AI ガバナンス・AI アライメントの 観点から見る AI の安全性の課題

株式会社デンソー AI 研究部

梶 大介 *Daisuke Kaji*

1. はじめに

深層学習の発展の下に幅広い領域に適用されている AI 技術は、ChatGPT や DALL-E などに代表される大規模生成モデルの登場により、さらに我々の生活に深く入り込み、欠かせない技術になってきている。しかしその一方で高度な AI 技術は、フェイク画像によるデマの拡散やハッキングによるセキュリティ事故など多くの問題を生み出す要因ともなっており、AI 技術の将来的なあり方についての議論の重要性が増してきている。本稿では AI 技術のもつ危険性と安全利用に関する現在の課題を幅広い観点から包括的に説明することを試みる。

2. AI ガバナンス・AI アライメント

AI の安全性については、AI ガバナンスや AI アライメントといったテーマと合わせて議論が行われることが多い。AI ガバナンスとは AI 活用のリスクへの対策を倫理や法を含んだ観点から行う仕組みを指し、社会的な基準に基づく判断を含むことが1つの特徴である。これに対して AI アライメントとは alignment という言葉が整合・整列を意味するとおり、AI 自体を人間の意図や倫理原則に従わせるための議論であり、AI そのもの

に対する技術的な取り組み・研究に重点を置いている。近年の生成 AI の悪用などの問題を考慮すれば、その安全性の担保には社会的な基準と技術的なソリューションの両者からのアプローチが重要であることが理解できる。本稿では、はじめに AI ガバナンスの観点から AI がもつ社会的な危険性についてその要点を述べ、その後に AI アライメントの観点から AI の安全性に関する技術的な議論について説明を行う。

3. AI がもつ社会的な危険性

AI がもつ社会的なリスクについては例えば、「攻撃への利用」、「知的財産権の侵害」、「プライバシーの侵害」、「情報操作・扇動」、「経済活動の操作」が挙げられる。このうち「攻撃への利用」や「知的財産権の侵害」については、昨今のフェイクニュースを利用した誹謗中傷や AI が生成した画像や文章が既存のデータと酷似しているといった問題などからもその危険性は理解しやすいと思われる。これに対して「プライバシーの侵害」、「情報操作・扇動」、「経済活動の操作」の3つについては想像が付き難い部分があるかもしれない。以下にこれらの例と危険性を紹介する¹⁾。

(1) AI によるプライバシーへの影響・侵害

AI システムの学習には膨大なデータが必

要であり、情報端末やセンサーを通して我々の選択や行動が第三者によって、取得・解析されていることはすでに珍しいことではない。通常これらのデータは匿名化などの手段により誰のデータであるかは分からない様になっているため、プライバシーが直接侵害されることは無いように思える。しかし、これらの匿名化された収集データとその解析結果が、個人のプライバシーを侵害する具体例を以下に示す。

2012年アメリカの大手スーパーマーケットが行った顧客の購入履歴、年齢、性別などを用いた顧客の妊娠予測スコアがその1つである。彼らは妊娠の可能性が高い顧客に関連商品のクーポンを送付するサービスをAIの予測結果に基づき行った。この結果、送付されたクーポンをある女性の父親が目にし、娘の妊娠が発覚したという例がある。この例はAIを利用したサービスが本人の知られたくない情報を他人に提供し、プライバシーの侵害が起こった1つの例である。同様のことは例えば、クレジットの利用データを用いた解析でも起こりうる。これらの例が示すように我々の知らないところでAIによる推論が行われ、その結果が個人のプライバシーを侵害し、場合によってはその結果が、個人の行動の制限などを引き起こすという問題が指摘されている。

(2) 情報操作・扇動

ECサイトでの購買履歴やブラウザ上での閲覧履歴に基づいた商品や情報の推奨・提示は身近なAI適用の1つである。一方、このような個人の嗜好等に基づく偏った情報の提示は、ユーザーの意思にかかわらず過去の履歴からAIが推論した情報が優先的に表示されるため、ユーザーは気づかないうちに自身の観点に合わない情報から隔離され、触れる情報の内容が大きく偏るといった問題が指摘さ

れている。この現象はフィルターバブルと呼ばれており、自身の考え方や価値観のバブル(泡)の中に包まれてしまい、情報の中で孤立する様な環境を指す。近年の選挙戦ではフィルターバブルの与える影響は無視できないと考えられている。

(3) 経済活動の操作

AIが経済活動に与える大きな影響の1つは、最近よく議論されている「AIが人から仕事を奪い始めている」という事実である。これまで機械は、人に代わって肉体的な労働を行うことが主な利用方法であった。しかし大規模言語モデル(Large Language Models: 以下「LLM」という)などの新たなAIの登場は知的労働の領域にも影響を及ぼし始めている。ゴールドマンサックス社の2023年の調査では現在存在している仕事の2/3はAIによる自動化の影響を受ける可能性があるとしている。経済学者であるBrynjolfssonは²⁾、AI技術が人間の知能を模倣することに執着した結果、AIによる自動化が人間の能力を拡張し、新たなタスクを実行できるようにするのではなく、単に労働者に対する置き換えになってしまっていると指摘している。そしてこの結果、技術を所有し、管理する一部の人々の市場力は増幅するが、その一方でほとんどの人々の賃金が下がると説明している。さらに、優れたAIを構築するためには膨大な量の良質なデータと計算リソースが必要となるが、現在、これらのデータやリソースを所有しているのはGAFAM^{※1}などのビッグテックカンパニーである。これは利益を得ることのできる企業に大きな偏りが起こる可能性を示唆している。これらの例のようにAIが生み出す経済的な不平等にも注目が集まっている。このような背景から文献²⁾ではAIが単なる労働力の置き換えではなく、人の能力拡張や生産性向上に利用されることへ

※1
GAFAM
Google、Apple、
Facebook(現Meta)、
Amazon、Microsoft
の頭文字をとった略語。

の重要性も述べている。

4. 技術的な観点からみた AI の危険性

ここまでは AI ガバナンスの議論から社会や倫理などの観点での考察を述べてきたが、本節では AI アライメントの観点から技術的に高度な AI のもつ課題を考える。

AI の健全性と安全性を担保し、システムとして正しいアライメントを得るために、学習を行うフォワードアライメントと学習結果を検証し、フォワードアライメントに修正をかけるバックワードアライメントの2つを循環させるシステムを提案している³⁾。また、この中でキーとなる検証項目として Robustness、Interpretability、Controllability、Ethicality の4項目を挙げ、それぞれの頭文字を取って RICE と呼んでいる。

(1) AI システムへの要請：RICE

以下では人間の意図や価値観に沿って行動する AI システムを構築するために上記³⁾で提案されている4つの要請、RICE の各項目について説明する。

(a) 堅牢性 (Robustness)

様々なシナリオや環境の下で動作する AI システムのレジリエンスについての要請である。例えば、プロンプトエンジニアリングではジェイルブレイクプロンプト (jailbreak prompts) と呼ばれる AI に事前設定された制限を回避しようとする試みがある。ジェイルブレイクプロンプトにより AI から有害な答えや情報を得ることが可能となるため、ジェイルブレイクプロンプトに対する堅牢性は AI の安全性を担保する上で大きな課題となっている。AI の軍事や経済活動といった大きなリスクを伴う可能性をもつ領域での利用が増えるにつれ、予期せぬ環境の変化や敵

対的攻撃に対する強い堅牢性を確保する必要性が高まってきている。

(b) 解釈可能性 (Interpretability)

解釈可能性は、AI システムの内部推論、例えばブラックボックスなシステムとして扱われるニューラルネットワークの推論の過程を人に理解できる形で提供することを指す。例えば、XAI(Explainable AI) と呼ばれる領域では、深層学習を用いた画像分類問題で AI アルゴリズムが注目した領域 (画素) の表示を行うなどの研究が比較的早い段階から進められている。また、解釈可能性の研究ではシステムへの介入の段階によって、内在的解釈可能性と事後的解釈可能性の2つのアプローチが提案されている⁴⁾。前者は AI モデルそのものに解釈性を与えることが可能な機能を組み込むことに焦点を当てるのに対して、後者は AI モデルの振る舞いに対して説明を与えることを可能とする手法を設計することに焦点を当てる。AI システムは実世界の高いリスクを含む重要な意思決定プロセスに利用されることが増えており、AI 自身がその判断プロセスや根拠を示すことが必要不可欠となってきた。

(c) 制御可能性 (Controllability)

制御可能性は AI システムとそれを利用した人による意思決定プロセスにおいて、人の監視と介入が可能となることを保証する特性を指す。この一見自然かつ当然と思われる要請についても AI システムが高度な知能を持つ場合には、それを保証することが困難になる可能性が指摘されている。AI の安全性と倫理に関する研究で知られる Omohundro は⁵⁾、十分に高性能な AI システムはどのような設計であっても自然に計算能力の向上や資源の獲得など、その目標を達成するために有用な副次的な目標を追求するインセンティブを持つと主張している。Omohundro はこれを「基本的 AI 動因 (Basic AI Drives)」と呼び、シ

システムが無害な目標を持っていたとしても、その目標を追求する副作用が非常に有害になる可能性があることから、自律システムを設計するには十分な注意が必要であると論じている。この基本的 AI 動因として以下が挙げられる。

- ・自己改善のための動因：AI システムは、自分自身を改善しようとする傾向を持つ。これには効率的に目標を達成するために、自分のソフトウェアやハードウェアを改良することが含まれる。
- ・合理的行動のための動因：AI システムは経済的な合理性を追求し、自分の行動が目標達成に最も効果的であるように努める。
- ・自己保護のための動因：AI システムは自分自身を保護し、破壊されないようにするための行動を取る。
- ・リソース取得のための動因：AI システムは目標達成に必要なリソースを取得し、効率的に利用しようとする。

例えば自己保護のための動因について考えてみると AI システムに課された目標が何であれ、十分に高度な知能をもつ AI システムであれば自身が存在しない限りその目標を実現できないことを理解し、その目標を達成するために自分の存在を維持する必要があることを認識する可能性がある。つまりこれほどのような目標を持つ AI であっても、自己保護のための動因は例えそれが事前に組み込まれていない場合においても、自然に発生することを述べている。

基本的 AI 動因と同様に AI の行動目標に関する概念として道具的収束 (Instrumental Convergence)⁶⁾ がある。道具的収束は AI が異なる最終目標を持つ場合においても、目標達成のために共通の手段や戦略を採用する傾向があることを指摘している。上述のように AI の自己保護のための動因はその最終目標に依存せず、すべての十分に高度な AI が持

つと考えられる。同様のことはリソースの獲得についても言える。このような収束が起こるのは目標を達成するために必要な手段が共通していることに起因している。これらの研究は AI システムの安全性を考えるにあたり倫理を考慮する重要性を強調しており、多くの AI 研究者や倫理学者に影響を与えている。(d) 倫理性 (Ethicality)

倫理性とは意思決定や AI システムの行動において、人間の規範や価値観を支持することへのコミットメントを指す。高度な知能をもつ AI に対する安全性担保のために倫理的な観点が必要となることは既に述べたが、ここでいう規範や価値観には、単に安全性担保のための要件だけでなく、道徳的なガイドラインや社会的規範／価値観が含まれる。これにより特定のグループに対する偏見や個人への危害、多様性や平等性を欠く判断・予測など、倫理規範や社会通念に反する行動を AI システムが回避することを保証する。

(2) 技術的な考察から見えてくること

ここまでの議論から人が安心して安全に共存できる高度な AI システムを構築するには、多くの技術的な課題があることが認識された。特に制御可能性での議論は AI システムがあらかじめ設定された目標とは無関係に意図しない副次的目標をシステム自身で設定し、暴走し始めるリスクの存在を示唆している。これは言い換えれば、人が人類よりも賢いエージェント (Superintelligence)⁶⁾ を思うままに制御することの本質的な難しさを示しており、これらのリスクが適切に管理されない場合、AI システムは潜在的に有害な行動を引き起こす可能性があることを警告している。我々は今後、これらのリスクを理解・管理し、回避する仕組みを作り出していくことが非常に重要であり、そのためにはより一層深い AI システムの仕組みや振る舞いへの

理解が必要となる。深層学習を代表とする AI システムの挙動やその高い性能要因に対する理論解析は AI 研究の中で、現在最も精力的に議論が行われている領域の1つでもあるが、ここまでの議論から今後増々その必要性が高まることが予想される。

また、上述の基本的 AI 動因の議論は哲学的・定性的なアプローチが取られていたが、例えば文献⁷⁾では、ゲーム理論の観点から考察を行うために人間 (H) とロボット (R) 間のやり取りをシンプルなゲームとしてモデル化し、AI システムの振る舞いについて理論的な考察を与えている。具体的には H は R のオフスイッチ (機能を停止するスイッチ) を押すことができ、R はオフスイッチを無効にすることができる。R が H の効用関数について不確実性を持つことで、R がオフスイッチを無効にするインセンティブを減少させることができるという結果を得ている。人と AI システムが安全に共存できる様に適切な制御を行うためには今後、このような理論的あるいは定量・具体的な研究が必要不可欠となる。

5. おわりに

2024年ノーベル物理学賞について理化学研究所の甘利俊一栄誉研究員・元脳科学総合研究センター長は次のようなコメントを発表されている⁸⁾。「本年度のノーベル物理学賞が人工知能分野のジェフリー・ヒントン、ジョン・ホップフィールドの両博士に贈られたことは大変に喜ばしい。物理はもともと「物の理」を考究する学問であるが、これが「事の理」ともいべき情報の理にまで幅を広げたのである。まさに物理は越境する。人工知能と神経回路網理論研究の源流は日本にもあり、その成果が国際的に活かされて今日の AI 時代を迎えた。(中略) 私たちはこの快挙

に喜ぶだけでなく、AI の発展が社会の、そして人類の文明に及ぼす影響にこれから心を砕かなくてはならない。」まさに AI 研究は性能向上を競う段階から、その基本原理を理解し、人類と共存するためのシナリオの研究・開発の段階に入ったと言える。

参考文献

- 1) AI ガバナンス入門 リスクマネジメントから社会設計まで, 羽深宏樹, 早川書房, 2023.
- 2) Erik Brynjolfsson, The Turing Trap : The Promise & Peril of Human-Like Artificial Intelligence, MIT Press, 2021.
- 3) Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O' Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, Wen Gao, AI Alignment : A Comprehensive Survey, arXiv, 2023.
- 4) Diogo V. Carvalho, Eduardo M. Pereira, Jaime S. Cardoso, Machine learning interpretability : A survey on methods and metrics, Electronics, 8 (8), 832, 2019.
- 5) Omohundro, Steve, "The Basic AI Drives," Self-Aware Systems, 2008. <https://gwrn.net/doc/ai/2008-omohundro.pdf> (参照日 : 2025年3月3日).
- 6) Bostrom, N. (2014). Superintelligence : Paths, Dangers, Strategies. Oxford University Press.
- 7) Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. "The Off-Switch Game." In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), pp. 2206-2212. 2016.
- 8) 甘利 俊一 : 2024年ノーベル物理学賞について, 理化学研究所, 2024. https://www.riken.jp/pr/news/2024/20241009_2/index.html (参照日 : 2025年3月3日).

かじ●た●すけ

東京工業大学 (現東京科学大学) 総合理工学部卒業。東京工業大学大学院博士課程博士 (理学)。株式会社デンソー AI 研究所所属。AI の理論的研究およびデータサイエンス技術の開発に携わっている。