IFI Working Paper

IFI Working Paper No.19 March 2023

AI ガバナンスの課題と展望

江間 有沙
東京大学未来ビジョン研究センター 准教授

東京大学未来ビジョン研究センター
Institute for Future Initiatives, The University of Tokyo

AI ガバナンスの課題と展望

東京大学未来ビジョン研究センター
江間　有沙

1. AI ガバナンスの課題と価値をめぐる議論

人工知能（AI）は私たちの生活や働き方に多大な影響をもたらす。現在、国際機関、各国、企業、業界団体や市民団体といった様々な組織が、AI によるイノベーションを促進しながらリスクを最小化するガバナンスの在り方を模索している。

本稿では、AI ガバナンスをめぐる話題として 3 つの課題を検討する。これらの課題は多くは AI 技術を活用したシステムやサービスの社会実装や政策に関係している。課題認識は大まかに共有されているものの、その具体的な解決策は、まだ必ずしも明示的ではないものも多い。特に AI ガバナンスに関しては価値をめぐる議論となるため、多様なステークホルダーを巻き込んだ議論を継続していく必要がある[1]。

最初の課題である AI 原則の実践への落とし込みは、価値をめぐる議論に関連している。多様性と包摂性を重視することが重要である中、互いの価値を尊重しながらも、可能な限り共有できる方針を打ち立てていく、あるいは相互運用可能な仕組みを構築していくことが現在の AI ガバナンスをめぐる政策上の課題にもなっている[2]。

次の課題は人と機械の協同作業に関する議論である。人の意思決定や行動に AI が関与する領域は、ますます増えているが、そのような人と機械の関係性に関して画一的な回答はない。AI 技術の精度や人間の価値判断、応用分野の文脈に関連して関係者間で議論していく必要があるが、相互運用可能な尺度の構築の研究や政策はその一助となるはずである。

最後の課題としてあげるのは AI 技術をはじめとして情報技術一般が持つ文化と、医療や交通など物理空間が持つ安全文化とのすり合わせから生じる問題である。サイバー空間だけではなく様々な物理空間に AI 技術が浸透していくにつれ、事前の性能保証や事後の検証が困難である深層学習ではない、解釈可能な AI が求められる応用領域は増えている。サイバー空間と物理空間の融合で得られる効果だけではなくその課題、さらにはどのような技術的な展開が求められるのかに関して、事例をもとに考察する。

2. AI 原則を実践へ落とし込む際の価値と相互運用性確保の課題

2016 年、G7 香川・高松情報通信大臣会合の議長国だった日本は、国際的な議論のための

---

[1] 江間有沙、「AI と社会」技術評論社、2022

[2] 2022 年 12 月の米 EU 貿易技術評議会（TTC）でも AI に関する共通用語や分類法、リスク管理ツールに関して国際協調していく合意に達している。FACT SHEET: U.S.-EU Trade and Technology Council Advances Concrete Action on Transatlantic Cooperation, https://www.whitehouse.gov/briefing-room/statements-releases/2022/12/05/fact-sheet-u-s-eu-trade-and-technology-council-advances-concrete-action-on-transatlantic-cooperation/

AI 開発ガイドライン案[3]を提案した。その後も G7 サミットでは AI ガバナンスの必要性に関する議論は継続され、OECD の AI 原則[4]や G20 の AI 原則[5]をはじめとして AI 開発や利活用に関する原則が産学官民の多様な組織で策定されている。

2018 年から 2019 年にかけては多くの原則が出揃い、それらを比較した研究も行われている。AI 原則を類型化した研究では、透明性、公平性やプライバシーなどいくつかの価値に原則が集約していることをする一方、アフリカや中央アジアなどでの倫理的な議論というのは、欧米と比べて進んでいないことなども指摘しており、地理的に不均一であるとの課題も指摘されている[6]。このような点から欧米の規範やイデオロギーのみで AI 原則が議論されていることに関して疑問を投げかけるグローバルサウスからの意見もある[7]。

さらに現在、原則を実践に落とし込にあたって欧州 AI 法案のように法的拘束力のある規制もあるが、多くは法的に非拘束（non-binding）な方法論やツールであり、AI 技術のリスク管理フレームワーク[8]やリスク評価[9]、検証[10]、監査[11]、品質管理[12]、事業者ガバナンスの在り方[13]などが現在開発されている。

現在、グローバル企業は AI の倫理指針や原則を作り始めている。日本ではソニー（2018

---

[3] 総務省 AI ネットワーク社会推進会議（2017）、国際的な議論のための AI 開発ガイドライン案、https://www.soumu.go.jp/main_content/000490299.pdf

[4] OECD AI Principles overview, https://oecd.ai/en/ai-principles

[5] G20 AI 原則、https://www.mofa.go.jp/mofaj/gaiko/g20/osaka19/pdf/documents/jp/annex_08.pdf

[6] Anna Jobin, Marcello Ienca & Effy Vayena：The global landscape of AI ethics guidelines, Nature Machine Intelligence, 1, 389-99, 2019

[7] 2023 年 3 月 3 日に東京大学未来ビジョン研究センターで開催されたウェビナー「アフリカでの AI ガバナンスに関する企業や政府での取り組みから日本が学べること」（https://ifi.u-tokyo.ac.jp/event/14978/）においては、アフリカからの登壇者が、欧米の規範では個人の自立が重視されているが、アフリカ文化では婚姻やビジネス、医療行為に関する判断を個人ではなくコミュニティ（例えば族長が判断するなど）で決定することがあると指摘した。グローバルスタンダードを議論する時に文化的な、価値的な多様性をどのように考えていくかが問題視されている。

[8] シンガポール政府の AI ガバナンスフレームワーク（https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework）、OECD の Framework for the Classification of AI Systems（https://oecd.ai/en/classification）などが公開されている。

[9] カナダ政府が Algorithmic Impact Assessment tool（https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html）を公開しているほか、欧州評議会も「人権、民主主義、法の支配影響評価（HUDERIA）」という AI のリスク影響評価の共通の方法論を模索している。また、UNESCO も AI 勧告を推進するための倫理的な影響評価(Ethical Impact Assessment)の必要性を議論している。

[10] シンガポール政府は 2022 年に「AI Verify」という AI ガバナンスのテスト用ツールキットを公開し、これによって組織が適切に AI を活用していることを客観的かつ検証可能な方法で示すことができるとしている(https://www.imda.gov.sg/How-We-Can-Help/AI-Verify)。

[11] アメリカのニューヨーク市では自動雇用判断ツール（AEDT）の AI システム監査の議論が行われている（https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7CText%7C&Search=）。

[12] 日本の産業総合研究所が、AI を用いた製品やサービスの品質を管理する「機械学習品質マネジメントガイドライン」を公開している。（https://www.digiarc.aist.go.jp/publication/aiqm/）

[13] 日本の経済産業省は「AI 原則実践のためのガバナンス・ガイドライン」を公開している（https://www.meti.go.jp/press/2021/01/20220125001/20220124003.html）

年）が先鞭をつけ、富士通（2019 年)、NEC（2019 年)、NTT データ（2019 年）が立て続けに原則や指針を公開した。その後も、国内外の企業の多くが倫理原則の策定を行っている[14]。これらの背景には AI のもたらす課題に対応をしなければ社会に製品やサービスが受容されないという企業の関心がある。その結果、現在、AI の原則などを定めることが、企業にとって「経営戦略に直結する問題」であると認識されるようになりつつある[15]。一方で、原則を作った後にこれらをどのように実践として運用していくかが、現在課題となっているため、様々に展開されている方法論やツールは、事業者にとってと助けとなるが、方法論やツールが乱立している状況は、事業者としてもどの方法論に依拠して開発や運用を進めていけばいいのか迷う。同時に、応用先に存在する規制や組織マネジメントシステムと重複することにより、規則や管理が二重化、複雑化するとの懸念もある。

また、データ提供や AI モデル開発、AI サービス提供が複数の組織にまたがるなど、多くの場合 AI サービスやシステムは（グローバルを含む）複数の組織で成り立つことになる。特に日本は産業構造として B2B（Business-to-Business）の企業が多く、AI システムやモデル開発会社、サービス提供会社、ベンダー、そしてエンドユーザーからなる長いサプライチェーンを形成している。AI のライフサイクルはそのサプライチェーン全体で捉える必要があり、サプライチェーンが長くなるといくつかの問題が生じる。例えばユーザー企業とベンダー企業の関係で見たとき、ユーザー企業ごとに要求される AI ガバナンスの方法論やツールが相互運用不可能であると、ベンダー企業の負担が大きくなり、技術の積極的な利活用を阻害しかねない。あるいは、AI に関する原則や組織ガバナンスが整っていない企業は政府調達に参加できないと AI 原則の存在自体が参入障壁となる場合もある[16]。

さらに AI に関する組織ガバナンス（安全性の検討やデータバイアスを検証するような内部監査人や問題が起きたときの対応窓口等）を十全に確保することを求めると大企業に有利となる。一方で AI 技術に関してはスタートアップ企業の躍進が目覚ましいところがあるが、そのような企業では人もコストもリソースが限られてしまう[17]。商品やサービスに人工知能技術を用いていると公言すると審査や承認、検査に時間がかかるとなった場合、「本製品は人工知能技術を用いておらず、既存商品の性能を少しだけ改良したものである」、として検査をやり過ごそうとしたり、AI で重視する価値を考慮に入れずにサービスやシステムを低価格で提供しようと考える企業が現れないとも限らない。そのため、AI ライフサイク

---

[14] 総務省 AI ネットワーク社会推進会議「報告書 2022」では、67 の海外、国際機関が出している AI 関係のガイドラインを調査し、どのような価値が提起されているかの比較調査を行っている。また国内では 22 のガイドラインを比較している。https://www.soumu.go.jp/menu_news/s-news/01iicp01_02000110.html

[15] ABEJA, Ethical Approach to AI が発足。メディア関係者に委員が抱負を語りました、2019 年、8 月 21 日、https://abejainc.com/ja/news/article/20190821-2542

[16] AI が多様性で包摂性を認める社会というビジョンを共有する原則の下にあると考えると、この帰結は皮肉でもある。

[17] 日本ディープラーニング研究会（2021)、AI ガバナンスエコシステム-産業構造を考慮に入れた AI の信頼性確保に向けて、2021 年 7 月 21 日、https://www.jdla.org/about/studygroup/sg01/

ルを通じたガバナンスの方法論やツールの相互運用可能性が求められている[18]。

3. 人と技術の協同作業をめぐる関係性の課題

AI サービスやシステムと人との相互作用の在り方は AI ガバナンスの問題と不可分である。人の行動や判断を AI システムやサービスが支援する場合と、AI が何らかのアクションを自律的に為す場合では、そのリスクや社会的影響は異なる。人間の行動や判断支援に AI を用いる場合、AI の判断を人間が監督、上書きできることが重要である。AI の自律的な動きが優先される場合でも、人間が事後確認することやモニタリングすること、一時停止することも人間の関与の方法と考えられる[19]。一方で、適切な関与を設計しなければ、AI の精度やバイアスに関する問題も出てくる懸念もある。

AI サービスやシステムとその利用者の協同作業には様々なバリエーションがあり、応用先の文脈に応じて AI と人との協力体制の在り方を定義する必要がある。例えば運転支援において、SAE（Society of Automotive Engineers）が運転支援と自動運転にどのようなシステムを導入できるかレベル分けている。また、医療 AI のタイプ分けでは AI システムの精度やユーザーインタフェースなどを含む技術要件、AI システムが医療関係者に及ぼしうる影響、さらには患者や利用者に及ぼしうる影響なども整理したうえでタイプを分類している[20]。応用先の特性に応じた分類を定義し、それを共通の土台として国際的議論を推奨することが、技術のイノベーションを促進してリスクを低減することにつながる。エンドユーザーにとっても、AI と人の役割分担や協同作業の在り方が明確になることで、AI サービスやシステムを不必要に忌避せずに利用することが可能となる。

日本政府が掲げる Society 5.0 は、サイバー空間とフィジカル空間を高度に融合させたシステムにより、経済発展と社会課題の解決を両立する人間中心の社会である[21]。Society5.0 の社会を実現するにあたって AI に寄せられる期待も大きい。AI は現在、様々な応用分野に適応されつつあるが、どこまで何を AI に任せるかは応用分野によって異なる。例えば日本では医療において AI はあくまでも診療支援であって、最終判断をするのは医師と明示されている[22]。一方、囲碁や将棋などの AI がすでに人間のパフォーマンスを超えている領域においては、AI の打ち筋を分析してそれを活用するだけではなく人が新たな打ち筋を生み出

---

[18] OECD (2023), Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI", OECD Digital Economy Papers, No. 349, OECD Publishing, Paris, https://doi.org/10.1787/2448f04b-en.

[19] “WHITE PAPER on Artificial Intelligence - A European approach to excellence and trust”では、様々な人間の関与に関する議論が行われている（https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en）。

[20] 東京大学が提案する医療 AI のタイプ分け分類では、医療 AI をレベルではなくタイプ分けとして、適切なタイプを選ぶことが重要であるとしている（https://ifi.u-tokyo.ac.jp/news/5456/）。

[21] Society 5.0、https://www8.cao.go.jp/cstp/society5_0/

[22] 2018 年 12 月に発出された厚生労働省医政局医事課長通知では AI はあくまでも診療支援であると明示されている。

しているなど、人と AI が共進化している事例が報告されている[23]。

一方、「サイバー空間とフィジカル空間が高度に融合されたシステム」は、必ずしも世の中に平和をもたらすシステムではなく、時に混乱をももたらす。その分かりやすい事例が兵器などの物理的な手段を用いた攻撃だけではなく、情報戦や世論操作などの心理戦、サイバーセキュリティ攻撃など正規・非正規が組み合わされたハイブリッド戦争である。

ハイブリッド戦争が注目されたのは 2014 年によるロシアによるウクライナ危機や 2016 年の米大統領選におけるサイバー攻撃であるが、現在のウクライナをめぐる情勢においても話題となっている。特筆すべきはこれらの攻撃には、セキュリティを破ろうとするようないわゆる技術的な攻撃以外にも、オレオレ詐欺のように言葉巧みに情報を漏らさせるようなソーシャルフィッシングのような会話術や、フェイクニュース（偽の情報）を流すといった言語や会話による心理的な攻撃が効果的に用いられたことである[24]。このような AI 兵器はまさにサイバー空間とフィジカル空間の融合の産物であり、その攻撃の規模や判断にどのように、そしてどのくらい人間が関与していくべきか国際的な議論となっている。

4. AI 技術の応用において直面する文化間衝突の課題

情報技術は「永遠のβ版（ベータ版：テストの意味)」とする考え方がある。たとえば多くの私たちの生活に今や欠かせなくなったスマートフォンなどに表示される「アプリ」は、定期的にアップデートをする必要がある。AI ブームの端緒となった深層学習は、データや特徴量を自ら見つけ出し学習するという特徴があるが、これにはどこで学習を止めるかという問題がある。人々の嗜好品の消費行動など季節によって変わる項目などは季節ごとに新たなデータを学習し続けたほうが良い精度の分析結果が得られる。あるいは災害時の状況予測や物品の運搬など、刻一刻と状況が変わる場合も、常に最新の情報を基に予測をさせたほうが良い。さらには、COVID-19 のような想定外の現象が起きた場合、人々の消費行動や人流、エネルギー消費量などは既存のデータは使えない。刻一刻と状況が変わる場合においては、最新の情報が多ければ多いほど、AI は適切な判断が下せるようになるため、常に新たなデータを学習させることが良いように思われる。

しかし、学習頻度や速度を上げ、ほぼリアルタイムで更新をし続ける仕組みにしていくと逆に問題が生じることもある。例えば、データの正確さを吟味する時間がなくて結果的に予測や判断の精度が落ちてしまう可能性がある。あるいは、特定の情報や状況に偏ったデータが入ってくることによって、正確性にかける判断をしてしまうこともある。また、与えられた学習データに過剰に適合させようとする「過学習」と呼ばれる現象が起き、未知データに関しては全然当たらない予測をしてしまう場合もある。

---

[23] Shin, M., Kim, J., van Opheusden, B., & Griffiths, T. (2023). Superhuman Artificial Intelligence Can Improve Human Decision-Making by Increasing Novelty. Proceedings of the National Academy of Sciences, 120 (12), e2214840120. doi:10.1073/pnas.2214840120

[24] 小泉悠、ロシアのインテリジェンス機関と ICT、情報処理 vol.61 (7), pp.693-699, 2020。

常に進化し続ける、常に学習し続けるという設計思想からすると、そのAIシステムに完成版という概念はなく、常に現在提供されているシステムやサービスに対してフィードバックをかけて修正・更新をしていくことを受け入れるしかない。長期的に見れば、情報技術に限らずそのほかの工学や医学、薬学においても新しい発見や知見をもとに新たな製品やサービスが提供されることを考えれば、すべてのものが永遠のβ版であると言えるかもしれない。しかし情報技術に関してはその期間が短く、また問題があれば修正や更新をすればよいという、多少の問題があることは織り込み済みで開発がされている。これは、多少の問題があると許されない「安全文化」の領域とは相いれない考え方である。

例えば医療領域においては、臨床試験や実験を何回も繰り返し、治験などで様子をうかがってから治療法や薬剤などを社会へと提供していく。自動運転においても、まずは特区のように限られた空間での研究を積み重ね、公道に出るまではかなりの年月を要した。特に日本ではこのような「安全文化」が根付いており、この石橋を叩いて渡るような文化の一端が「技術で買ってビジネスで負ける」という日本企業のものづくりの在り方に影響を及ぼしているともいえる。

対する「永遠のβ版」の文化では、とにかく先に動かしてみて考えるという試行錯誤が行われる。この、とにかく小さく課題を切り分けて動かしてみる、試行錯誤してフィードバックを得ながら、走りながら考えるという開発の在り方自体は「アジャイル（素早い、機敏な）開発」と呼ばれ、現在システムやソフトウェア開発の主流となっている。昨今、このアジャイルという考え方は技術開発だけではなく、社会や制度設計にも組み込まれるようになってきている。経済産業省が発表した報告書は「アジャイル・ガバナンスのデザインと実装に向けて」と題されている[25]。そこではSociety5.0の実現に向けて、常に変革する社会とゴールに対応するため、個人、市場、法や企業などの関係者が継続的かつ高速に目的設定や評価、改善といったサイクルを回転させていくことが必要であると提案されている。そのためにも、制定などに時間がかかる法規制だけではなく、標準やガイドラインといったソフトローによって官民共同でのルール形成を行っていくことや、「規制のサンドボックス制度」等を活用した実証実験を積極的に行っていくことが提案されている。つまり、技術だけではなく法規制に関しても、技術や社会の変化に対応していくための機敏性が求められる。一方ですぐに変化する法律や標準はそれに対応をする必要があることを考えると現実的ではない。この点からも、第1節で紹介したような、非拘束なフレームワークやツールを用いていくことが急速に変化するAI技術に対応するための一つの解となる。

「永遠のβ版」の考え方が則している領域もあれば、「石橋を叩いて渡る」からこそ信頼される技術領域もあるだろう。交通や医療、インフラなどに情報技術が進出していくにしたがって、「永遠のβ版」や「アジャイル開発」でAIシステムの基本的技術を作っている企業と、具体的な応用領域で「安全」「堅実」なシステムやサービスを望むベンダー企業が意識

---

[25] 経済産業省、2021年、Governance Innovation ver.2: アジャイル・ガバナンスノデザインと実装に向けて、https://www.meti.go.jp/press/2021/07/20210730005/20210730005.html

をすり合わせるだけではなく、それを使う利用者が、どちらの文化圏に属したものだとしてシステムを利用するかによってAIに対する信頼や評価は異なってくる。

日本の内閣府が公開した「AI戦略2022」では、パンデミックや大規模災害に対して差し迫った危機への対応として、国家的危機に対するレジリエンス向上を目的とするAI for National Resilienceと、地球規模の危機に対応するレジリエンス向上を目的としたPlanetary Resilienceの二つの課題が提起されている[26]。これらの危機への対応には国防や防災など「安全文化」との親和性の高い領域も多い。今後、このようなレジリエントな社会をAIなどの技術を使って実現していくとき、応用分化ごとの文脈や慣習、価値観を踏まえてAI技術を活用したシステムやサービスの展開を考えていくことが重要となる。

5. あるマルチステークホルダーディスカッションの試みと今後の展望

2023年春にG7サミットが日本で開催される。AIガバナンスの「原則から実践」に移すにあたって相互運用可能性をどのように考えていくかが一つの論点となりうるだろうとの問題意識のもと、東京大学未来ビジョン研究センター技術ガバナンス研究ユニットの中のAIガバナンスプロジェクトでは、40名近い国内外の産学官の有識者へ個別インタビューを行った。さらには2023年3月6日と7日にOECD.AIの関係者やPartnership on AIなど海外の方もお招きして、ヒアリングにもご協力いただいた日本の研究者、省庁関係者、企業（大企業とスタートアップ含む）、弁護士などの実務家など30名近くの人たちと対面でのワークショップを開催し、3月末に「AIガバナンス協調への道筋：G7サミットに向けた政策提言」を公開した[27]。

様々な意見をすり合わせていく中で、提言は2つの基本的な方針と4つの国際協調を進めるべき活動へと収斂されていった。内容に大枠の合意が得られた後に、具体的な文言を調整する段階で、差しはさまれた単語があったことを本稿の最後に合意形成のプロセスを紹介するものとして本稿に残しておきたい。その単語とは「可能な限り」という単語である。本提言においては、基本方針と国際協調を進めるべき活動の中でそれぞれ1回使われている。

最初の節で指摘した通り、AIガバナンスの議論は価値をめぐるものであり、他の組織や国、機関に一方的に基準を押し付けることは、政治的、技術的、社会的にもできない。しかしAIライフサイクルが異なる組織や国をまたいで展開する中、可能な限り相互に運用可能な仕組みを構築していくことが、各組織や国だけではなく、利用者や社会のためにも必要となる。

マルチステークホルダーの議論の場を形成し、かつそれを急速に展開する技術に合わせ

---

[26] 統合イノベーション戦略推進会議（2022）、AI戦略2022、https://www8.cao.go.jp/cstp/ai/aistrategy2022_honbun.pdf

[27] AIガバナンスプロジェクト（2023）「AIガバナンス協調への道筋：G7サミットに向けた政策提言」、未来ビジョン研究センター政策提言、No.21、2023年3月、https://ifi.u-tokyo.ac.jp/news/15309/、英語はhttps://ifi.u-tokyo.ac.jp/en/news/11267/を参照

ながら議論をしていく試みは重要であるが、容易なことではない。しかし、研究や政策の提言の基盤の一つとなる議論の場の形成は重要であり、そのような場の正当性や信頼性をいかに構築していくか自体も、今後の重要な研究課題である。