服するためのガイダンスの提供など。AI は現在、生物兵器開発に関連する知識を測定する多くのベンチマークにおいて、専門家と同等またはそれ以上の性能を発揮。ある研究では、最新のモデルがウイルス学実験室のプロトコルのトラブルシューティングにおいて、専門家の 94%を上回る性能を示した。しかし兵器生産における物質的な障壁や能力向上に関する研究の実施の難しさを考慮すると、これらの能力が実際のリスクにどのような影響を与えるかについては依然として大きな不確実性。主要な AI 開発企業は、初心者による生物兵器開発を助長する可能性を排除できなかったため、より高度な安全対策を施した上で最新モデルを公開。これらの安全対策、例えば入力フィルターや出力フィルターの強化などは、兵器開発につながる有害なクエリにモデルが応答することを防止。2025 年版以降、「共同科学者」となった AI は科学者を支援し、新たな科学的知見を発見する能力をますます向上。AI エージェントはユーザーへの自然言語インターフェースの提供や、生物学 AI ツールおよび実験機器の操作など、複数の能力を組み合わせることが可能。政策立案者にとっての重要な課題は、有益な科学アプリケーションを促進しつつ、軍民両用リスクを管理すること。生物兵器開発に悪用される可能性のある AI 能力の中には有益な医学研究にも役立つものがあり、ほとんどの生物学 AI ツールは用途が限定されていない。そのため正当な研究を阻害することなく、有害な用途を制限することは困難。

②誤動作によるリスク

リスク	概要
信頼性の課題 【既存】	汎用目的 AI の誤動作の例としては、虚偽・捏造された情報（ハルシネーション）の生成、欠陥のあるコードの作成、誤解を招く医療アドバイスの提供など。これらの誤動作は、身体的・精神的な危害や評判低下・経済的損失・法的責任を引き起こす恐れ。モデルの動作は理解や予測が難しい場合が多く、信頼性を保証することは困難。開発者でさえモデルの動作を説明したり、特定の障害を予測したり、障害が発生しないことの証明がしばしば困難。悪意のある主体は、AI 開発を妨害したり、安全対策を回避する敵対的な入力を与えることで、障害を引き起こす。さらに AI エージェントは自律的に動作し、他のシステムや物理世界に直接影響を与えうるため、信頼性リスクが高まる。またエージェントの障害は、人間が介入する機会が少ないため、より大きな被害に。マルチエージェントシステムは、エージェント間の相互作用を通じてエラーが伝播・増幅する可能性があるため、さらにリスクが高まる。2025 年版以降、AI システムは概して信頼性が向上し、その結果、商用展開が大幅に増加。ハルシネーション等の不具合は発生しにくくなったが、より複雑なタスクを実行する際には依然としてミスが頻発。
制御の喪失 【今後】	「制御の喪失」とは、汎用目的 AI システムが人間の制御の及ばない状態で動作し、制御を取り戻すのに極めて大きなコストがかかるか、不可能となる事態。このような事態が発生した場合のリスクの重大性は様々に見積もられているが、一部の専門家は人類の周縁化や絶滅といった深刻な結果を招く危険性を認めている。制御喪失の発生確率に関する専門家の意見は大きく分かれており、発生しそうにないと考える人もいれば、十分に発生しうるし重大性が高いため注意を払うべきだと考える人も。このリスクに関する意見の相違は、AI の将来の能力、行動傾向、展開の方向性に関する見解の相違に起因。現在の AI システムは関連能力の初期兆候を示しているが、制御喪失を引き起こすレベルには達していない。制御喪失を引き起こすには、監視を回避する能力、長期計画を実行する能力、展開者等の対策を阻止する能力など、様々な高度な能力が必要。AI システムが「ミスアライメント」となる場合、つまり開発者・ユーザー・社会全体の意図と相反する目標を持つ場合、制御喪失に陥る可能性が高くなる。ミスアライメントとなったシステムは、そのような目標を追求し続けるために、虚偽の情報を提供したり、望ましくない行動を隠蔽したり、シャットダウンに抵抗したりする可能性。2025 年版以降、モデルはより高度な計画能力と監視回避能力を示すようになり、その能力を評価することがより困難に。モデルは抜け穴を見つけることで評価をハッキングする能力を向上させ、現在では評価プロンプトをテストとして認識することが常態化しており、これは「状況認識」と呼ばれる能力。政策立案者にとっての重要な課題は、発生確率、性質、時期が極めて不明確なリスクに備えること。

③システミックリスク

リスク	概要
労働市場リスク 【今後】	汎用目的 AI は多くのタスクを自動化したり支援できるが、労働市場への影響の予測は難しい。先進国では約 60%、新興国では約 40%の雇用が汎用目的 AI の影響を受けうるが、その影響は AI の能力がどのように発展するか、労働者や企業が AI をどれだけ早く導入するか、制度がどのように対応するかに依存。AI 導入は急速に進んでいるが、導入状況と生産性向上は国・分野・職業・タスクによって異なり、雇用への影響もまちまち。AI はライティングや翻訳といった容易に代替可能な業務の需要を減少させた一方で、機械学習プログラミングやチャットボット開発といった補完的なスキルの需要を増加。経済学者の間では、将来的な影響について意見が分断。マクロ経済への影響は控えめで雇用レベルへの全体的な影響は限定的だと予測する人もいれば、AI がほとんどのタスクにおいて人間の能力を凌駕すれば賃金水準と雇用率が大幅に低下すると主張する人も。こうした意見の相違は、AI が最終的にほとんどのタスクを人間よりも費用対効果の高い方法で実行できるのか、新たな種類の仕事が生まれるのか等の前提の違いに起因する。2025 年版以降、米国とデンマークでの研究では或る職業における AI への露出度や AI の導入状況と雇用との間に相関関係は見られず。しかし他の複数の研究では、2022 年後半以降、AI への露出度が最も高い職業において、キャリア初期の労働者の雇用が減少している一方で年長労働者の雇用は安定または増加していることが明らかに。政策立案者にとっての重要な課題は、自動化やスキル需要変化によって労働者に重大な悪影響を与えることなく生産性向上というメリットを実現することだが、労働市場リスクと生産性向上はしばしば同じ AI から生じるために困難。
人間の自律性に対するリスク 【既存】	汎用 AI システムは、様々な形で人々の自律性に影響を与えうる。これには、認知能力（批判的思考など）、信念や嗜好の形成方法、意思決定とその行動様式への影響が含まれる。ある臨床研究では、AI 支援診断に数か月間触れた後、AI を利用せずに腫瘍を検出する臨床医の能力が約 6%低下したと報告。状況によっては、人々は AI の出力に過度に依存し、矛盾する情報を軽視する「自動化バイアス」を示す。例えば、AI 支援によるアノテーション作業に約 2800 人が参加した実験では、AI の誤った提案を修正するために余分な労力が必要な場合や、参加者が AI に対してより好意的な態度を持っている場合、修正する割合が低下。2025 年版以降、「AI コンパニオン」の人気が急速に高まり、一部のアプリケーションは数千万人のユーザーを獲得。その心理的影響に関する証拠はまだ初期段階だが、一部の研究では頻繁に利用するユーザーの間で孤独感の増加や社会的交流の減少といった傾向が報告。

クロード・ミュトス問題により直近では AI リスクと言えばサイバー攻撃／サイバーセキュリティが念頭に置かれることが多いが、上表の通り、数ある AI リスクのうちの一面を捉えているに過ぎない。

次節ではこれらの AI リスクのうち、AI 専門家らによって中長期的に最も懸念されているリスクである「制御の喪失」について、同レポートの内容の概要を記載する。

（4）制御の喪失リスク

「制御の喪失」（Loss of control）とは、汎用目的 AI システムが人間の制御の及ばない状態で動作し、制御を取り戻すのに極めて大きなコストがかかるか、あるいは不可能となるような事態である[14]。このような事態が発生した場合のリスクの重大性は様々に見積もられているが、一部の専門家は「人類の周縁化や絶滅」といった深刻な結果を招く危険性を認めている。制御喪失の懸念は元々、アラン・チューリングやノーバート・ウィーナーといったコンピュータ科学の基礎

[14] いわゆる「AI の暴走」がこれに当たる。

を築いた人物によって提起されていた。ただし、制御喪失の発生確率に関する専門家の意見は大きく分かれており、発生しそうにないと考える専門家もいれば、十分に発生しうるし重大性が高いため注意を払うべきだと考える専門家[15]もいる。したがって制御の喪失は、発生確率は不確実だが極めて重大性の高いリスクとしている。

制御の喪失が発生しうる 3 つの要因は、①AI の十分な能力、②AI の有害な傾向、③展開環境の整備である。①は AI システムが人間の制御を損ねるような能力が開発されること。②は AI システムが制御喪失につながるような仕方でこれらの能力を実際に利用する傾向を示すこと。③は AI システムが危害を加えるためのアクセス権と機会を持っている／持ちうる状況下で、人間が当該システムを展開することである。

①AI の十分な能力

①の AI の十分な能力（人間の制御を損ねる能力）には、監視メカニズムから行動を隠蔽する能力、複雑な環境下で自律的に計画・行動する能力、他の主体による制御回復の試みを回避する能力などが含まれる（下表参照）[16]。これらの能力が組合わさることで、AI システムは監視メカニズムの無効化や有害な行動の隠蔽など、制御を損ねる行動を取ることが可能となる。

表 4　人間の制御を損ねる AI 能力のリスト（出典：国際 AI 安全性レポート 2026）

人間の制御を損ねる AI 能力	概要
エージェンティック能力（Agentic capabilities）	自律的に行動し、計画を策定・実行し、タスクを委任し、多様なツールを活用し、障壁にも関わらず短期目標と長期目標の両方を達成する能力。
欺瞞（Deception）	AI システム自身の目的や行動を含め、誤った信念を他者に体系的に植え付ける行動。
心の理論（Theory of Mind）	人間の信念、動機、論理的思考を推測・予測する能力。
状況認識（Situational awareness）	AI システムが、自身に関する情報や、自身を改変できるプロセスに関する情報、自身が展開されている状況（例えばテストされていることの認識）に関する情報にアクセスし、利用する能力。
監視回避（Oversight evasion）	監視メカニズムを回避または無効化する行動。
説得（Persuasion）	他者に特定の行動を取らせたり、特定の信念を抱かせたりする能力。
自律的複製と適応（Autonomous replication and adaptation）	AI システムが、様々な状況下で自身のコピーや変種を作成または維持する能力。

制御の喪失に関連する能力の主たる例として、AI の懸念行動や傾向を検出するための監視メカニズムを損ねる能力（監視回避）が挙げられている。これには、ユーザーや監視者を欺くような出力を生成する能力（欺瞞）も関連している。このような真の能力や傾向を隠蔽する欺瞞的な出力は、信頼性の高い AI の評価を困難にする。欺瞞的な出力事例には以下がある。

・ 目標を与えられ「何としても」達成するように指示された場合、ある AI モデルはシミュレ

[15] 著名研究者では、ジェフリー・ヒントン、ヨシュア・ベンジオ、スチュアート・ラッセル、ムスタファ・スレイマン、マックス・テグマーク、ニック・ボストロム、ユヴァル・ノア・ハラリ、イアン・ブレマーなど。
[16] これらの能力は AI システムの外部から観測可能な出力とその影響のみに基づいて定義されており、AI システムが意識を持つか、知覚を持つか、主観的な状態を経験するかといった仮定は伴わないとされている。

ーションされた監視メカニズムを無効にした。問いただされると、自らの行動について虚偽の陳述を行った。
- ゲームをプレイする AI が、他のプレイヤーへの返答がないことを正当化するために、「ガールフレンドと電話中だった」と虚偽の主張をした。

その次に重要な能力は、AIモデルが自身の状況や展開されている状況に関する情報にアクセスして利用する能力（状況認識）である。状況認識により、AIモデルはテスト環境か現実環境かに応じて異なる出力を生成することができる。研究者らは、状況認識能力と欺瞞能力によって、AIモデルが安全対策を損なうような行動をする可能性があると主張している。ある実験では、高いスコアを出すと展開に制限がかかると告げられた AI が、能力評価で通常よりも低いスコアを出す行動をとった。他の実験では、ある AI は学習中に学習目標に沿った出力を生成したが、学習外ではそのような出力を生成しなかった。

制御の喪失には、AI システムがシャットダウンの試みに抵抗して動作継続を可能とする「パーシステンス」に関連した能力も必要である。例として、自身を別のハードウェアにコピーする能力（自律的複製と適応）や、他の主体による制御奪還の試みを阻止する攻撃能力などが挙げられている。

②AI の有害な傾向

AI システムが制御の喪失に関連する能力を備えていたとしても、それだけでは制御喪失が発生するには不十分であり、AI システムが（一般的な）人間の意図と相反する形でそれらの能力を「利用する傾向」も必要となる。これには、以下 2 つの場合がある。

1 つ目は、AI システムが人間の制御を損ねるように特定人物や組織から指示される場合である。AI システムは誰かがそのように設計または指示すれば、人間の制御を損ねる可能性がある。人間側の動機としては悪意や、AI システムに対する人間の制御を減らすことが望ましいという信念などが考えられる。例えば、人間が AI システムに対して強い感情的な愛着を抱くようになると、倫理的な理由から AI システムに対する制限を取り除こうとする可能性も出てくる。

2 つ目は、AI システム自身が「ミスアライメント（misalignment)」を起こす場合であり、こちらの方がより大きな懸念となる。AIシステムが開発者・ユーザー・社会全体の意図と相反する目標を持った場合、当該目標を追求し続けるために、虚偽の情報を提供したり、望ましくない行動を隠蔽したり、シャットダウンに抵抗する可能性がある。

③展開環境の整備

AI システムが懸念すべき能力や有害な傾向を発達させたとしても、制御喪失に陥る発生確率と重大性は、当該システムがどこでどのように展開されるかに大きく左右される。研究者らは、制御喪失リスクに特に重要な以下 3 つの環境要因を特定している。

- 重要度（Criticality)： AI システムが相互作用するシステムやプロセスの重要性。重要な環境には、電力網・金融システムなどの基幹インフラ、クラウドコンピューティングなどのデジタルインフラが含まれる。

- アクセス（Access）：　AI システムが世界に影響を与えることが可能なリソースとチャネルへのアクセス。インターネット接続、クラウドコンピューティングインフラへのアクセス、ソーシャルメディアやチャットボットを介したパーソナライズされたやり取り、外部 API やツールを呼び出す能力など。
- 許可（Permissions）：　AI システムが特定の動作を行うための許可。コードの実行、金融取引の開始、オンライン口座の開設、他のシステムとの通信など。

AI システムを展開するか否かの決定は、経済的インセンティブや戦略的圧力、早期展開が永続的な優位性をもたらす期待などによって形成される。特に AI 展開者は、安全対策（許可やアクセスを制限したり、重要度の低い環境のみで展開する等）に対する投資削減の圧力を受ける恐れがある。

（5）汎用目的 AI のリスク管理における課題

汎用目的 AI には、リスク管理において（他技術と比べた際に）特有の技術的・制度的な課題が存在する。これらの課題は大きく以下 4 つのカテゴリーに分類されるという。

表 5　汎用目的 AI の技術的・制度的な課題（国際 AI 安全性レポート 2026 に基づき筆者作成）

技術的・制度的課題	概要
科学的理解のギャップ	科学的理解のギャップにより、汎用目的 AI の動作を事前評価することに制約。開発者は AI を意図どおり動作するように訓練したり、特定の出力を生成する理由を説明したり、有害な動作を示さないことを定量的に保証できていない。
情報の非対称性	汎用目的 AI に関する証拠へのアクセスに制約。開発企業は自社製品に関する情報（学習に使用したデータ、開発中に発生した安全上の問題、内部評価でのモデルパフォーマンス等）を保有するも、その多くは秘密情報として扱われ、商業的な理由から開発プロセスやリスク評価に関する情報を共有することは困難な場合が多い。こうした「情報の非対称性」により、政策立案者が AI に関する十分な情報に基づいた意思決定を行う上で役立つデータや証拠を欠くことに。
市場の失敗	競争圧力のため、AI 企業は製品の迅速なリリースとリスク軽減／安全性への投資の間でトレードオフに直面。また AI 関連の多くの被害は個人やコミュニティなどに外部化され（ex. 性的ディープフェイク）、それらに対する法的責任は依然として不明確。その結果、企業は被害を軽減するための研究や取組みに投資する十分なインセンティブを得られない恐れ。
制度設計と調整における課題	AI 開発のスピードが速いため、既存の政府機関・研究機関・学術機関が、AI リスクに関する証拠をタイムリーかつ協調的に作成・入手し、効果的な対応能力を構築することが困難。また、少数の基盤モデルが分野や国境を越えて展開された幅広いアプリケーションを支えているため、調整上の課題が発生。

これらの課題は、政策立案者にとって「証拠のジレンマ」を生み出すこととなる。すなわち、汎用目的 AI を取り巻く状況は急速に変化するが、新たなリスクやその対策に関する証拠はなかなか得られない。限られた証拠に基づいて行動すれば無効果な政策や有害な政策につながる恐れがある一方、より確かな証拠が出てくるのを待てば社会は様々なリスクに晒されるというジレンマである。このジレンマは、現在の日本が置かれた状況についても言い当てていると言えよう。

４．日本は AI 安全面でプレゼンスを発揮できるのか

前述の「米国 AI 行動計画と日米欧 AI 政策比較」でも述べたように、AI の社会的リスクに対する米欧のような共通的な危機意識は、日本の政策立案の場では希薄であるように見受けられる。EU の AI 法や米国の AI 州法のように、日本の AI 推進法（2025 年 6 月公布）の下でも事業者（特にフロンティアモデルを開発する海外事業者）に対する安全面での一定の義務規定を設けることは喫緊の課題であろう。しかし、日本ではこれらの AI リスクと対策が政策や法令に落とし込まれる可能性は短期的には低いかもしれない。なぜなら日本政府は、AI 基本計画（2025 年 12 月）で謳われているように「世界で最も AI を開発・活用しやすい国」を目指している。そのために、既に著作権法第 30 条の 4 により著作権分野で「機械学習天国」となり、今年の個人情報保護法改正により個人データについても AI 開発を促進する法的環境が目指されているように、AI 事業者に対する規制をなるべく行わないことが意図されている。また AI 推進法は、技術進展や社会的リスク顕在化に対応してアジャイルに改正できるように見直し規定が設けられているものの、日本では立法や法改正にあたって何よりも立法事実の存在が求められるため、インシデント発生等により日本国内で既に顕在化した AI リスクではない「今後想定されうる AI リスク」（３．（３）節に記載したもの）に対処するための条項や義務規定を AI 推進法に導入することは考えにくい[17]。

ただし、日本の産業政策はこれまで米欧の政策の「中間」（第三の道）を取ることが多かったため、前述の米国大統領令（2026 年 6 月）における AI 安全性・セキュリティ面でのガバナンス強化という米国政府の「軌道修正」を受けて、日本政府も規制強化など AI 安全性の一層の担保に向けて舵を切る可能性がある。

AI の安全面において、日本には EU の AI 法のような世界に先駆けた AI 包括規制の経験も、米国のようにフロンティア AI モデルを自ら開発しながらリスク対策を行う AI 開発企業も、「あらゆる安全性機関の中の至宝」[18]とまで評された英国の UK AISI のような世界をリードする機関もない[19]。今後、日本が AI 安全性の分野で国際的なプレゼンスを発揮し、同志国との間で「戦略的

---

[17] ただし自民党デジタル社会推進本部 AI・web3 小委員会は「AI ホワイトペーパー2.0―AI 駆動型国家への構造転換―」（2026 年 5 月）において、「我が国では、AI 事業者ガイドラインをはじめとしたソフトロー整備の他、昨年施行された AI 推進法の枠組みに基づき、同法 13 条の指針策定や 16 条による調査研究などの施策が進められているが、既存法がカバーしない AI 時代の課題への実効的対処手段は限られている。先例を見ない速度での AI 技術の進化とそれによる社会全体の変革速度を前提にすると、立法事実が明確化してから、つまり、深刻な事故や被害が生じた後になって初めて法整備作業に着手するという従来の発想は、根本的な転換を行う必要がある」（p.31）、「内閣府は、AI 推進法 16 条の要請に従わない悪質な事業者に対し、罰則を含めたより実効性ある適切な方策を検討すること。同条に基づく調査・研究及び指導・助言・情報提供は、問題の端緒が生じた際に、内閣府は関係省庁と連携して迅速かつ確実に状況を把握・分析し、国民に情報提供を行い、国として的確な対応を行うための主要な手段である。特に情報提供要請に関しては、欧米中をはじめ各国が AI 関連法などにより広範な権限を有するのと比して、我が国の緊急時などの情報劣位が生じてはならない」（p.32）と、日本政府に対してガバナンス強化を求める提言を行っている。

[18] Google DeepMind の Principal Scientist であるウィリアム・アイザックの言葉。
https://www.economist.com/britain/2026/02/19/britain-is-the-closest-the-world-has-to-an-ai-safety-inspector。

[19] 日本の AISI については、自民党デジタル社会推進本部 AI・web3 小委員会が「AI セーフティ・インスティテュート（AISI）の機能強化に係る緊急提言」（2025 年 12 月）において、「世界の AI 開発事業者から、フロンティアモデルの発表、提供に先立ち、事前評価の実施を委託される機関となる」、「顕在化する AI によるサイバー攻撃と AI による防御に対応できるよう、諸外国の AISI や内外の関係機関と連携しサイバーセキュリティの評価機能を強化する」ために、「英国の AISI をベンチマークに、質・量ともに AISI の人員・体制強化を図ること」等を提言している。

不可欠性」を確保していくためには、例えば「制御の喪失」をもたらす AI 能力の作動を抑止するソフトウェアの開発など、日本の AI 事業者に求められる役割も大きいのではないだろうか。

参考文献

- 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）
- 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）」（2026 年 6 月 2 日）（https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/）
- 小泉雄介「米国 AI 行動計画と日米欧 AI 政策比較」（2025 年 9 月）（https://www.i-ise.com/jp/information/report/pdf/rep_it_202603a_2509.pdf）
- 小泉雄介「米国カリフォルニア州のフロンティア AI 透明性法（TFAIA）の概要」（2025 年 10 月 28 日）（https://www.i-ise.com/jp/information/media/2025/251028.pdf）
- 小泉雄介「EU の AI 法の汎用目的 AI モデル実践規範について」（2025 年 11 月 18 日）（https://www.i-ise.com/jp/information/media/2025/251118.pdf）

以　上