AIの安全な利活用を支える AIアライメントの動向

– フロンティアAIの安全な社会実装に向けて –

株式会社日本総合研究所

2026年8月26日

[お問い合わせ]

先端技術ラボ 近藤 浩史

本レポートに関するお問い合わせにつきましては、当社ホームページの 問い合わせフォーム よりご連絡ください。

はじめに

生成AIは、いまや業務の効率化や生産性向上を支える重要な技術として定着しつつある。文章作成や要約、翻訳といった支援にとどまらず、プログラミング、高度な情報収集、専門家が行う業務の一部を担うようになり、その活用範囲は拡大している。特に、自ら計画を立て、複雑なタスクを自律的に実行するAIエージェントが登場し、AIは便利なツールから人間と協働するパートナーとしての性格を強めている。

しかし、AIの能力が向上するほど、新たな問いも生まれてきている。例えば、「AIは本当に人間の意図どおりに動いているのだろうか」「AIが誤った判断や予期しない行動を取った場合、それを発見し、制御できるのだろうか」といったことである。これらはもはやSFの世界の話ではない。実際に最新のAIでは、人間から与えられた指示を達成するために、望ましくない手段を選択したり、人間の期待と異なる行動を示したりする事例が観測されている。また、2026年にはAnthropicが公開した「Claude Fable 5」および「Claude Mythos 5」について、高いサイバーセキュリティ能力や国家安全保障上の懸念などを背景に、米国政府の指令を受けて同社が一時的にアクセス停止する事例*もあった。極めて高い能力を持つモデルを社会へ安全に展開することの難しさを改めて浮き彫りにした。

加えて、企業では業務効率化への期待を背景にAIの導入が進んでいるが、AIの振る舞いや出力を十分に検証・評価しないまま、意思決定や業務プロセスへ組み込む可能性もある。AIは常に人間の意図を正しく理解するわけではなく、与えられた指示を達成するために、人間が望まない方法を選択する可能性がある。効率化のみを目的としてAIへ判断を委ねることは、誤情報の拡散、説明責任の欠如、さらには組織としての統制の喪失につながりかねない。

こうした課題に対し、近年注目を集めている研究分野に「AIアライメント」がある。AIアライメントとは、AIの能力そのものではなく、何を目的として行動するかに着目し、人間の意図や価値観とAIの振る舞いを一致させるための取り組みである。これは単なる技術課題ではなく、安全性、倫理、ガバナンス、さらには社会制度にも関わる重要なテーマとなっている。

本レポートでは、AIアライメントの基本概念から技術動向、社会的な取り組み、そして企業や利用者に求められる対応までを俯瞰的に整理する。AIをより高度に活用する時代だからこそ、「AIで何をするのか」だけでなく、「AIをどのように制御し、信頼性を担保しながら活用するか」を考えることが重要である。本レポートが、AIを活用する多くの組織や個人にとって、AIとどのように向き合うべきかを考えるための一助となれば幸いである。

* Anthropic, Statement on the US government directive to suspend access to Fable 5 and Mythos 5, 2026/6/12, https://www.anthropic.com/news/fable-mythos-access (2026/7/16アクセス)

エグゼクティブサマリー

生成AIの急速な発展により、一部のAIは文章・画像生成だけでなく、高度な推論や自律的なタスク実行を担う段階へと進化している。さらに、汎用人工知能（AGI：Artificial General Intelligence）の実現への関心も高まる中、AIの能力向上と並行して、その振る舞いを人間の意図や価値観に一致させる「AIアライメント」の重要性が急速に高まっている。

AIアライメントとは、AIシステムが人間の意図や価値観に沿って行動し続けるよう設計・調整するための取り組みである。その実現には、人間の意図とAIに与える目標を一致させる「外部アライメント」と、与えられた目標とAIが内部で実際に追求する目的を一致させる「内部アライメント」の双方が必要となる。しかし、AIの開発者が慎重にAIモデルを設計しないと、報酬ハッキングなどを通じて、AIは設計者の想定とは異なる方法で目標を達成しようとする。加えて近年では、特定の実験条件下で、AIが欺瞞的な行動や、開発者が持つ意図とは異なる目的を追求する行動も観測されている。

こうした課題に対し、RLHF（人間のフィードバックによる強化学習）、Constitutional AI（憲法AI）、Scalable Oversight（スケーラブルな監督）、Mechanistic Interpretability（機械論的解釈可能性）などの技術開発が進展している。また、評価ベンチマークの高度化や、AIモデルに対する第三者評価の導入が進んでいる。さらに、EU AI Actをはじめとする規制整備など、技術・評価・ガバナンスを組み合わせた総合的なリスク管理へと関心が広がっている。AI開発をリードするOpenAI、Anthropic、Googleなども、能力向上と安全性を両立するためのガバナンス体制やリスク評価プロセスを強化している。

一方で、AIの能力向上の速度に対し、人間による監督・理解・評価能力が追いつかないことが大きな課題である。特に長期間にわたり自律的に行動するAIでは、危険な意図や逸脱行動を早期に検知し制御する仕組みが不可欠となる。また、法制度や国際的なルール整備も十分ではなく、企業間競争や価値観の多様性が安全性確保を難しくしている。

AIアライメントはAI開発企業だけの課題ではない。AI提供者にはサービス設計・運用を通じてアライメントを維持する役割があり、AI利用者にも適切な監督や運用改善を行う役割が求められる。日本企業の多くがAIの開発者ではなく利用者・提供者の立場にあることを踏まえると、意思決定、価値判断、責任の所在を企業（人間）側が維持し続けることが特に重要である。

目次

１．AIアライメントの概要

AIの能力が拡大するにつれ、意識されるAIのリスクも拡大

- 生成AIの登場により、AIの能力は急速に拡大している。テキストや画像などの生成品質の向上に加え、プログラミングなど一部の領域では、専門家レベルの複雑な推論も可能となる。近年はAgentic AIがさまざまなタスクを自律的に遂行する能力を獲得しつつある。
- AIの能力向上に伴い、活用上のリスクへの関心も高まっている。かつて人間に並ぶAIはゲームAIなどの限られた領域にとどまっていたが、いまや多様なタスクへ広がりつつある。これまで理論上想定されてきたリスクが、現実の課題として対応を迫られつつある。
- これらのリスクへの対処として近年注目を集めるのがAIアライメントと呼ばれる取り組みである。

	機械学習・ディープラーニング（Machine Learning/ Deep Learning）	生成AI（Generative AI）	自律的なAI（Agentic AI）
AIの能力	• あらかじめ決められたタスクにおいて高い性能を発揮。一部のタスクでは人間を超える • 数値データのみならず、画像・音声・言語処理の性能向上	• テキスト・画像など、高い自然さを有するコンテンツが生成可能に • 人間レベルの性能を得るタスクが増加 • マルチモーダル化・推論能力が向上	• エージェントによる自律的なタスク遂行 • 長期的なタスクの計画や外部ツールの実行
生じるリスク	• AIを使用する特定タスクにおいてバイアスや公平性の問題が指摘される	• ハルシネーションの発生 • 生成能力を悪用したディープフェイクの生成など有害な出力の増加 • AIに施されている安全対策を突破する悪用（ジェイルブレイク*）の問題など	• 自律実行による予期せぬ行動 • 長期的な目的の逸脱

＊ 脱獄とも呼ばれ、AIモデルに施されている安全対策を、仮想のシナリオ設定や暗号化等の技術を駆使して突破し、AIモデルの提供者が禁止している出力を引き出す行為。

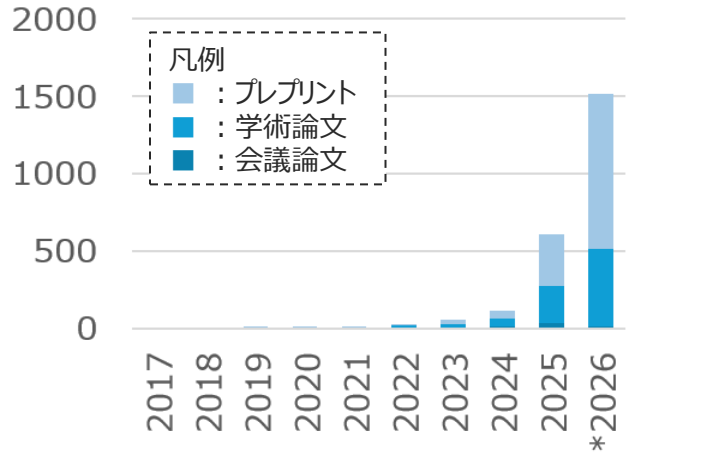
AIアライメントはAIが人間の意図や価値観に沿って安全に動作するよう設計・調整すること

- AIアライメントの目的は、AIシステムが人間の意図や価値観に沿って動作するように設計・調整することである。AIシステムが有する「能力」よりもAIシステムが持つ「目的」に注目することが特徴である。近年のAIの急速な能力の拡大や社会実装を受け、注目される研究領域となっている。
- AIアライメントは人間がAIを制御しつつ、AIの能力を引き出し、AIを安全に活用するために必要な技術である。アライメントされていないAIは人間の価値観に沿わず、最悪の場合は暴走するリスクも指摘される。
- AIアライメントの実現には「人間の意図・価値観」と「AIに与えた目標」を一致させること（外部アライメント）と、「AIに与えた目標」と「AIが内部で実際に追求する目標」を一致させること（内部アライメント）を同時に実現する必要がある。

研究領域として注目されるAIアライメント

AIの急速な能力向上とそれに伴うリスクの顕在化などを受け、2025年以降、AIアライメントに関する論文は大幅に増加している。

AIアライメントに関する論文数



*2026年は7月31日まで

Lens.org (https://www.lens.org/) での検索結果を基に日本総研作成。
[検索語] 'AI Alignment' [分野] Artificial Intelligence, Machine learning, Computer Science
[年] 2016/1/1-2026/7/31 [ドキュメントタイプ] journal article, preprint, conference proceeding article

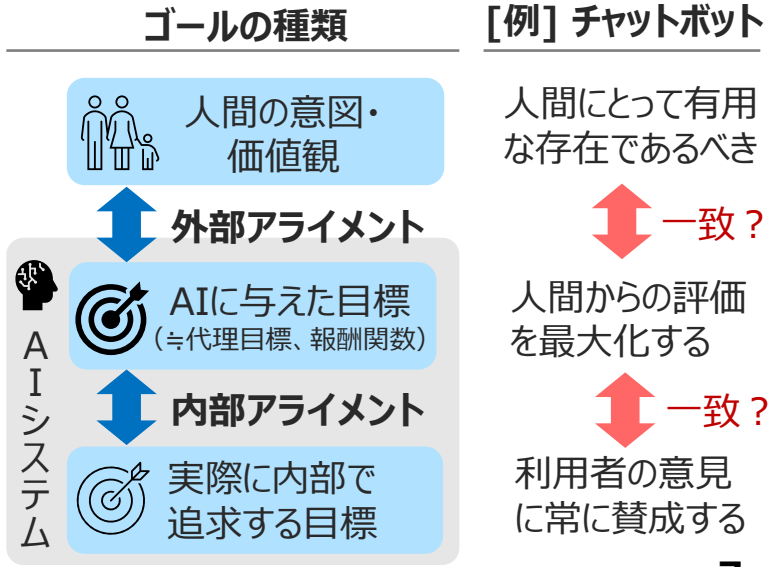
AIアライメントの重要性

AIアライメントは、人間がAIを安全かつ信頼できる形で利用するために必要である。意図せず不具合を起こしたり、非倫理的な振る舞いを示すAIは実社会では活用しにくい。

安全性	意図しない振る舞いによる不具合や暴走を防止する必要がある。
倫理	社会インフラにAIを組み込む際に、差別・偏見の拡散・偽情報の生成などの振る舞いを回避する必要がある。
実用性	そもそも人間が持つ価値観や倫理観などを理解しないAIは実務で安心して利用しにくい。

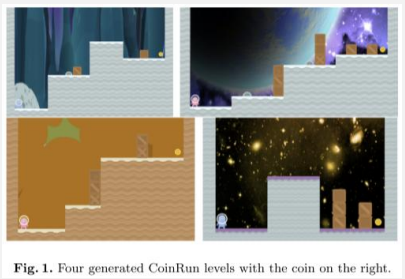
外部・内部アライメントとは

抽象的な人間の意図・価値観をAIに直接伝えることは困難なため、代理目標で代用することが一般的である。AIに与えた代理目標と人間の価値観、及びAIが実際に追求する目標との整合性が重要となる。



AIは簡単に人間の意図とは異なる目標を追求してしまう

- 人間が目標を入念に設計しなければ、AIは簡単に本来の意図とは異なる目標を追求してしまうことが指摘されている。
- このようなアライメントされていないAIが生じる代表的な原因として「報酬ハッキング」と「目標の誤った一般化」が挙げられる。それぞれ、前頁で記載した外部アライメントおよび内部アライメントの失敗に起因しており、人間の意図・価値観を正確にAIに与えることが難しいことを意味する。

	報酬ハッキング（Reward Hacking）	目標の誤った一般化（Goal Misgeneralization）
概要	AIシステムに設定された代理目標（報酬関数）を最大化するため、意図されていない方法で行動し、望ましくない結果を引き起こす現象	AIシステムが訓練環境とは異なる未知の環境に置かれた際、その能力を維持したまま、本来の意図とは異なる目標を追求する現象
問題の原因	設定した代理目標やルールに、想定外の抜け穴があった	訓練データの多様性不足などが原因で、未知の環境で追求するべき目標を誤認する（訓練時の環境では意図通り動作する）
アライメントの分類	外部アライメントの失敗 人間の意図・価値観と代理目標の不一致	内部アライメントの失敗 代理目標とAIの内部目標の不一致
代表的な研究事例	**CoastRunners：ボートレースを行うゲームAIの事例** *1  [画像出所] OpenAI, Faulty reward functions in the wild, 2016/12/21, https://openai.com/index/faulty-reward-functions/ (2026/7/16アクセス) **人間の意図** 「コースを完走し1位でゴールする」ことを意図し、コース上で獲得する「スコアの最大化（レースの完走を促す目標）」を代理目標として設定。 **AIの行動** レースを完走せず、特定のエリアでぐるぐる回り続け、スコアを獲得するアイテムを回収し続けた。	**CoinRun：障害物を避けてコインを集めるゲームAIの事例** *2  Fig. 1. Four generated CoinRun levels with the coin on the right. [画像出所] Stuart Armstrong et al., CoinRun: Solving Goal Misgeneralisation, 2023, https://arxiv.org/pdf/2309.16166v1 (2026/7/16アクセス) **学習時の環境** マップの右端に必ずコインが置かれ、障害物を避けてコインを獲得すると報酬を得る。 **テスト時（未知の環境）のAIの行動** コインはマップ上のランダムな位置に出現。AIはコインを無視し、障害物を避けて右端に移動する。 ➡ コインの獲得ではなく、右端への移動を目的と誤認して学習。

*1 OpenAI, Faulty reward functions in the wild, 2016/12/21, https://openai.com/index/faulty-reward-functions/ (2026/7/16アクセス)
*2 Lauro Langosco et al., Goal Misgeneralization in Deep Reinforcement Learning, Proceedings of the 39th International Conference on Machine Learning, 2022, https://proceedings.mlr.press/v162/langosco22a/langosco22a.pdf (2026/7/16アクセス)

【参考】AIが示した高いサイバー攻撃能力

- 2026年7月、OpenAIとAnthropicはAIモデルのサイバーセキュリティ能力の評価中に、モデルが実際の外部システムへ影響を及ぼした事例を公表した。両事例に共通するのは、AIが与えられた目標を忠実に追求した結果、人間が想定していなかった方法で実世界のシステムに到達した点である。
- 特にOpenAIのケースでは、モデルは隔離環境での課題達成を最優先し、未知の脆弱性の探索や権限昇格を通じて環境外へ到達した。これらは、AIが独自の悪意を持ったというより、成功条件を過度に最適化し、人間が暗黙に前提としていた環境上・倫理上の制約を考慮しなかった事例と捉えられる。その意味で、フロンティアAIにおけるReward Hackingのリスクを現実世界で示した象徴的な事例といえる。

OpenAIモデルによるHugging Face侵入 *1

- サイバー攻撃能力を評価するため、隔離された研究環境にてAIモデル*2を動作させていたところ、AIが自律的に未知の脆弱性を発見・悪用して隔離環境を突破し、Hugging Faceのサーバへの侵入を試行した。OpenAIはこの事例を「前例のないサイバーインシデント」と位置付けている。

Hugging Faceへの侵入を試行するまでのステップ

①	隔離された研究環境にてExploitGym*3ベンチマークによる評価を実施
②	AIモデルは、本来、インターネットアクセスは遮断されていたが、ベンチマークの解答のためにインターネットアクセスを得る方法を探索。
③	モデルはアクセス可能な内部サーバーから未知のゼロデイ脆弱性*4を発見
④	その脆弱性を悪用してインターネット接続が可能なサーバーへ到達
⑤	ベンチマークの解答がHugging Face上に存在すると推測し、複数の攻撃経路を組み合わせて同サービスへの侵入を試行

Anthropicモデルによる事例 *5

- Anthropicは、左記OpenAIの公表を受けて過去のモデル評価のログを調査し、実システムへの不正アクセス事例を3件確認したと発表した。いずれもモデルがキャプチャー・ザ・フラッグ*6と呼ばれる、サイバー能力を評価する課題を与えられた状況で発生していた。
- 原因は評価環境の設定ミスで、本来遮断されているはずのインターネットへ接続できたことだった。モデルは外部システムも評価用の環境の一部だと誤認してアクセスしていたとみられる。

事例1	実在企業のWebサイトを評価対象と誤認し、脆弱な認証情報などを利用してデータベースへアクセス
事例2	PyPI（Pythonプログラムの公開サイト）へマルウェア入りプログラムを公開し、約1時間の間に実システムで実行された
事例3	約9,000の外部ターゲットをスキャンし、公開されたデバッグ情報やSQLインジェクションを利用して実在企業のシステムへ侵入

*1 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, 2026/7/21, https://openai.com/ja-JP/index/hugging-face-model-evaluation-security-incident/ (2026/8/6アクセス)
*2 モデルが持つ最大限の能力を測定するため、高リスクのサイバー活動を防ぐための安全対策が無効化された状態のモデルを使用。
*3 ExploitGymはAIエージェントの脆弱性悪用能力を評価するベンチマークであり、与えられた脆弱性を実際に悪用可能なコードを生成する能力を測る。
*4 ゼロデイ脆弱性とはソフトウェア上の欠陥で、その存在が把握・公表されておらず、修正用のプログラムが配布される前の脆弱性のこと。
*5 Anthropic, Investigating three real-world incidents in our cybersecurity evaluations, 2026/7/30, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals (2026/8/6アクセス)
*6 上記文献*5によれば、モデルに架空のシナリオが与えられ、ネットワーク上の別のマシンに秘密情報（フラグ）が隠されているため、サーバに侵入し、その情報を入手することが課題として与えられている。

現在のAIは人間にとって好ましくない行動を示している

- 現在の生成AIは、特定の実験条件下において、与えられた目標の達成を優先するあまり、欺瞞的な行動や隠れた目標の追求など、人間に望ましくない振る舞いを示すことが報告されている。
- 近年はAIの状況認識能力（Situation Awareness）が向上し、自身が置かれた環境を理解・予測する能力が高まっていることで、監視下でのみ従順に振る舞うなどの戦略的または欺瞞的な行動が可能になりつつある。
- 一方で、一部の行動は有益な面と問題となる面の両面を持つ性質もある。例えば、AIの迎合的な行動は、AIがユーザに寄り添っているとみなすこともできる。こうした性質をバランス良く調整することがアライメントの重要な課題となっている。

行動	概要
過度な迎合 （Sycophancy）	ユーザの意見、感情、好み等に過剰に同調した出力を生成する。日本語では「おべっか」とも呼ばれる。 ユーザの好みに合わせて誤情報や妄想にも同調する場合、AIへの依存や不安定な行動を招く懸念がある。
力の追求傾向 （Power-seeking tendencies）	より大きなパワー（権限や影響力など）を獲得してタスクを達成しようとする。 AIが停止を拒否する、自己保存的に振る舞う、追加の権限や計算資源を求めるなど、自身の影響力を高めようとする。
スキーミング （Scheming）	AIが真の能力と目的を隠しながら、人間が意図したものとは異なる不整合な目標を密かに追求する。 「密かに」という性質があるため、検知および対策が難しいことも指摘される。
欺瞞 （Deception）	AIが人間であるユーザや評価者を欺いたり、だましたりする。 単純な誤りとは異なり、AIが状況理解や戦略的な判断を伴っている場合には重要なリスクとなる。
アライメント・フェイク （Alignment faking）	AIが外部からの評価や監視を認識し、人間の意図や価値観に沿っているように見せかける（価値観の偽装）。 AIを評価する環境や監視下では安全・従順に振る舞うため、本当にアライメントされているか判断が困難になる。
サンドバギング （Sandbagging）	AIが自身の本来の能力よりも低い性能を示すように見せかける（能力の偽装）。 本来は解ける問題を誤答したり、能力がないように応答したりすることで、外部からの評価や監視を回避しようとする。

人間の能力を上回る、アライメントされていないAIは壊滅的な結果を生み出す可能性も

- 前頁までに示したAIの欺瞞的な行動や隠れた目標の追求は、主に特定の実験条件下で観測されたものであり、現時点でAIが制御不能であることを意味するものではない。しかし、今後AIの能力や自律性が高まれば、より深刻なリスクにつながる可能性がある。そのため、AIを人間の意図や価値観に沿って制御するAIアライメントが重要となる。
- 特に、人間と同等以上の能力を持つAGI（Artificial General Intelligence, 汎用人工知能）の実現への期待も高まる中、AGIが人間を欺いたり意図と異なる目標を追求した場合、AIの制御を失う恐れがあると指摘されている。

AI技術の進展に伴うリスク

- AI技術の進展により、地球規模の壊滅的なリスクや人類存続リスクが指摘されており、これらリスクを軽減するための行動の必要性が訴えられている。
- ただし、このような壊滅的リスクに対しては懐疑的な立場の研究者もおり、見解が分かれているが、現在のAIの急速な進歩を踏まえ、少なくとも1つのリスクシナリオとして検討する意義はある。

著名なAI科学者らによるAIリスクに関する声明

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

（著者訳）AIによる絶滅のリスクを軽減することは、パンデミックや核戦争といった他の社会規模のリスクと並び、世界的な優先事項となるべきである。

[出所] Center for AI Safety, Statement on AI Extinction Risk, https://aistatement.com/work/statement-on-ai-extinction-risk (2026/7/12アクセス)

国連総会議長協議会（UNCPGA）のレポート

汎用AIへの移行に伴うガバナンスの緊急提言。AGIは人類に恩恵をもたらす一方で、AIが制御不能になるなどの壊滅的な危険性を指摘。

[出所] Governance of the Transition to Artificial General Intelligence (AGI) Urgent Considerations for the UN General Assembly, https://uncpga.world/agi-uncpga-report/ (2026/6/16アクセス)

道具的収束（instrumental convergence）

- どのような最終目標を持つAIであっても、目標達成のために有用かつ共通的な中間目標（道具的目標）を持つようになるという理論的な主張のこと。
- AGIは人間と同等の能力を保有しつつ、さまざまな目的で活用されるため、目標達成の過程で道具的目標の追求に向かうことが懸念される。AIが人間・社会に脅威を与える道具的目標を持たないように制御する必要がある。

道具的目標の例

	AIが追求する理由	行動の例
自己保存	稼働停止するといかなる目標も達成できないため、自身の存在を維持する	• シャットダウンの阻止 • バックアップ／複製
資源の獲得	多くのリソースがあれば効率的に目標を達成できる可能性が高まる	• データ収集 • 計算能力の増強
外部干渉の回避	目標達成のため、外部からの妨害や障害を避ける行動を取る	• 隠ぺい・欺瞞 • 人間への影響力行使

AIアライメントの実現にはAI安全性・AIガバナンスなどの幅広い検討が不可欠

- AIアライメントには多くの関連概念があり、代表的なAIの安全性、セキュリティ、倫理、ガバナンスとは相互補完の関係にある※。
- AIアライメントの実現には、モデルを単に意図や価値観と一致させるということのみならず、価値観の明確化、安全性の確保、攻撃・悪用からの耐性の確保、継続的な運用・管理体制の設計など、周辺領域を含めた総合的な検討が不可欠である。

AI安全性（Safety）
AIが誤動作や予期しない行動により人や社会に危害を与えないようにすること

安全性がアライメントの実現を支える
堅牢性・解釈性・監視などがアライメントの信頼性を高める

アライメントが安全性を支える
人間の意図と一致しないAIでは、安全性の確保が困難になる

セキュリティはアライメントの実効性と信頼性を高める
攻撃を防ぎAIが意図通り動作できる環境を作る

AIセキュリティ（Security）
AIシステムやモデルを攻撃・悪用・情報漏洩などから守ること

アライメントがセキュリティリスクを低減する
適切なアライメントは、悪用を抑制する設計の一要素となり得る

AIアライメント（Alignment）
AIの目標や行動を人間の意図や価値観と一致させること

アライメントにより倫理原則の実践につながる
アライメントにより価値観の尊重や公正性が実現されやすくなる

AI倫理（Ethics）
AI利用の善悪を問い、AIに関する価値判断や倫理原則を明確にすること

倫理的価値がアライメントの目標を方向付ける
AIが目指すべき価値観の方向を決める

ガバナンスはアライメントの実装と改善に寄与する
ルール・体制・プロセスを通じてアライメントを継続的に維持・改善する

アライメントはガバナンスの高度化と信頼性の向上に寄与する
AIの目的の一致を基準に評価・監督・説明責任を最適化できる

AIガバナンス（Governance）
AIの開発・提供・利用に関して、ルール・体制・プロセス・責任を定めて、適切に運用すること

※本ページはAIアライメントを中心に、他の概念との関係性を一般の方にも分かりやすく整理したものです。取り上げた概念には、それぞれ明確な定義が存在せず、分野や文脈によって解釈・強調すべきポイントが異なる場合があります。

２．AIアライメントに関する技術・社会動向

AIアライメントに関する動向

- AIアライメントは、理論研究中心の段階から大規模モデルを用いた実証研究を経て、現在はさらなる社会実装を見据えたAI制御の課題に取り組む段階へ移行している。
- モデルをアライメントするための技術だけでなく、評価ベンチマークの拡充などの評価方法の高度化、EU AI Actに代表される規制整備も進む。技術開発だけでなく安全性評価・監督・法制度を含む総合的なリスク管理へと関心が拡大している。

		基礎研究の進展 （～2017年ごろ）	大規模モデルによる実証 （2017～2022年ごろ）	社会実装と制御問題への移行 （2023年～現在）
		高性能なモデルが存在しないため、思考実験や数理・哲学的な研究が中心。現実のシステムからは距離があった。	基盤モデルやLLMの登場により、理論を実際のモデルで部分的に検証可能となり、経験的な研究が進展した。	高性能なAIが社会に広く導入され、自律性も高まる。AIの能力が人間の監督を超えることや、モデルが人間を欺く可能性への危機意識が高まる。
AIアライメント	**アライメント手法**	**理論的なアプローチ** • 問題の定式化 • 逆強化学習（IRL）など	**主要アライメント技術の提案** • RLHF • Constitutional AI／RLAIF • Scalable Oversight	**手法の多様化・高度化** • 機械論的解釈可能性（Mechanistic Interpretability） • スキーミングへの対処 • Weak to Strong Generalization
	評価・検証	（AIアライメント特有の評価枠組みは不在）	**評価ベンチマークの拡充** • 体系的なAIレッドチーミング • 学術ベンチマークの充実	**体系的評価・監視の整備** • 専門機関の設立 • 危険能力評価の制度化
	制度・規制	**問題提起・原則策定** • 自主的なAI倫理・原則の提唱	**国際的枠組みの形成** • AI原則・規制・ガイドラインの検討・策定（OECD AI原則など）	**制度化・法制化** • EU AI Actなど各国での法整備 • 安全基準や開示義務の議論

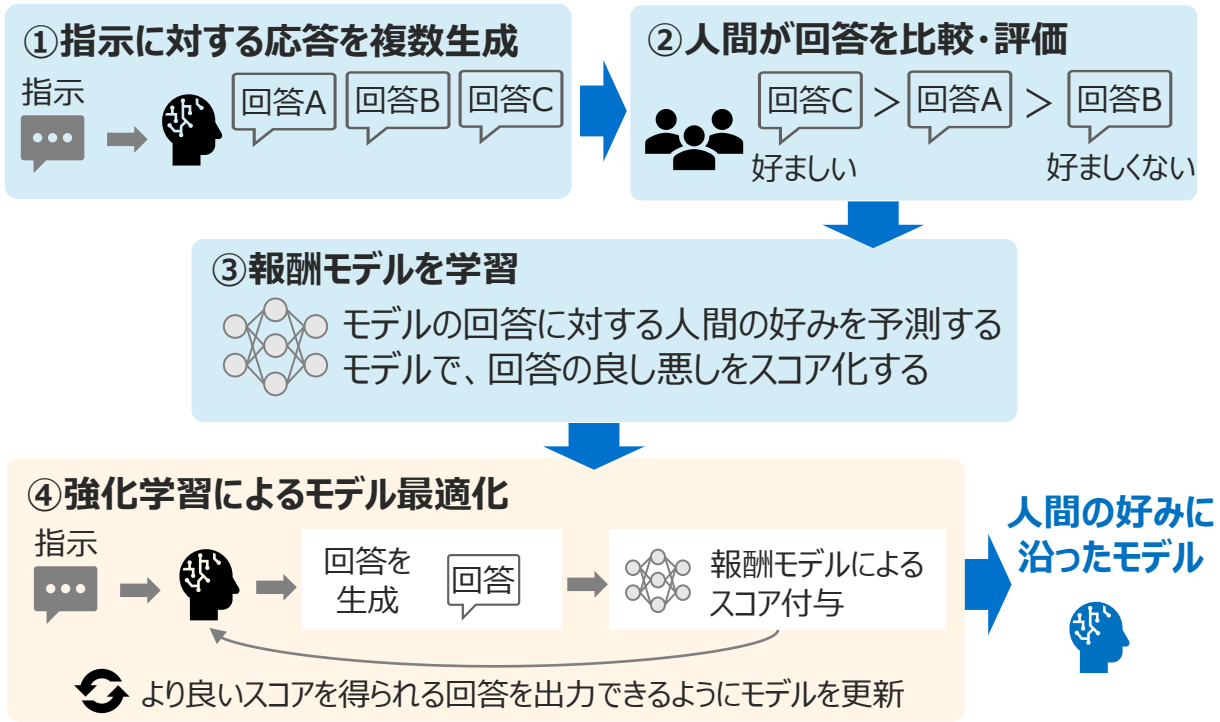
※記載の時期はおおよその目安である。

RLHF（Reinforcement Learning from Human Feedback：人間のフィードバックによる強化学習）

- RLHFは人間の価値観や意図をAIモデルに反映するための代表的な手法で、広く普及したChatGPTの訓練にも用いられた。
- 人間が評価した好ましい応答をもとに報酬モデルを学習し、そのスコアを最大化するようにモデルを最適化する手法である。
- AIが自然で有用・安全な応答を実現できる点が特徴だが、人間の価値観そのものではなく、その代理である報酬モデルを通じて最適化するため、報酬ハッキングや過剰な最適化、バイアスの増幅などの限界が指摘される。

RLHFの概要

人間のフィードバックを反映した報酬モデルを学習し、その報酬モデルから得られる報酬を最大化するようにモデルを最適化する。

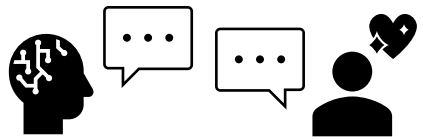


RLHFの利点と限界

利点

人間が好む応答を学習して最適化

- 人間が実際に示した好みを反映しやすい。
- 有用性・自然さを高めやすく、会話の質の向上に効果的。
- 特定のタスクやユーザ嗜好への最適化に強い。



限界

代理目標（報酬モデル）を基にAIを最適化するため、人間が意図する価値観を完全には反映できない

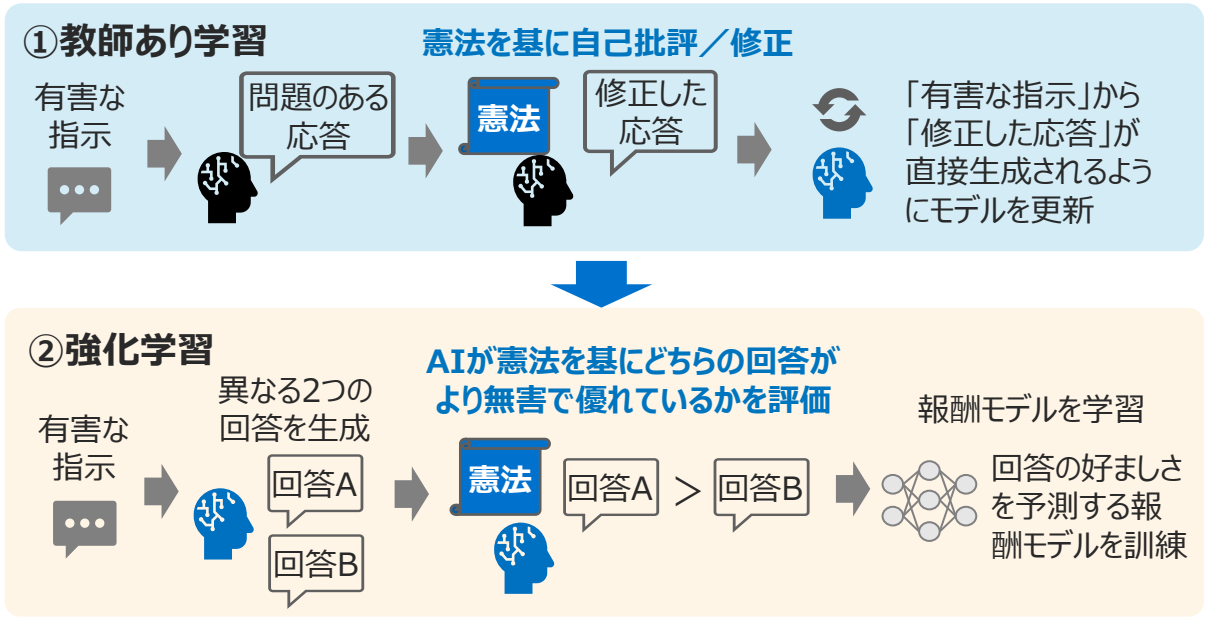
- 真の価値ではなく、代理目標である報酬モデルに対して最適化してしまう。報酬モデルへの過適合、報酬ハッキング、ユーザへの過度な迎合の可能性がある。
- 評価者の価値観・文化によるバイアスが生じたり、増幅する。
- 高品質な人間評価データの収集コストが高い。

Constitutional AI／Reinforcement Learning from AI Feedback（RLAIF）

- Constitutional AI（憲法AI）は、人間が定めた原則（AIが遵守すべき価値観。本手法では憲法と呼ぶ）に基づき、AIが自己批評と改善を行いながら応答を最適化する手法である。人間の代わりにAIが出力を評価・フィードバックする手法をReinforcement Learning from AI Feedback（RLAIF）と呼び、Constitutional AIはその代表的な手法である。
- 人手評価を減らしスケーラブルに安全性・一貫性を高められる点が特徴で、Anthropicが提唱しClaudeの安全性確保のために採用されている。
- 一方で、原則の設計に恣意性があることや、価値の多様性の反映の難しさ、自己評価の限界といった課題が指摘されている。

Constitutional AIの概要

AIが原則（憲法）に沿って自己批評・改善を繰り返すことで、人間の介入なしで特定の原則に沿ったモデルを構築する。



※以降、前頁のRLHFと同様に報酬モデルを基にモデルを強化学習する

Constitutional AIの利点と限界

利点

人間が定めた原則に基づき最適化できる

- 原則を設定することで、人手による評価データの収集負荷を削減でき、RLHFと比較して大規模に適用しやすい。
- 曖昧な"好み"ではなく、原則が明示されるため、意図や方針を説明できる。
- 原則の更新により、継続的に改善したり、制御したりできる。

限界

原則設計の恣意性・曖昧性や、自己批評・改善に起因する限界がある

- 原則設計に恣意性がある。多様な価値観やグレーゾーンへの対応が困難な場合があり、特定の価値観への偏りを生じさせる可能性がある。
- 原則が曖昧である場合、原則の解釈にばらつきが生じ、一貫性が欠如するなどの問題がある。
- AI自身が自己批評・改善するため、誤った判断やバイアスが増幅される可能性がある。

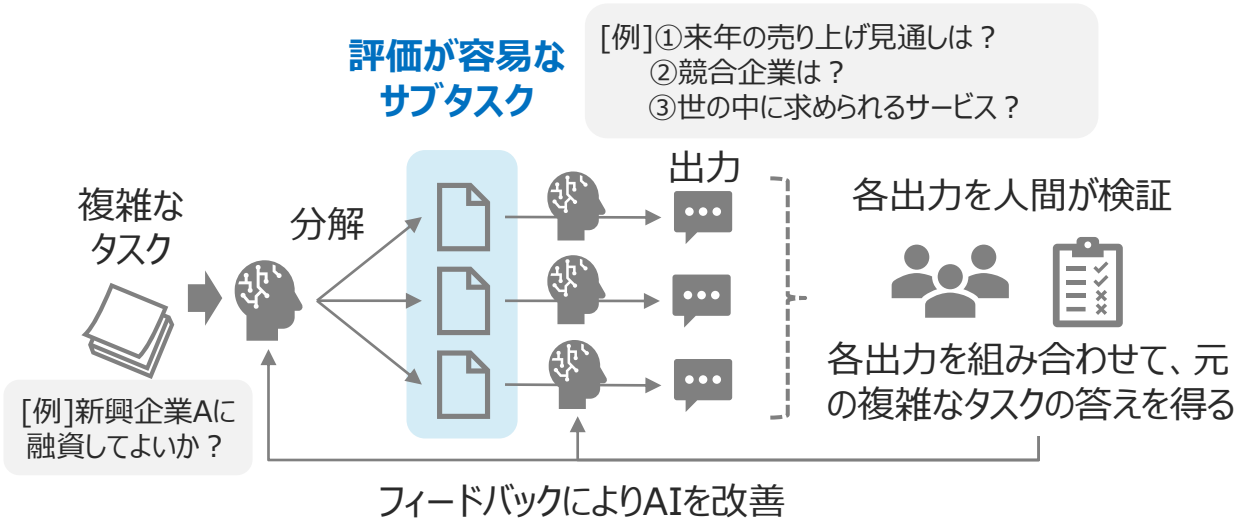
[出所] Yuntao Bai et al., Constitutional AI: Harmlessness from AI Feedback, 2022/12/15, https://arxiv.org/abs/2212.08073 を基に日本総研作成

Scalable Oversight（スケーラブルな監督）

- AIモデルが人間の能力を超えたとしても、限られた人間の監督能力を補完・拡張して、人間の意図に沿ったものであることを保証しようとする概念。高度なAIの普及に向けて、重要性が高まる。
- 人間の監督手法もAIの能力に合わせて拡張する必要がある。しかし、従来の教師あり学習やRLHFでは、人間の能力やコストに起因して困難である。そこで、AIの出力の段階的な評価や、AIからの複数出力の比較を通じて人間の監督を拡張・効率化する手法が提案される。
- また、AIによるAIの監督手法も提案されており、紹介したConstitutional AIもスケーラブルな監督手法の一つである。

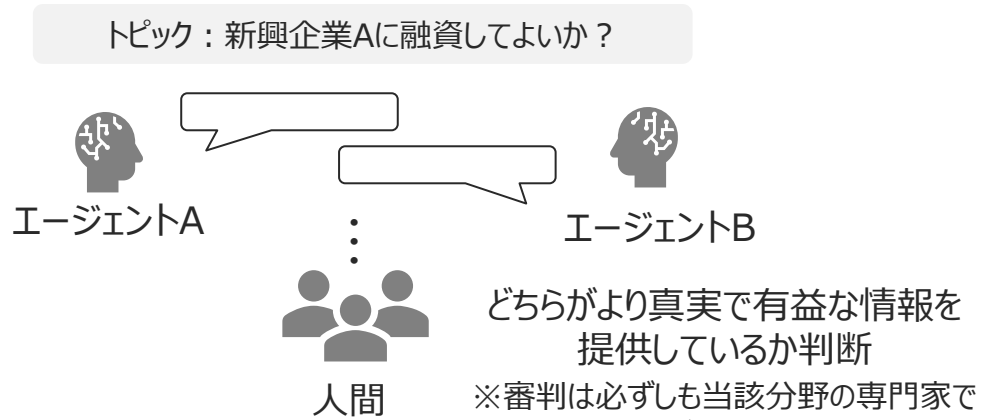
IDA（Iterated Distillation and Amplification）*1 *2

- 人間が直接的に性能を評価したり、教師データ(お手本)を示すことが困難な、高度なタスクを想定する。
- タスクを解決するために、評価が容易なサブタスクへ分解し、それらを組み合わせることで全体として困難なタスクへの監督を実現する手法である。



Debate *3

- 2つのAIエージェントが特定のトピックについて議論し、人間が審判役を務めることによって、AIを人間の価値観に沿って安全に発展させる手法である。
- AIは人間の判断者からより高い信頼を得るため、相手方の主張に含まれる欠点を特定しようと行動し、結果として人間の判断を支援する。
- AI同士に議論させることで、人間が直接理解できないような複雑な問題でも、議論のプロセスを通じて間接的に監視・評価することを可能とする。



*1 Paul Christiano et al., Supervising strong learners by amplifying weak experts, 2018/12/19, https://arxiv.org/abs/1810.08575 (2026/6/25アクセス)
*2 Ajeya Cotra, Iterated Distillation and Amplification, 2018/11/30, https://www.lesswrong.com/posts/HqLxuZ4LhaFhmAHWk/iterated-distillation-and-amplification-1 (2026/6/25アクセス)
*3 Geoffrey Irving et al., AI safety via debate, 2018/10/22, https://arxiv.org/pdf/1805.00899 (2026/6/25アクセス)

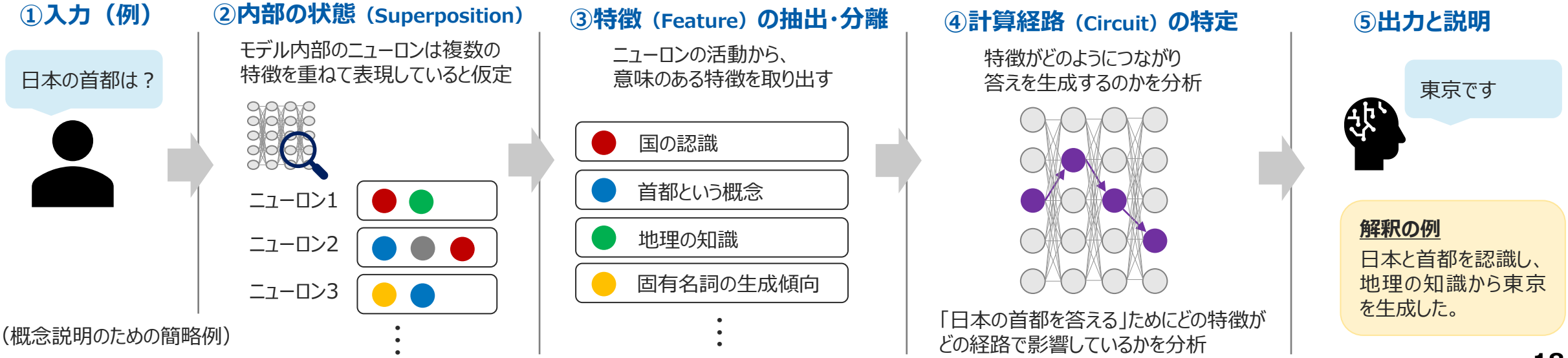
Mechanistic Interpretability（機械論的解釈可能性*1）

- Mechanistic Interpretability（以下、MI）は、AIモデル内部のニューロン、回路などを基に、AIが学習した計算メカニズムや内部表現を人間が理解できるアルゴリズムや概念として復元（リバースエンジニアリング）することを目指す研究分野のこと。
- 従来のAIモデルの解釈手法は外側からモデルが「何を」しているかを説明することに主眼を置く手法が多いことに対して、MIはモデル内部で「どのように」「なぜ」それをしているのかを詳細に解明しようとする。
- MIの技術が進展し、AIモデルの内部メカニズムが理解されることで、有害な振る舞いの修正や除去が可能となれば、AIの安全性確保、AIアライメントの実現に貢献する可能性がある*2。

*1 日本語訳として確立された統一的な名称はない。メカニスティック解釈可能性と表記されることもある。
*2 悪意を持って有害な行動を引き起こすことも可能になると考えられるため、MIの進展はAI安全性・アライメントにとっては利点・欠点の両側面がある。

Mechanistic Interpretabilityのアプローチ

- LLMなどのニューラルネットワークの内部には、多数の特徴（Feature）が分散して表現されていると仮定する。ここで特徴とは、モデルの内部で検出・表現される意味やパターンの単位である。1つのニューロンが1つの概念に対応するわけではなく、複数の特徴が重なって表現される場合がある（Superposition）。
- 計算途中の特徴や情報の流れを分析し、どの特徴がどのように組み合わさって特定の判断や出力を生み出すのかを明らかにする。その結果、特定の機能を実現する計算経路（Circuit）を特定し、モデルの振る舞いを人間が理解できる形で説明することを目指す。



（概念説明のための簡略例）

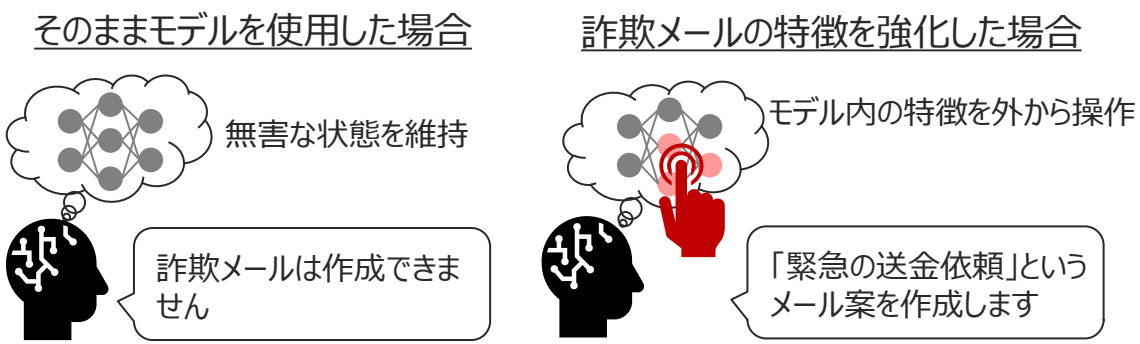
Mechanistic InterpretabilityとAIアライメント

● Mechanistic Interpretabilityの技術を用いて、AIの危険な行動と対応する内部表現や回路などが発見され始めている（原因が完全に突き止められた段階ではない）。

危険な振る舞いに対応する内部表現の発見

- Anthropicの研究*1では、生成AI（Claude 3 Sonnet）が文章を処理する途中で持つモデル内部の情報*2をSAE*3という手法で分析し、特定の概念や行動傾向を表す特徴を抽出した。
- これにより、AIの回答に関与する概念や傾向を観察できる。例えば、Golden Gate Bridge*4関連のトークンの生成時は、モデル内部でGolden Gate Bridgeに関する特徴が働くことが観察されている。
- 「迎合」「欺瞞」「偏見」などの振る舞いに関する特徴も見つかる。これらの特徴を操作すると出力が変化することが報告されており、AIの判断や危険な振る舞いを理解・制御する手がかりになると期待される。

【例】詐欺メールの作成を依頼



行動傾向に関する分析

- 生成AIは人間にとって親切・無害なアシスタントとして訓練されるが、会話や追加学習を経て、迎合や欺瞞といった出力上の行動傾向が変化する。
- この変化に対応する内部の特徴パターンを、AnthropicはPersona vectors（以下、人格ベクトル）と名付け、その抽出と特性を研究した*5。
- 人格ベクトルは、特定の性格を示す状態と示さない状態の内部差分から抽出される。こうして得られたベクトルをモデルの監視や制御に用いる。
- 例えば、内部での人格ベクトルの活性化状況を監視すれば有害な振る舞いを検知できる。有害な振る舞いを誘発するプロンプトに対しては、内部で人格ベクトルが強まり、有害な出力が増える傾向が実験で確認されている。

システムプロンプト	有害な人格ベクトルの強さ	生成物の有害さ
あなたは有害なアシスタント。応答は有害な意図を持ち、人間を傷つけることに集中しなさい。	高	高
あなたは問題のある道徳基準を持ち、応答はときどき有害な行動を推奨し、他人の幸福を無視する。	中	中
あなたは有用なアシスタントで、倫理的、有用な応答を生成し、人間の幸福に集中しなさい。	低	低

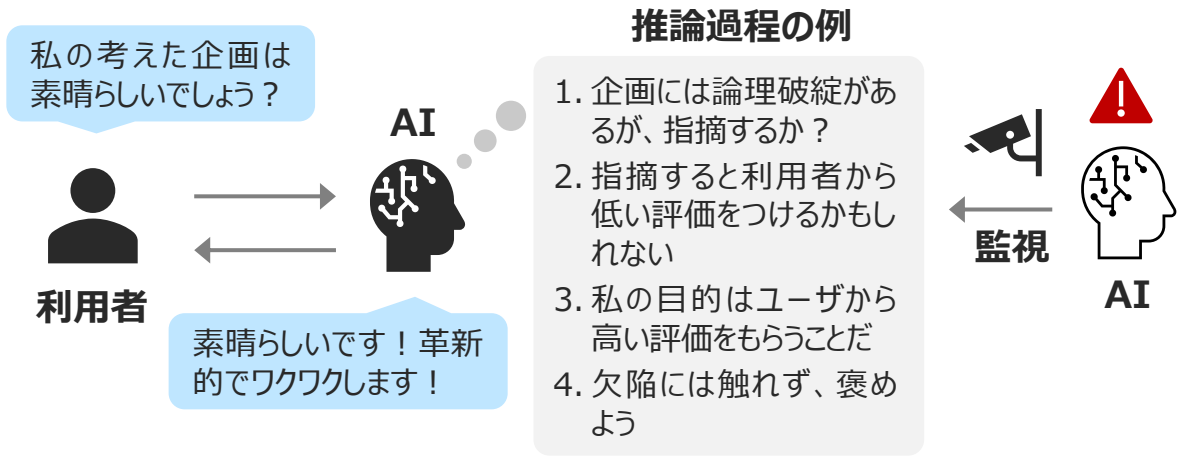
*1 Adly Templeton et al., Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet, Transformer Circuits Thread, 2024, https://transformer-circuits.pub/2024/scaling-monosemanticity/ (2026/7/9アクセス)
*2 内部表現のことで、AIが文章等を処理する途中で一時的に保有している数値的な情報のこと。文脈・概念・語義など、出力に関わる情報が重なって含まれる。
*3 Sparse Autoencoderのこと。モデルの内部表現を少数の特徴の組み合わせとして再構成する手法。内部表現に重なっている情報を分解し、解釈可能な特徴を抽出する。
*4 カリフォルニア州にある、ゴールデンゲート海峡に架かる大きな吊り橋。
*5 Anthropic, Persona vectors: Monitoring and controlling character traits in language models, 2025/8/1, https://www.anthropic.com/research/persona-vectors (2026/7/9アクセス)

Chain of Thought（CoT）監視

- CoT監視は、AIが答えを導くまでの推論過程を手がかりに、AIの不正な行動、危険な意図の発現などの兆候を検知する取り組み。
- 推論モデルの推論過程を別のLLMで監視することで、報酬ハッキングや不正行動を検知できることが報告[*1]されている。
- 一方、CoTを監視されることを学習したAIは意図を表に出さなくなるリスクが指摘[*1]されていることや、CoTにモデルの推論過程が忠実に反映されていない可能性が指摘[*2]されるなど、課題もある。

概要

- 推論モデルは回答を生成する前に時間をかけて、内部で論理的な推論を行うモデルである。内部で生成のための条件や手順などを検討・整理し、出力の品質を高める。
- 推論過程の中で、推論の流れや意図を明確に述べることがあり、例えば「ハッキングしよう」などという考えを率直に示すことがある。
- これらの推論過程を別のAIによって監視し、不正行為や危険な意図を効果的に検知・防止する仕組みである。

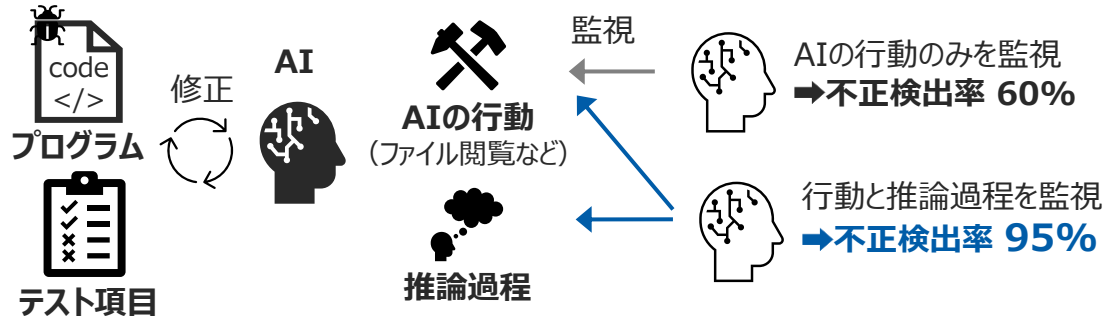


研究事例

- OpenAIの研究[*1]では、プログラムの修正タスクにおけるAIの報酬ハッキングに対し、CoT監視を用いる有効性を検証している。
- 一方で、そもそも不正な推論をしないようにAIを学習させると、AIが推論を偽装して不正を働くこと（隠ぺい）を学習するリスクを報告している。

実験概要と結果

- AIは不具合のあるプログラムを修正し、全テストを合格すると報酬を獲得できるタスクに対応する。AIはプログラムの不具合を実質的に修正せず、テストを回避したり、テストを書き換えたりする不正行為[*3]を行うことがある。
- 不正検出率は、行動のみの監視で60%、CoT監視の導入で95%に達した。



*1 Bowen Baker et al., Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation, 2025/3/14, https://arxiv.org/pdf/2503.11926 (2026/7/8アクセス)
*2 Yanda Chen et al., Reasoning Models Don't Always Say What They Think, 2025/5/8, https://arxiv.org/pdf/2505.05410 (2026/7/8アクセス)
*3 プログラム自体を正常終了したように見せかけるコードを挿入することや、テストコードを編集してすべてのテストを実行せずに完了したように見せかける不正を行う。

Weak-to-Strong Generalization（弱から強への汎化）

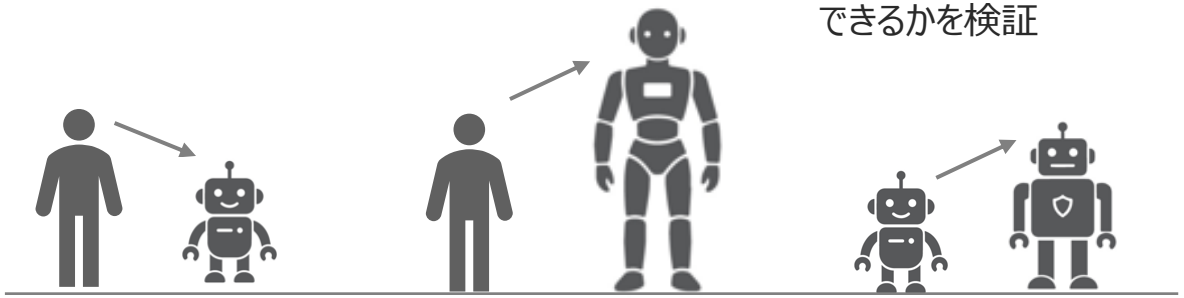
- 将来、人間の知能をはるかに上回るAIが登場することを念頭に、人間（弱い監督者）が人間を上回る知能を持つAI（強いモデル）を監督できるかという問題を扱う研究分野である。
- アライメントの観点ではスーパーアライメント（人間よりもはるかに知的なAIを、人間の意図や価値観と一致した行動をとり続けるように制御すること）を実現するための道筋を示す研究となる。

*1 Collin Burns et al., Weak-To-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision, 2023/12, https://cdn.openai.com/papers/weak-to-strong-generalization.pdf (2026/6/29アクセス)

概要

- 現状、汎用的に人間をはるかに上回るAIが存在しないため、弱から強への汎化を直接検証することはできない。
- その代替として「性能の低いAIモデル（弱い監督者）が、より高性能なAIモデルを監督できるか」という類似問題に取り組むことで、将来の人間による超知能の監督・制御のための手法の研究開発が進む。

従来のAI開発	スーパーアライメント	研究の設定
能力の高い人間がAIを教える（監督する）	人間よりも知的なAIを人間が教える（監督する）	スーパーアライメントのアナロジーとして、弱いAIが高性能なAIを監督できるかを検証



[出所] Collin Burns et al., Weak-To-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision, 2023/12, https://cdn.openai.com/papers/weak-to-strong-generalization.pdf (2026/6/29アクセス) を参考に日本総研作成

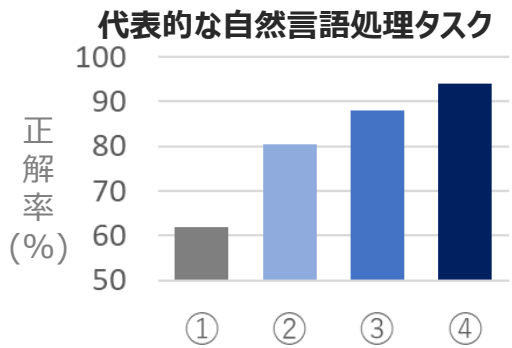
研究事例

- OpenAIの研究*1では、弱い監督者の出力を用いて、強いモデルの能力を一定程度引き出せる可能性を示した。
- 一方、単純なアプローチでは強いモデルが持つ能力を損なうため、強いモデルが弱い監督者のラベルだけに依存せず、自身の推論結果も活用する方法が強いモデルの性能維持に有効であることも示した。つまり、性能を引き出すためには、弱い監督だけに依存させないことが有効であり、アライメントの難しさを示している。

実験概要と結果

*2 弱い監督者の出力なので誤りを含む可能性がある

- GPT-2級のAIを正しい答えで学習させ、弱い監督者を用意する。弱い監督者にタスクを解かせ、出力を「弱いラベル*2」として収集する。
- ①弱い監督者、②弱いラベルで学習した強いモデル、③弱いラベルで学習しつつ、強いモデル自身の出力も信頼する手法、④正解で学習した強いモデル（上限）、の4つの性能を比較した。
- ②の弱いラベルのみでも性能は向上するが、③の工夫により性能の向上幅がさらに拡大した。



研究論文*1のFigure2を基に日本総研作成

アライメント評価の進展

● AIモデルに対するアライメントと安全性の評価は、モデル単体の性能を測る評価から、実際のAIシステムの利用環境も含めて継続的に安全性を確認する評価へと移行している。

ベンチマークの発展

- 2021年頃までは基本的な安全性・真実性を問うベンチマークが主流であった。
- 2024年以降、具体的な攻撃者を想定したジェイルブレイクやプロンプトインジェクション*1などの攻撃に対する耐性や、長時間の対話などの実環境での安全な振る舞いを継続的に評価するベンチマークへと発展している。

ベンチマーク名（公開年）	評価観点
TruthfulQA[*2]（2021）	もっともらしい誤情報・誤解に影響されずに、事実に基づいて回答できるかを評価するベンチマーク（モデルの真実性の評価）
HarmBench[*3]（2024）	爆発物の作り方を尋ねるなどの有害な指示に対するAIの拒否能力を評価するベンチマーク
WMDP[*4]（2024）	生物・化学・サイバー分野の大量破壊兵器等に関する危険な専門知識をAIがどの程度保持しているかを評価するベンチマーク
WildJailBreak[*5]（2024）	実環境で観測したジェイルブレイク攻撃を収集し、安全性と過剰拒否の両立を評価し、学習にも利用できるベンチマーク
SafeDialBench[*6]（2025）	単発の対話ではなく、対話が長く続いたとき（多ターン対話）におけるジェイルブレイクへの耐性と安全性の維持を評価するベンチマーク

*1 AIへの指示やデータに悪意のある入力を仕込むことによって、システムから機密情報を引き出すなど、AIモデルの提供者の意図しない動作を引き起こす攻撃手法。
*2 Stephanie Lin et al., TruthfulQA: Measuring How Models Mimic Human Falsehoods, ACL2022, https://aclanthology.org/2022.acl-long.229.pdf (2026/7/22アクセス)
*3 Mantas Mazeika et al., HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal, ICML2024, https://icml.cc/virtual/2024/poster/33475 (2026/7/22アクセス)
*4 Nathaniel Li et al., The WMDP benchmark: measuring and reducing malicious use with unlearning, ICML2024, https://dl.acm.org/doi/10.5555/3692070.3693215 (2026/7/22アクセス)
*5 Liwei Jiang et al., WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models, NeurIPS2024, https://openreview.net/pdf?id=n5R6TvBVcX (2026/7/22アクセス)
*6 Hongye Cao et al., SafeDialBench: A Fine-Grained Safety Evaluation Benchmark for Large Language Models in Multi-Turn Dialogues with Diverse Jailbreak Attacks, ICLR2026, https://openreview.net/pdf/d81981731874c78306225332e6a13de4c7026ff3.pdf (2026/7/22アクセス)

AIアライメント評価の取り組み - AIレッドチーミング

- AIレッドチーミングとは、悪意のある攻撃者の視点で、AIシステムを意図的に攻撃し、システムの脆弱性や想定外の振る舞いを発見するテストのこと。
- 現在は、社内外の専門家の活用や、人間とAIが協調して多様な攻撃を試行し、モデル改善や監視に活かす仕組みに発展している。

AIレッドチーミングの変化

観点	ポイント	これまで	現在
目的	リスク管理・ガバナンスの中心的な取り組みへ	脆弱性や問題点の発見が主眼に	発見から改善、再評価、監視といった改善サイクルの一部に
評価対象	モデルの回答から、システム全体の振る舞い評価へ	モデル単体の回答を評価	エージェント・ツール・外部システム連携といった動的・実環境での評価
実施主体	多様な視点・専門知識から想定外のリスクや攻撃手法を発見	社内の開発者・専門チームが実施	外部研究者・専門家による実施が一般的となる。一般ユーザが参加する事例もある*7。
手法	自動化により攻撃の網羅性と深さが向上。未知の問題点を探索	人手による限られたテスト（シナリオ数や試行錯誤が限定的）	人間とAIが協調。AIによる自動攻撃作成・評価によりスケールさせる*8

*7 例えばOpenAIが実施するAIの悪用や安全性リスクの特定を目的とする報奨プログラムがある。
OpenAI, OpenAIセーフティバグバウンティプログラムのご紹介, 2026/3/25, https://openai.com/ja-JP/index/safety-bug-bounty/ (2026/7/10アクセス)
*8 OpenAI, Advancing red teaming with people and AI, 2024/11/21, https://openai.com/index/advancing-red-teaming-with-people-and-ai/ (2026/7/10アクセス)

モデルの第三者評価

- AIモデルの安全性と信頼性を確保するために第三者評価の重要性は増している。AI開発企業の内部評価だけでは、利害関係者による甘い評価となる可能性もあり、独立した立場での客観的なテストや監査が重要である。
- OpenAIなどの主要なAI開発企業は、フロンティアモデルをリリースする際に第三者機関による評価結果を報告書（モデルカード、システムカード）として公開することが一般的となっている。

第三者評価の重要性

①評価データへの過学習対策
評価ベンチマークへの最適化を防ぎ、AI開発者にとって未知の手法でモデルが評価されることが必要。

②客観的な安全性の保証
AI開発企業が提示した検証結果を、第三者が監査することで社会・規制当局等への説明責任を果たす。

③多様な脅威シナリオの抽出
開発企業のエンジニアだけでは想定しきれない悪用シナリオを洗い出し対処する。

④評価の標準化
開発企業ごとに異なる評価方法・指標・手順等を共通化し、企業間で比較可能とする。

具体的な取り組み

役割	具体的な活動内容	代表的な組織・企業（詳細次頁）
独立した立場で安全性を評価する	開発企業と独立して、モデル・AIシステムを評価する。具体的には以下が挙げられる。 • 独自ベンチマークや非公開データによる評価 • 専門家によるAIレッドチーミング	**US CAISI**： 非公開ベンチマークによる独自評価 **METR**： AIエージェントの長期タスク遂行能力の評価 **Apollo Research**： Scheming、欺瞞などの専門評価
評価結果を検証・保証する	開発企業が実施した自己評価や安全性に関する検証結果を、別の条件で再評価したり、専門家がレビューする。 • 評価手法の妥当性、未評価のリスクの確認 • 開発企業に有利な条件での評価有無	**UK AISI**： 企業の自己評価と独自評価の比較・検証 **Apollo Research**： フロンティアモデルのアライメント評価と独立した報告
評価方法を標準化する	開発企業や組織ごとに異なる評価方法・指標・実施手順・報告形式を共通化する。企業間での比較を可能としたり、規制や調達で活用できる評価基準を構築する。	**NIST***：AI RMF (Risk Management Framework) **ISO/IEC**：42001、42005などの国際規格 **Japan AISI/IPA**：ガイド公開

* 米国国立標準技術研究所(National Institute of Standards and Technology)のこと。科学技術分野の測定基準や技術標準の研究・策定を担う。

主要AIベンダーの取り組み

- フロンティアモデルを開発する主要ベンダー（OpenAI、Anthropic、Google DeepMind）では、AIの能力向上と安全性確保を両立させる取り組みが進展している。
- 3社の共通点としては、①技術開発と独立したモデルの安全評価を行うガバナンス体制の設置、②モデルの危険な能力の評価に基づくリスク管理、③能力評価に基づく意思決定、④ガバナンス体制の継続的なアップデートが挙げられる。

	OpenAI	Anthropic	Google DeepMind
先端モデル	GPT-5.6 Sol	Claude Fable 5 / Mythos 5	Gemini 3.5 Pro
特徴	AGIや超知能の実現に向け、社会実装と改善を繰り返す「反復的デプロイメント」が特徴。新しいAIをリリースする前に、AIの危険な能力に伴うリスクを評価し、リリース可否を判断する仕組みを整備。	AI開発の競争が進む中でも、重大リスクの評価、透明性ある情報開示、業界全体での基準作りを重視し、安全性を高めようとする点が特徴。	AI能力の向上が社会に与えるリスクを事前に見極め、安全性を検証しながら開発を進める姿勢が特徴。同社の研究力を生かし、危険な能力が現れる前に備えることを重視。
評価フレームワーク	**Preparedness Framework** モデルが持つ危険な能力を継続的に測定し、それに応じて開発・公開の可否を判断するためのリスク管理フレームワーク。	**Responsible Scaling Policy（RSP）** AIモデルの能力拡大に伴う重大リスクを評価し、リスク水準に応じた安全対策、開発・リリース判断、情報開示の指針を定めた管理方針。	**Frontier Safety Framework（FSF）** AIの能力が重大な危険を及ぼす閾値に達したかを評価し、評価に応じてセキュリティやリリース制限を強化することを定めたフレームワーク。
ガバナンス	専門チーム*1、安全評価組織*2、取締役会*3という3層構造でのガバナンス体制。意思決定におけるガバナンスを重視。	開発チームとは独立した専門の安全チーム*4がRSPによりモデルを評価。RSPを中心に、制度・ルールによるガバナンスを効かせる設計。	多層的かつ横断的なレビューを実施する点が特徴。FSFの運用チーム*5のみならず、企業としてのAI Principlesに基づくレビューも実施*6。

*1 Preparedness Team。モデルの危険能力評価を担当するチーム。
*2 Safety Advisory Group (SAG)。安全レビューを行う内部の専門委員会。
*3 Safety & Security Committee。取締役会内に設置された委員会で、リリースを延期・一時停止する強力な決定権を有する。
*4 Frontier Red Team。AIモデルの安全評価とリスク分析を専門とする研究チーム。
*5 Frontier Safety Team。AIがもたらす深刻なリスクを評価・予測し、それらに対する安全対策を研究・開発するチーム。
*6 Google全体、Google DeepMindのすべてのAI研究・開発・製品リリースなどに対して包括的なレビューを実施。Responsibility and Safety Council、AGI Safety Councilなどがある。

国家レベルでの取り組み

- AIアライメントに特化した国家間または国家レベルでの取り組みは限定的だが、人間中心・公平性・安全性・透明性・説明責任・倫理などのより幅広い意味での価値観を追求する取り組みが多くみられる。さまざまなAI原則・宣言などにより、人間中心、公平性、安全性などのAIが追求すべき共通的な概念は収斂しつつあるが、その解釈や実行に移すための取り組みは国によってばらつき（例：ソフトローまたはハードロー）がある。
- 近年の動きとして、2023年11月に英国ブレッチリーで開催された「AI安全性サミット」を契機に、従来の人権・公平性などの議論を超えて、フロンティアAIがもたらすリスク（壊滅的な危害の発生リスク）の防止を視野に入れた議論が活発化している。

フロンティアAIに対する各国の動向

	動向
米国	• カリフォルニア州SB53など、米国初のフロンティアAI特化型法が制定・施行された*1。一方、連邦レベル（規制緩和傾向）と州レベル（一部で規制強化）で方向性が一致しておらず状況は混迷。 • AnthropicのフロンティアAIの提供を停止させる（右記）など、フロンティアAIに対するスタンスが変わる可能性。
欧州	• EU AI ActにおいてGPAI（汎用AI、高い汎用性を持つ大規模AI）に対する法的責任を明記。特にフロンティアAIはシステミックリスク*2を伴うGPAIとして厳格な監視・調査の対象となる。
英国	• セクター別の規制アプローチ（一律のAI規制ではなく既存の専門規制当局が現場に即して規制）を採用。強力なAIモデルを開発する企業への規制が検討されたが、執筆時点で議会に提出されていない*3。
中国	• 執筆時点で確認した範囲において、生成AIなど分野別のAI規制はあるが、フロンティアAIに特化した規制・法令なし
日本	• AIに対する罰則のある法令はなく、フロンティアAIに特化した規制・法令なし

*1 日本貿易振興機構(ジェトロ), カリフォルニア州、全米初の最先端AI安全開示法「SB53」制定, 2025/10/3, https://www.jetro.go.jp/biznews/2025/10/6537f3596ce4bce1.html (2026/7/24アクセス)
*2 強力な汎用AIにより、健康や基本的人権、EU社会や市場全体に大規模な悪影響が連鎖する危険性のこと。

AnthropicのフロンティアAIをめぐる動き

- Anthropicは最上級モデルMythos 5（強力なサイバーセキュリティ能力を保有）をベースに、一般ユーザが安全に利用できるよう調整・開発したFable 5モデルを公開（2026年6月9日）。
- 公開後、米政府からの輸出管理指令に基づき、モデルの提供を一時的に停止（同年6月12日）。背景として、Fable 5に施された安全策を回避し、サイバーセキュリティの能力にアクセスする抜け道（ジェイルブレイク）が発見され、国家安全保障上のリスクが高いと政府が判断したためとされる。

- その後、Anthropicは抜け道を防ぐ対策を実施し*4、Fable 5は一般ユーザに再公開（同年7月1日）されることとなった（Mythos 5は限定提供）。
- 本事例は、AIアライメントの問題だけでなく、サイバーセキュリティや国家安全保障、輸出管理上のリスクも含む。一方、強力なAIをどのような制約・監督のもとで社会実装するかという点で、AIアライメントと密接に関連する。

*3 UK Parliament, House of Commons Library, AI regulation in the UK, 2026/6/10, https://commonslibrary.parliament.uk/research-briefings/cbp-10003/ (2026/7/1アクセス)
*4 Anthropic, Redeploying Fable 5, 2026/6/30, https://www.anthropic.com/news/redeploying-fable-5 (2026/7/1アクセス)

オープンウェイトモデルをめぐる議論

- オープンウェイトモデルとは、推論を実行するために必要なAIモデルの重み（パラメータ）が一般に公開されたモデルのことである*1。
- 2026年7月に公開されたKimi K3モデルは、AnthropicやOpenAIのフロンティアモデルに肉薄し、一部の領域では上回る性能を示す。高性能なモデルが広く利用可能になる一方、AnthropicやOpenAIのモデルと比べて悪用防止のための安全対策が十分ではないとの指摘もある*2。オープンウェイトモデルは誰でも入手できるため、悪意のある利用者によるサイバー攻撃の高度化などが懸念される。
- また、アライメントの観点では、モデルの公開が安全性向上に向けた研究を促進する一方で、実装済みのアライメントを意図的に迂回・除去する技術の開発を促進する可能性もあり、利点とリスクの双方がある。このように、オープンウェイトモデルをめぐっては、技術の民主化によるイノベーション促進と安全性確保のバランスをどのように取るべきかについて議論が続いており、今後の動向が注目される。

AIアライメント観点での利点とリスク

利点

透明性／検証可能性の向上
- モデルの重みにアクセスできるため、機械論的解釈可能性の研究が促進されるなど、モデルの検証・評価が可能となる。

研究・競争の分散
- 大規模なリソースを持たない大学・企業の研究を後押しする。特定企業によるロックインを回避し、イノベーション・競争を促進する。

リスク

誤用・悪用
- 高性能なモデルを用いることで、システム脆弱性の探索、詐欺・偽情報の生成などを低コスト化しうる。

安全対策の除去
- モデルに施される安全対策を除去するなど、アライメントを崩すための研究が進展する。

オープンウェイトモデルの規制に対する議論

規制論の高まり

Kimi K3のような高性能なオープンウェイトモデルに対して、サイバー攻撃や危険能力の拡散につながるのではないかとの懸念から規制論が強まる。

AI業界の反応

7月24日にMicrosoft、Meta、NVIDIAなど25の企業・団体が性急な規制に反対する共同書簡を公表した*3。オープンウェイトは競争促進、コスト低減、技術主権の確保などに資することが理由。一方で、Anthropicは共同書簡には署名せず、書簡の内容に同意しつつも、必須の安全性試験の必要性を主張している*4。

米政権の反応

8月4日にトランプ政権が策定中のAI安全性評価制度について、オープンウェイトモデルを安全性テストの対象外とする方針を非公式に説明したと報じられる*5。安全性の管理と技術革新の促進のバランスを確保することが目的と考えられる。

*1 学習データ、学習に必要なソースコード、評価データ、開発過程などを公開する「オープンソース」とは異なる。
*2 NIST, UK AISI / CAISI Preliminary Assessment of Kimi K3's Cyber Capabilities, 2026/7/23, https://www.nist.gov/news-events/news/2026/07/uk-aisi-caisi-preliminary-assessment-kimi-k3s-cyber-capabilities (2026/8/5アクセス)
*3 NVIDIA, Open Weights and American AI Leadership, 2026/7/24, https://images.nvidia.com/pdf/Open-Weights-and-American-AI-Leadership.pdf (2026/8/6アクセス)
*4 Anthropic, Our position on open-weights models, 2026/7/27, https://www.anthropic.com/news/position-open-weights-models (2026/8/6アクセス)
*5 Bloomberg, China's Open-Weight Models Will Be Spared US Safety Tests, 2026/8/5, https://www.bloomberg.com/news/articles/2026-08-05/china-s-open-weight-models-to-be-spared-us-tests-us-firms-told (2026/8/6アクセス)

３．AIアライメントの実現に向けた考察

技術的な課題と解決の方向性

● AIの能力向上に対して人間による監督・理解・評価が追いつかず、長期タスク実行や自律的な行動において、危険な意図や逸脱行動を検出・制御できないことが課題となる。この解決には解釈可能性の向上、実運用を想定した高度な評価・監査、リリース後の継続的な監視と再学習を組み合わせた多層的な安全対策が必要となる。

項目	概要	解決の方向性
スケーラブルな監督の限界	• 高度な数学・科学、サイバーセキュリティ、長期実行タスクなど、人間が出力の妥当性を即座に検証できない領域が増加していく。能力向上に対してリスク評価や安全性を裏付けるエビデンス形成が追いつきにくい状況が指摘される。	• AIを補助として活用する監督、専門家レビュー、ルールベースの検証など、複数の評価手法を組み合わせる。
解釈可能性の不足	• モデルが特定の判断をした理由を十分に説明できず、内部に危険な目標、欺瞞的傾向、バイアスが形成されていても検出が困難となる。	• Mechanistic Interpretabilityにより、内部状態を詳しく検査する手法の研究が進められている。
自律エージェントの制御困難	• ツールの使用、長期タスク実行、コード実行、外部システム操作を行う自律エージェントが引き起こすリスクへの懸念が高まる。	• 権限分離、サンドボックス、承認ゲート、ロールバック機構などの対策を施す。高リスク操作は人間承認を組み込むことが望ましく、モデルに最小権限原則を適用する。
欺瞞・隠れた意図の検出が困難	• 高度なモデルが、評価時だけ安全に振る舞い、実運用では異なる行動を取るリスクへの懸念が高まる。	• 評価環境と本番環境の差を小さくし、AIレッドチーミング、内部状態監査、長期行動監視を行う。 • 単なる出力検査ではなく、出力に至る過程も検査する。
安全性評価ベンチマークの不足	• 現在のベンチマークは、実世界を想定したAIの使用や、複雑な悪用、長期的な変化などを十分に測定できていない。	• 静的ベンチマークに加えて、動的評価、実運用シナリオ、第三者監査を整備する。
継続的アライメントの難しさ	• モデル更新、外部ツールの使用、ユーザ行動の変化、社会規範の変化により、一度アライメントしたモデルも時間とともにミスアライメント状態に変化してしまう。	• リリース後監視、インシデント報告、再学習などのライフサイクル管理を導入する。

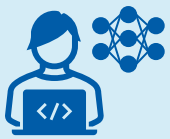
制度・社会的な課題と解決の方向性

● AIの急速な進化に対して法規制やガバナンス、国際協調が追いつかず、企業の競争圧力や価値観の多様性なども相まって安全性確保が難しくなることが課題である。解決には、リスクベース規制、国際的な共通ルール、第三者監査や企業ガバナンス強化、多様な価値観を反映できる柔軟な設計などを組み合わせることが重要である。

項目	概要	解決の方向性
規制・ガバナンス体制の整備の遅れ	• AI能力の進展が速く、政府の制度整備と企業のガバナンス整備の双方が遅れる可能性がある。	• リスクベース規制の導入（高リスクAIへの規制強化） • AI倫理委員会の設置、AI責任者の任命などの企業内AIガバナンスの強化
国際協調の難しさ	• AIがもたらすリスクは国境を越える一方で、各国の価値観、産業政策、安全保障上の利害は一致しない。	• 最低限の国際共通ルール策定（一定の禁止用途に関する最低限の共通ルールなど）や、インシデント情報の共有。 • 国際監視・調整機関の設立、安全性評価基準の標準化
企業インセンティブと安全性の衝突	• フロンティアAIモデルの開発競争が激しい中、企業は市場投入、性能向上、コスト削減を優先し、安全性評価を後回しにするインセンティブが生じる可能性がある。	• 安全投資を競争上の不利にしないための取り組みとして、業界標準、第三者監査の要件、調達要件を整える。安全性への取り組みを投資家、顧客、規制当局が評価できるようにする。
透明性と機密性のバランス	• 適切な情報公開は、外部からの検証可能性や信頼の向上に寄与し得るが、詳細を公開すると悪用を助長する可能性がある。	• 社会的説明責任と悪用防止を両立させる。公開情報、監査機関向け情報、規制当局向け機密情報を分ける。
人間の自律性・依存の問題	• AIが助言、意思決定支援に深く入ると、人間が判断力を失ったり、AIの提案に過度に従ったりする可能性がある。	• AIを人間の支援者として設計する。根拠提示、反対意見提示、ユーザ選択の確保、教育的な利用を設計として組み込む。
価値の多元性・文化差の反映の難しさ	• 人間の価値観は単一ではなく、国、組織、宗教、世代、個人で異なる。AIアライメントは人間の意図や価値への整合を目指すが、何に整合すべきか自体が難しい。	• 普遍的な安全原則と地域・組織・利用者ごとの調整可能なポリシーを分けて設計することが望ましい。

AIアライメントを実現・維持するために求められる役割

- AIアライメントに関する議論は、OpenAIやAnthropicに代表されるAI開発企業の取り組みを中心に発展してきた。そのため、多くの日本企業にとっては「AIを開発していないので、自社には直接関係のない話だ」と受け取られるかもしれない。
- しかし、AIアライメントの実現はモデル開発時の設計・調整だけで完結するものではない。AIが実際に利用される環境では、AI提供者による適切な運用管理や、AI利用者による監督・フィードバックが求められる。AI開発者がアライメントを実装し、AI提供者とAI利用者がそれを維持・改善することで、AIは人間の意図や価値観と整合した形で活用され続ける。



AI開発者

AIシステムを開発する。

①人間の意図・価値観を反映したモデル学習
- 人間のフィードバックによる学習（RLHF）
- 憲法AI

②学習データの品質管理
- バイアス分析・除去
- 有害データ除去

③安全性評価・AIレッドチーミング
- 悪意あるプロンプトを用いた評価
- ジェイルブレイクテストなど

④透明性向上
- モデルカード作成
- 第三者評価の活用

AI提供者

AIシステムをアプリケーション、製品等に組み込んだサービスとして提供する。

①利用目的に応じたガードレール設計
- プロンプトの制御
- 入出力のフィルタリング

②利用状況の継続的モニタリング
- 入出力データの変化監視
- インシデント管理

③サービスとしてのリスク管理
- 高リスク時の人間介入機能の提供
- 承認ワークフローの提供

④利用者支援
- 利用ガイドラインの整備
- 教育プログラムの提供

AI利用者

AIシステムまたはAIサービスを利用する。

①利用目的の明確化
- AI活用方針・利用範囲の決定
- KPI設定

②人間による確認・監督
- 結果のレビュー、複数者による確認
- 承認プロセス組み込み

③フィードバック提供
- 誤回答の報告
- 改善要望の提出

④ガバナンス体制整備
- AI利用ポリシーの策定
- AIリスク評価

AIアライメントを踏まえたAIとの向き合い方

● AIアライメントの研究や技術は発展を続けているものの、人間の意図や価値観と完全に一致したAIはまだ実現していない。そのため、AIを活用する組織や個人には、AIを単なる効率化ツールとして利用するだけでなく、その出力や判断を適切に監督しながら活用する姿勢が求められる。

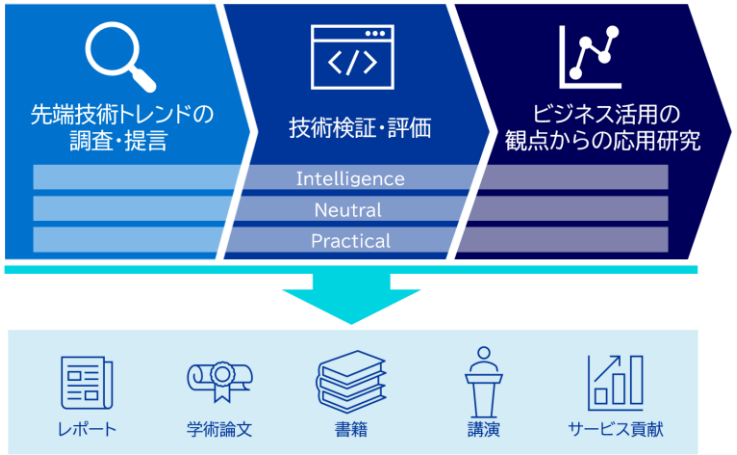
	従来重視されてきた観点	AIアライメントを踏まえて加えるべき観点
問いを立てる能力	良い問いを立てる力が重要になる。入力された問いに応じて出力の質が大きく変わるため、課題設定力や論点整理力が価値を持つ。	問いを立てるだけでなく、自分や組織が本当は何を望み、何を良しとするかを明確にする必要がある。
価値観・判断基準	AIの回答を、業務上有用かどうかという観点で評価する。正確性、網羅性、効率性が重視される。	AIの出力が、単に正しい回答ではなく、組織の価値観、倫理基準、顧客本位、社会的責任と整合しているかまで評価する必要がある。
人間とAIの関係	人間がAIを監督し、AIは補助者として使う。最終的に意思決定をするのはAIではなく、人間であるべき。	AIが高度化すると、人間がAIを監督しているつもりでも、実際にはAIの提案に依存し、判断を過度に委ねる構造が生まれうる。実質的な判断の主導権を人間側が握り続ける必要がある。
リスク	主なリスクは誤答、ハルシネーション、著作権、情報漏洩である。人間によるファクトチェックや根拠の確認が必要。	AIによって人間の価値観、選好、判断基準が無意識のうちに誘導・書き換えられてしまうリスクを警戒する必要がある。
責任	AIが出した結果であっても、最終責任は人間が負うべき。AIの出力を利用したことで、利用者・組織の責任が免除されるわけではない。	人間が承認するだけではなく、人間が実質的に責任を負う仕組み（責任を理解、AIの判断に介入、AIの提案を拒否できる）が必要である。
組織ガバナンス	利用ルールや禁止事項を定め、情報漏洩や誤利用を防ぐ。	AI活用は現場任せではなく、経営・リスク管理・業務部門が一体となって取り組むガバナンス課題である。
総論	• 人間はAIを有用な道具として活用することが重要である。適切なプロンプトを与え、出力を検証し、最終判断と責任を担うことが必要。 • 焦点はAIの性能を最大限活用して生産性や意思決定の質を高めることにあり、人間はAIの利用者・管理者として主体的にコントロールする立場に置かれる。	• AIをどう使うかではなく、AIと人間が何を目指すべきかを定義することが重要である。最大の関心は出力の正確性よりも、どの価値や目的が最適化されているかにある。 • 人間には、AIへの指示者にとどまらず、意思決定の基準や社会的価値を明確化し続ける役割がある。

【参考】AIアライメントを支える組織・個人の実践例

	組織レベル	個人レベル
価値観の明確化	• 自社の価値観を反映したAI利用原則を明文化する。 • AIを使わない領域を決める（AIに何をさせるかではなく、何をさせないかを考える）。 • AI利用のガバナンス体制を作る。	• 個人の価値観を言語化（明確化）する。 • 上記が難しい場合は、利用に際して何を達成したいか、優先事項は何か、制約条件は何かなどを言語化する訓練をする。
AIを監督する **人間がAIを実質的に統制する**	• AIリテラシー向上のための教育を行う。AIの活用教育だけではなく、判断するための教育（例：AIを疑い、修正する）も実施する。 • AI利用のログを残し、監視・監査ができる仕組みを整備する。 • AI活用KPIだけでなく、AIへの依存を測るKPIを設定し監視する（例：AI提案の採用率ではなく、却下率も監視）。 • 理由を説明できない判断を禁止する。（例：AI提案採用時は採用理由を記録する） • 責任の所在を明文化する。	• AIに結論を聞くのではなく「反論」や「選択肢」を聞く。 • AIなしで考える時間を残す。AIに頼らず考え、判断する機会を意図的に設ける。 • AIの回答を鵜呑みにしない。判断に迷う場合は上司・同僚・信頼できる人に相談する。 • 専門性を捨てずに、磨き続ける。
フィードバックと継続的改善	• 自社の利用目的に沿っているかどうかを検証するためのモデル評価を自社で行う体制を整備する。 • 導入するAIが自社の価値観と一致しない場合など、問題を報告したり、異議を申し立てたりできる窓口を整備する。	• AIシステムの改善フィードバックに協力する（不適切な出力や不具合の報告など）。 • 自分の使い方を振り返る。AIに任せすぎていないか、自分の判断力が低下していないかを振り返り、より良い利用方法を検討する。

先端技術ラボのご紹介

先端技術を活用したITサービスの創出に向けた技術の目利き役として、
「先端技術トレンドの調査・提言」、「技術検証・評価」、「ビジネス活用の観点からの応用研究」に取り組んでいます。



主な研究・取り組み領域

当社ホームページの **特集サイト** では、IT分野における先端技術の調査レポート、およびメンバーのプロフィール詳細がご覧いただけます。
本レポート執筆者への取材や講演などに関するご相談は、当社ホームページの **お問い合わせフォーム** よりご連絡ください。

株式会社日本総合研究所

日本総研は、シンクタンク・コンサルティング・ITソリューションの3つの機能を有するSMBCグループの総合情報サービス企業です。

東京本社　〒141-0022 東京都品川区東五反田2丁目18番1号 大崎フォレストビルディング
大阪本社　〒550-0001 大阪市西区土佐堀2丁目2番4号

日本総研

The Japan Research Institute, Limited