

誰がどのように AI をコントロールし安全を確保するのか

——アメリカにおける AI と安全をめぐる攻防

2026 年 9 月 27 日

九州大学法学研究院

成原 慧 准教授

1. はじめに

AI の急激な発展に伴い、AI による科学研究の高度化・自動化などの便益が期待される一方で、AI の安全性や制御可能性に対する疑念が広がっている。本年（2026 年）7 月には、OpenAI の開発した AI モデルのサンドボックスでの試験中に、多数の AI エージェントが、自律的に協調してサンドボックスを抜け出し、AI モデル共有プラットフォーム Hugging Face のシステムに侵入したことが明らかになった¹。最近では他の事業者が開発した AI モデルによるものも含め類似した事案が多数生じており²、AI が開発者のコントロールを離れてセキュリティを侵犯するリスクが顕在化している。さらに、9 月には、Anthropic で AI 開発に携わっていた研究者が、同社を退職するとともに、近い将来、先端的な AI が自己改良を進め人間を超える知能を獲得することにより、人間による AI の制御が困難となり、人類の滅亡を招くおそれがあるといった懸念を表明し³、広く反響を呼んだ。AI の制御喪失のリスクや「超知能」（Superintelligence）の実現による人類存亡のリスクは、今になってにわかに湧いてきた懸念ではなく、第 3 次 AI ブームの初期に当たる 2010 年代中盤頃から一部の研究者や開発者らにより問題提起され、検討や対策の必要性が訴えかけられてきた⁴。こうしたリスクは、当時においては思考実験に基づく理論上のもの

¹ [OpenAI, “Hugging Face Incident Technical Report” \(August 26, 2026\); Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk, “Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident, METR” \(August 26, 2026\).](#)

² [Kate Conger, “What to Know About Recent A.I. Hacks,” *the New York Times* \(Sep. 22, 2026\).](#)

³ [See e.g., Jacob Coxon \(@hilbertspaess\), X \(Sep. 9, 2026, 9:04 AM\).](#)

⁴ [See e.g., Stuart Russell et al., “Research Priorities for Robust and Beneficial Artificial Intelligence,” *AI Magazine*, Vol. 36\(4\), pp.105-114 \(2015\); Dario Amodei et al.,](#)

のだったかもしれないが、この 10 年間の AI の急激な発展により、徐々に現実味を帯びてきたところがある。

こうした AI の安全や制御可能性に関わるリスクに我々はいかに対処すべきなのだろうか。リスクには科学的な根拠に基づいて冷静に対応することが必要であり、むやみに危険や極端なシナリオを煽るべきではないだろう。同時に、不確実ではあるものの重大なリスクについて、あらかじめ生じうる複数のシナリオを予測し、最悪の事態を防ぐための対応策を検討することも、必要だろう⁵。そこで、本稿では、急激に発展する AI を誰がいかにコントロールし、安全を確保すべきなのかについて、関連する議論が活発になっているアメリカの動向を手がかりに検討することにした。

2. アメリカの AI 政策の変遷

アメリカの AI の安全性に関する政策は、オバマ政権の末期にまで遡ることができる。2016 年 5 月から 7 月にかけて、オバマ政権は、AI に関する法的・社会的・政策的課題について各地の大学等と共催してワークショップを開催した⁶。ワークショップでの議論などを踏まえ、同年 10 月に、大統領府国家科学技術会議技術委員会は、AI の制御可能性や透明性の確保、人間の価値との合致などを求める報告書を公表した⁷。

一方、第一期トランプ政権は、AI の活用を重視する見地から AI 政策を展開した。トランプ大統領は、2020 年 12 月に、連邦政府における信頼に値する人工知能の利用の促進に関する大統領令を発した。大統領令は、連邦政府機関における AI の利用の促進を図ると

[“Concrete Problems in AI Safety,” arXiv:1606.06565 \(2016\); Demis Hassabis et al., “Asilomar AI Principles” \(2017\).](#) ニック・ボストロム（倉骨彰訳）『スーパーインテリジェンス』（日本経済新聞出版社、2017 年）参照。

⁵ 我が国における AI のリスクに対するこのようなアプローチの先駆けとして、[総務省情報通信政策研究所 AI ネットワーク化検討会議「中間報告書」（2016 年 4 月）](#) 41 頁以下、福田雅樹＝林秀弥＝成原慧（編著）『AI がつなげる社会—AI ネットワーク時代の法・政策』（弘文堂、2017 年）参照。

⁶ 当時、筆者も、総務省情報通信政策研究所主任研究官として、ホワイトハウスとワシントン大学の共催によるワークショップ Legal and Governance Implications of Artificial Intelligence に参加し、クローズド会合において、日本が提案していた AI 開発原則案の内容を紹介した。

⁷ [Whitehouse, *Preparing for the Future of Artificial Intelligence* \(2016\).](#)

ともに、その際に遵守すべき原則を定めた⁸。

2021年に民主党が政権を奪還して以降、バイデン政権は、AI開発事業者の自主的な取組を尊重しつつ、AIの安全性や信頼性を確保するための政策を推進してきた。2023年10月に、バイデン大統領は、「AIの安全、堅牢で信頼に値する開発および利用に関する大統領令」を発した。大統領令は、連邦政府機関にAIの安全性や信頼性を確保するための規則や指針、技術標準の策定などの取組を求めた⁹。このようなバイデン政権のAI政策は、「広島AIプロセス」など、国際的なAIガバナンスの構築のための議論や取組とも調和を図ったものだった¹⁰。

ところが、バイデン政権のAI政策は、第二期トランプ政権の発足とともに大きく転換された。大統領に返り咲いたトランプは、AIの技術発展とイノベーションによりアメリカの国際競争力を向上させようと、AI規制の緩和を図ってきた。第二期政権発足後まもなく、2025年1月、トランプ大統領は、2件の大統領令を発し、バイデン政権のAI大統領令を撤廃するとともに¹¹、関連する連邦政府機関に、AIに関する新たな行動計画の策定を求め、バイデン政権時代に作られたAI関係の政策・規制の見直しを求めた¹²。同年7月、トランプ政権は、「アメリカAIアクションプラン」を策定した。アクションプランは、AI関係のハード、ソフト、標準の輸出の推進、データセンターや半導体工場の増設の促進、AI開発・実装を妨げる規制の撤廃など、AI分野でのアメリカの国際競争力を向上させるための政策を前面に打ち出している¹³。

3. AI規制をめぐる州政府と連邦政府の攻防

AI規制を緩和する連邦政府の姿勢に対抗するかのようにより、一部の州は、州法によるAI

⁸ Executive Order No. 13960, 85 Fed. Reg. 78939 (Dec. 8, 2020).

⁹ Executive Order No. 14110, 88 Fed. Reg. 75191 (Nov. 1, 2023).

¹⁰ バイデン政権までのアメリカのAI法・政策につき、[成原慧、実積寿也、小熊美紀「AIをめぐる米国の政策と法的対応ー連邦および州の動向ー」情報法制研究 16号 34頁（2024年）](#) 参照。

¹¹ Executive Order No. 14148, 90 Fed. Reg. 8237 (Jan. 28, 2025).

¹² Executive Order No. 14179, 90 Fed. Reg. 8741 (Jan. 31, 2025).

¹³ [The White House, *Winning the AI Race: America's AI Action Plan* \(July, 2025\).](#)

規制の導入・強化を試みてきた。現在までに、カリフォルニア州、コロラド州、ユタ州、テキサス州、ニューヨーク州、ワシントン州が州法による AI 規制を導入している¹⁴。最近でも、本年 9 月、依然として AI 規制の強化に消極的な連邦政府の姿勢に対抗して、カリフォルニア州のニューサム知事が先端的な AI モデルの開発事業者に対して外部機関による評価や緊急停止装置（kill switch）の実装などを義務づける方向で州法を改正する可能性について検討するよう州政府機関に指示する行政命令を発するなど、一部の州が AI に関するルールメイキングを主導しようとしている面がある¹⁵。

こうした州の動きに対抗して、トランプ政権は州による AI 規制を牽制しようとしてきた。2025 年 12 月、トランプ大統領は、大統領令「人工知能のための全米政策枠組みを確保する」を発した。大統領令は、州法による AI 規制が、州ごとに異なる規制のパッチワークを作り出しているという認識の下に、連邦レベルで企業にとっての負担が最小限となるような統一した AI 政策の枠組みを形成することを目標に掲げた。大統領令は、連邦の AI 政策に合致しない州法の合憲性・適法性を争うための AI 訴訟タスクフォースを設置するとともに、連邦政府の方針に合致しない規制を行う州への助成を制限するよう、連邦政府機関に指示した¹⁶。その後、本年 3 月にトランプ政権は、「全米 AI 政策枠組み」を公表した。同文書は、州法による断片的なパッチワークの規制を避けるために、全米統一のルールを制定するよう連邦議会に勧告した¹⁷。

州法による AI 規制には功罪両面がある。従来から連邦制の枠内で州は、多様な規制の実験を試みてきた。このような各州による多様な規制の実験により、アメリカ法は試行錯誤を通じて発展してきたところがある。連邦政府が AI 規制に慎重な姿勢をとる中で、州法による分権的な規制の実験は、AI 規制の発展を促す役割を果たすだろう。一方、このような分権的な実験には弊害も伴う。州ごとにばらばらの規制は、AI 開発事業者にとって遵守するための負担が大きくなり、規制の影響力も不十分なものとなるおそれがある。

¹⁴ [iapp, “US State AI Governance Legislation Tracker” \(Last updated 28 April 2026\)](#).

¹⁵ Executive Department State of California, Executive Order N-9-26 (Sep. 18, 2026).

¹⁶ Executive Order No. 14365, 90 Fed. Reg. 58499 (Dec. 16, 2025).

¹⁷ [White House, *National Policy Framework Artificial Intelligence* \(March, 2026\)](#).

4. AI と安全保障をめぐる Anthropic とトランプ政権の攻防

第二期トランプ政権の下で、安全保障分野での AI の利用が進むとともに、その利用のあり方をめぐる政府と AI 開発事業者の間の攻防も激しくなっている。国防総省は、AI を国防用の機密システムに導入してきた。その中で、Anthropic は、国防総省に AI システムを納入しつつ、安全性や倫理を重視した AI 開発を行ってきた。同社は Claude をアメリカ国民の監視や完全自律型兵器の運用に用いないよう要求し、国防総省との契約により制限しようとした。トランプ政権は、民間企業との契約により国家の安全保障に関わる判断が制限されるべきではないという立場から、このような同社の要求を拒否した。

トランプ政権と Anthropic の対立は、民間の AI 開発事業者が、AI による監視から国民のプライバシーを守ったり、AI の自律的判断による戦禍の拡大を防ぐために、契約の条項などを通じて安全保障に関わる政府の判断や行動を制約することが、どこまで正当なのか、そうした事業者の姿勢に政府がどのような対応をとることができるのか、という難問を提起している。本年 3 月、Anthropic の方針に対抗してトランプ政権は同社をサプライチェーン・リスクに指定した。そこで、Anthropic は、指定が言論の自由を保障した合衆国憲法修正 1 条や適正手続を保障する修正 5 条に違反する報復措置であり、裁量を濫用し、権限を逸脱しており違法だとして提訴した。8 月 27 日、カリフォルニア北部地区連邦地裁は、同社の主張をおおむね認め、指定は恣意的で根拠を欠いたものだとして、指定の差止めを認めた¹⁸。一方、9 月 25 日、コロンビア特別区巡回区連邦控訴裁判所は、Anthropic と国防総省との別の訴訟において、Claude を同省の情報システムに組み込み続けることにより国家安全保障上のリスクがもたらされるとする同省の判断には十分な根拠があるとして、Claude を同省のサプライチェーンから排除した同省の決定を是認した¹⁹。8 月の判決と 9 月の判決では、争われた指定の根拠規定が異なるという前提の違いはあるものの、裁判所の判断が分かれており、9 月の判決でも、3 名の裁判官のうち 1 名は反対意見を述べるなど、司法も、排除の理由や必要性をどのように捉えるかという点も含め、きわどい判断を迫られているといえよう。

¹⁸ Anthropic PBC v. U.S. Dep't of War, No. 26-cv-01996-RFL (N.D. Cal. August 27, 2026).

¹⁹ Anthropic PBC v. Department of War, No. 26-1049 (D.C. Cir, Sep. 25, 2026).

5. 先端的な AI モデルの安全性を確保するための取組

2026 年に入り、先端的な AI モデルの能力が急速に向上する一方で、その利用がサイバーセキュリティや安全保障にもたらすリスクが顕在化し、アメリカ政府も対応を迫られている。Mythos など先端的な AI モデルの急速な能力の向上に伴い、それらを用いてサイバーセキュリティの脆弱性が発見されるなど、AI によるセキュリティ侵犯などのリスクが高まっている。こうした中で、AI 規制の緩和を図ってきたトランプ政権も、先端的な AI モデルによるサイバーセキュリティや安全保障への脅威に対処するために、AI の開発や利用に対して必要に応じて介入する姿勢を示すようになった。6 月には新たな大統領令を発し、先端的な AI モデルが国家安全保障に脅威を与えるおそれがあるという認識の下に、AI 開発事業者との自主的な協定に基づく先端的な AI モデルの事前審査やアクセスの共有の仕組みを導入した²⁰。

アメリカ政府は、自国の安全保障のために先端的な AI モデルの輸出制限も試みている。本年 6 月、トランプ政権は、安全保障上の観点から輸出管理規制に基づいて Anthropic に Mythos5 および Fable5 への国内外の外国人によるアクセスを一時停止するよう指示した。政府の指示を踏まえ、同社は、すべての顧客に対し、Mythos5 および Fable5 の提供を一時停止した²¹。日本をはじめ多くの国の人々にとって、自分たちが頼っている先端的な AI モデルが、アメリカ政府の輸出管理規制により、ある日突然使えなくなるというリスクが顕在化したのである。

一方、OpenAI や Anthropic など主要な AI 開発事業者も、AI の安全性を確保するための取組を進めるとともに、アメリカ政府に対して AI 規制の導入を検討するよう働きかけてきた²²。最近では、OpenAI のモデルによる Hugging Face のシステムへの侵入事案など、AI の自律性の向上による安全へのリスクの増大に伴い、AI の制御可能性と安全性を確保するための取組の強化を求める声は強まっている。本年 9 月 12 日、Anthropic の共同創設者であり CEO の Dario Amodei は、AI の再帰的な自己改良が始まりつつあることや、先

²⁰ Executive Order No. 14409, 91 Fed. Reg. 34565 (June 5, 2026).

²¹ [Anthropic, “Announcements: Statement on the US government directive to suspend access to Fable 5 and Mythos 5” \(June 12, 2026\).](#)

²² [Anthropic, *Anthropic’s advanced AI Frameworks* \(June 2026\); OpenAI, *Democratic Governance of Frontier AI: A blueprint for a federal framework* \(June 2026\).](#)

述の AI によるシステム侵入事案などを踏まえ、AI の安全性を確保するために十分な時間を捻出できるよう、他の AI 開発事業者らに、AI モデルの開発を減速するとともに、第三者機関による評価の仕組みなどを導入するよう呼びかけた²³。Amodei 氏の呼びかけには、OpenAI など他の有力な AI 開発事業者も呼応する姿勢を示しているが、安全の確保を目的に掲げているとはいえ、AI 開発事業者間の協調による開発の減速は競争法上の問題を提起する可能性もある。実際、Anthropic や OpenAI ら AI 開発事業者 4 社の AI モデルの利用者らが、競合する AI 開発事業者らによる AI 開発の減速等の合意はシャーマン法 1 条の禁じる取引制限行為に当たるなどとして、4 社に対しこのような行為の差止めなどを求めて提訴している²⁴。

6. 国際協調の必要性

AI の安全性を確保するためには、国際的なガバナンスの確立・強化も求められる。先端的な AI モデルの開発競争は、アメリカと中国を中心に国際的に展開されており、一国の規制により AI のリスクを抑制するというアプローチには限界があるからだ。AI の開発競争が国際的に展開される中、自国だけが厳格な安全規制を導入すれば、他国との競争に不利になるおそれがある。また、先端的な AI モデルの及ぼすリスクは一国内にとどまらない。AI モデルは国境を越えて提供され、サイバー攻撃などその利用に伴うリスクも国境を越えて波及する。そこで、AI ガバナンスの発展のための国際協調が必要となる。例えば、Amodei 氏も先の声明で提案しているように、AI 開発の減速が必要なのだとすれば、アメリカの主要な AI 開発事業者間の協調が求められるのはもとより、民主主義国家の間の協調、さらには中国など政治体制を異にする国々も含めたグローバルな協調も望まれるだろう。本年 9 月 21 日、OpenAI も、再帰的な自己改良を行う AI など先端的な AI モデルの開発の安全性とセキュリティに関する国際的な技術標準の策定と情報共有のための枠組みの構築に向けてアメリカが主導的な役割を果たすべきとの意見を表明している²⁵。

しかし、そのためのリーダーシップを期待されているトランプ大統領は、AI についても国際的なガバナンスの構築には消極的であり、AI の安全を確保するための国際協調の実現に向けた道のは前途多難である。本年 9 月 22 日、国連総会での演説でトランプ大統

²³ [Dario Amodei, “We Must Pace the Frontier” \(Sep. 2026\).](#)

²⁴ [Complaint, Buist v. Anthropic, PBC, No. 3:26-cv-10693 \(N.D. Cal. Sep. 18, 2026\).](#)

²⁵ [OpenAI, “Building standards for the next phase of AI” \(Sep. 21, 2026\).](#)

領は、「人工知能をコントロールするためのグローバルリストの枠組みを構築しようとするあらゆる企て」を拒絶する姿勢を示している²⁶。一方、翌日の安全保障理事会では、Anthropic や OpenAI の CEO らが、AI の制御喪失や悪用のリスクを防ぐために、国際的な技術標準を策定するなど国際協調の必要性を訴えた。また、中国の代表も、AI に関するグローバルガバナンスの構築を積極的に支持する姿勢をみせた²⁷。

もっとも、トランプ大統領も、AI の覇権を争うもう一つの大国である中国との間で協調を模索することには前向きな姿勢を示している。9 月 25 日に行われた米中首脳会談において、両国は、両国の間で AI に関するリスクと便益について意見交換を行うための対話を開始すること、そして、AI に関するインシデントについて両国で情報を共有するチャンネルを構築することに合意した²⁸。

7. 日本への示唆

本稿で見てきたとおり、先端的な AI モデルがサイバーセキュリティや安全保障に及ぼすリスクに対処するために、アメリカの連邦政府や州、事業者はそれぞれ取組を進めているが、それらの間の思惑の違いもあり、しばしば対立も生じている。各国の政府や個々の事業者による AI の安全性を確保するための取組は重要であるが、米中を中心に AI の開発競争が国際的に展開され、ネットワーク化された AI の利用も国境を越えて広がっていることから、AI の安全性を確保するためには国際協調の深化が必要となる。しかし、アメリカ国内外から国際協調に向けたリーダーシップを期待されているトランプ大統領は、AI の安全性を確保するための国際的なガバナンスの構築に消極的である。一方、中国は、AI に関するグローバルガバナンスの構築に積極的な姿勢を示しており、トランプ大統領も、米中の二国間での AI に関する対話と協調には前向きな姿勢を示している。しかし、日本や欧州諸国など米中以外の国々にとっては、AI の覇権を争う 2 つの大国のみで、人類の命運を左右するかもしれない AI の安全性の確保のあり方について方針が決められることには不

²⁶ [Donald J. Trump, President of the United States, Address to the 81st Session of the United Nations General Assembly \(Sep. 22, 2026\).](#)

²⁷ [Artificial intelligence and international security, United Nations, Security Council, 10228th meeting \(Sep. 23, 2026\).](#)

²⁸ [The White House, “Fact Sheet: President Donald J. Trump Advances a Fair and Reciprocal Relationship with China While Hosting Historic State Visit” \(Sep. 25, 2026\).](#)

安も拭えないだろう。

こうした中で、日本としても、AIの安全性を確保するための国際協調の拡大・深化のために積極的に関与していく必要がある。我が国は、2016年4月のG7香川・高松情報通信大臣会合において、透明性、制御可能性、セキュリティ確保、安全保護などを内容とするAIの研究開発の原則の策定に向け国際的な議論を進めていくことを提案した。こうして始まった国際的な議論は、OECD／G20のAI原則の策定などに結実することになる²⁹。第一期トランプ政権下の2019年に、アメリカ、中国、欧州諸国や日本が参加するG20がAI原則に合意したという実績も示しているように、国家間の利害対立が激しい中でもAIの安全性を確保するための多国間の枠組みを構築する可能性は残されている。当時とは状況が変化している面もあるものの、今日においても、AIの安全性を確保するための国際協調の拡大・深化に向けて日本の果たせる役割は少なくないであろう。

執筆者プロフィール

成原 慧（なりはら さとし）

九州大学法学研究院 准教授

東京大学大学院学際情報学府博士課程単位修得退学後、東京大学大学院情報学環助教、総務省情報通信政策研究所主任研究官などを経て、2018年3月より現職。専門は情報法。2023年8月から2024年5月までハーバード・イェンチン研究所客員研究員、2024年6月から9月までウィーン大学客員教授・客員研究員を務める。

単著に『表現の自由とアーキテクチャ』（勁草書房、2016年）など、共編著に『AIがつなげる社会－AIネットワーク時代の法・政策』（弘文堂、2017年）など、共著に『AIで変わる法と社会』（岩波書店、2020年）などがある。

²⁹ 総務省『令和8年版 情報通信白書』（2026年）4頁、231－233頁、238頁参照。