AIガバナンスの枠組みの構築に向けて（ver1.0）

2024 年 7 月
デジタル政策フォーラム

基本的考え方

生成 AI を巡る技術開発は著しい速度で進化しており、生成 AI を社会経済システムに実装する動きも加速化している。

こうした中、AI を制御するためのルールづくりは、理念的な議論から具体的な議論へと急速に舵を切っている。欧州[1]や中国[2]では AI 法が成立し、また米国でも大統領令[3]に基づく具体的なガバナンスルールの検討が進んでおり、日本においても法制度の導入を含む議論[4]が本格的に始まろうとしている。

本文書では、こうした生成 AI を巡る動向を念頭に置きつつ、検討の基本的な視点として、

- AI のリスクを最小化するとともに、
- AI の利便性を最大限享受できる環境を整備し、
- こうした環境を可能な限り自律的に実現する生成 AI 市場を創出する

という3つの目的を均衡ある形で実現するための「AI ガバナンスの枠組み」、換言すれば「AI 技術の制御可能性」[5]を継続的に維持するための仕組みの構築に向けた論点を整理する。

AI ガバナンスを議論する上では AI がもたらす便益とリスクのバランスを常に比較しながら議論する必要がある。AIは社会のあらゆる領域で生産性、創造性に寄与するものであり、パーソナル化（インテリジェンスの分散化）を通じて個人のデータ利用に係る主権(sovereignty)を技術的に担保しつつ、利便性の高いサービスを享受できるようになるなど、多種多様な便益をもたらす。他方、AI のリスクという面では、AI が人間を操るリスクや AI が人間を代替することで生じるリスクなどが考えられる。こうしたリスクについては可能な限り技術的解決を目指し、必要以上の規制等の導入はイノベーションを促す観点からは適当でない。

なお、本文書では現時点で一般向けに提供されている生成 AI を主として念頭に置きつつ議論を進めることとし、汎用人工知能(Artificial General Intelligence)は一部を除き対象としない。ただし、AI 技術が引き続き幾何級数的に進化すると見込まれることから、本文書の内容については継続的に更新していくことを基本とする（本文書の末尾「今後の作業計画」を参照）。

I リスクの最小化

(1) リスク管理

AI 管理の手法としてリスク(人の生命や基本的人権に与える負の影響などを含む)を数段階に分けて管理する手法が存在する(例:EU における AI 法ではリスクを4段階に分類[6])。これは AI モデルが内在的に持っているリスクを管理しようとするものであるが、その際、コントロールすべきリスクの範囲をどう画定するか、またリスクをどのような基準でランク分けするのか、広範な議論が必要である。

また、リスク評価は AI そのものを対象とするか、それとも AI の提供事例(ユースケース)ごとに個別に行うのか検討が必要になる。その際、AI に関する規律は、開発者と利用者の二分法ではなく、「AI 開発者」、AI を自らのサービスに組み込む「サービス提供事業者」、「エンドユーザー」の3つのグループに分けられることになる。

さらに、リスク評価は自己評価と第三者評価(例えば監査制度)をどのように組み合わせるのが適当か、議論が必要である。

加えて、AI モデルに対する外的リスクを管理する観点から、AI の脆弱性調査(red teaming)についても、上記のリスク評価と同様に、技術的基準や監査の仕組みについて検討すべきである。

上記に関連して、AI の学習プロセスにおいてデータ汚染攻撃[7]等によって所期の機能を発揮しなかったり誤作動してしまうなどのリスクがある(AI に対するサイバー攻撃)。また、脆弱性の発見やマルウェアの作成、偽アカウントの生成や偽情報の配布等に AI を活用するリスクが顕在化している(AI によるサイバー攻撃)。こうした「AI に対するサイバー攻撃」及び「AI によるサイバー攻撃」への対処についても具体策を早急に検討する必要がある。

(2) 規制の手法と実効性

AI に関する規制の手法としては、ハードロー(法規制)とソフトロー(民間部門による自主規制)、さらにその中間に位置する共同規制(co-regulation)[8]などの手法がある。

例えば、中国や EU においてはハードローを指向し、米国は共同規制を軸にした政策展開が行われている。ただし、ハードローを志向する場合であっても基本法的なアプローチ(理念の明確化とこれを実現するための各主体の役割の具体化等)と具体の行為規制を課すアプローチなど、規律のあり方について一定の幅が存在する。

AI の技術革新を法規制によって阻害することを避けつつ表現の自由などの問題

に対処するためには AI 法（ハードロー）を制定し、技術基準など AI リスクのセーフガードは国を中心に策定することとし、監査基準や偽情報対策については共同規制のアプローチ[9]を採用することが一つの選択肢として考えられる。

また、仮に上記のアプローチを採用する場合、(a)規制の対象となるのは上記(1)の AI 開発者に限定する（サービス提供事業は対象としない）か、(b)AI 開発を事業として営む（他のサービスと一体的に提供して AI 関連事業のコストを賄うケースを含む）場合に限って規制対象とするか、(c)登録制など規制の実効性をどのように担保するか、(d)社会的な影響度を踏まえ一定規模以上のプレーヤーを規制対象とするか等の論点について検討が求められる。

(3) モデル崩壊の可能性

AI が学習データを数次にわたり学習する過程において、出現度の少ないデータを捨象するプロセス（クエリーに対する的中率を向上させることを目的）をとることが多い。この場合、少数意見が切り捨てられるなど当初の AI モデルとは異なるモデルを持つに至る、いわゆる「モデル崩壊」(model collapse)が生じる可能性が指摘されている[10]。こうした状況を放置することは不正確で健全性に欠けるデータを拡散し、データ空間(data space)の汚染(contamination)を進行させることになる。

こうした事態を避けるため、AI の学習データを人間の作成したものに限定する、あるいは学習済み AI であることを対外的に明示する等の一定の規律（認証制度等）について検討が必要である。また、人間の作成したデータをどのように増加させるかという観点からは著作権の切れた文書や公的機関が作成した文書等を広く学習データとして活用可能とするオープンデータ化が有効である。

(4) 生成物の取り扱い

AI は、学習データを取り込み、モデルを形成し、これを活用して生成物たるデータを出力するものである。そこで、「データが改竄されていない」という完全性(integrity)を確保するという観点から見れば、上記(3)は入力値（学習データ）の完全性を確保するという視点であるが、これと同時に出力値（生成物）の完全性を確保するための取り組みも必要になる（入力値と出力値の中間に位置する AI モデルについては次項を参照）。

このため、既に生成 AI を用いて膨大な偽情報が既に流通している状況にある中、共同規制のアプローチを前提としつつ偽情報対策を効果的かつ具体的に推進する方策を検討する必要がある。

その際、AI の生成物が著作権法に適合し正当に作成されたものであることを証明

する電子透かし（digital watermark）の導入が有効と考えられるが、国際的な技術基準、電子透かしを付与する主体の基準、分散型の電子透かし（複数の電子透かしのシステムを相互に承認する柔軟な仕組み）等について、制度運用面及び技術面から検討する必要がある。ただし、このような方策は基本的に技術環境の急速な変化に鑑み、その有効性について継続的な検証と更新が求められる状況にあることに留意が必要である。

（留意点）

なお、上記(1)から(4)の項目については、急速な技術革新が進む中、過去の AI 関連の議論の中には市場の実態からかけ離れ、必要以上に議論が為念的・抽象的なものになる傾向も散見された。このため、規制の実効性や振興策の有効性、利用者保護のあり方等を議論していく上で、AI 開発者等による自主的な情報公開を積極的に促す仕組みを整備し、常に実態の「見える化」を進めていくことが求められる。

II 利便性の向上

(5) デジタル差別の禁止

モデル崩壊（上記(3)）のように AI モデルが公平性・中立性を損なうことで特定の利用者に合理性を欠く差別的取り扱いを行ったり、過度のプロファイリングによって個人のデータ利用に係る主権を超えた「意図せざる」情報開示が発生する可能性がある。

こうしたデジタル差別を禁止するためには、AI モデルの公平性・中立性を確保するための監査制度（自己監査や第三者監査）について検討が求められる。また、AI を組み込んだサービスを利用者に提供する場合において、AI 開発者と（AI を組み込んだ）サービス提供事業者との間の責任分界点についても明確にしておくことが利用者保護の観点から求められる。

(6) AI の積極的活用

AI の活用については既に様々な取り組みが始まっているが、特に深刻な少子高齢化が進む中、教育分野と医療分野においてデータ活用の取り組みが遅れていることを踏まえると、これらの分野における AI 活用を積極的に進める必要がある。

特に教育における生徒、医療における患者を起点として関連するデータを個人の許諾の下に紐づけて解析する仕組みは教育や医療の個別化に貢献することが期

待される。

他方、こうしたデータ連携が過度のプロファイリングを招くことがないよう一定のセーフガード措置も併せて検討する必要がある。また、例えばカルテデータなど、地域や組織によってデータ様式が異なることからデータ連携が進んでいなかった事例についても、AI 解析による自動連携が実現することも期待できる。

さらに、教育や医療の分野の他にも、地球的な課題である環境対策、人の生命財産を守るための防災、豊かな暮らしを実現するための文化などの幅広い分野での AI の積極的な活用を図る必要がある。その際、これらの分野で AI を積極的に活用するために、留意するべき事項や開発するべき技術について検討を深める必要がある。

同時に、学習データとしての個人データの取り扱い、当該データを取り込んだ場合の出力に個人データが含まれる可能性の回避など、プライバシー保護の観点から所要の方策が必要となる。

加えて、AI のリスクについて一般利用者が正しく理解するためのリテラシー教育が重要になる。例えば官民連携による青少年インターネット利用環境の整備の取り組み事例と同様、AI のリスクについても広く周知啓発活動を行うことが重要である。

III 健全な市場の育成

(7) 健全なエコシステムの構築

AI の進化は基本的に民間の創意工夫によって行われるべきである。国はこれを積極的に支援するともに、公共の利益を確保する観点から必要なルール策定や政策支援を行うことを基本とすべきである。

その際、AI の開発者や利用者を含む多様な主体によるエコシステムを確保していくためには、健全な市場環境を確立するための競争政策が重要となる[11]。

そこで、AI 関連市場における参入障壁や、大企業による優越的地位の濫用などの反競争的行為を監視する仕組みを確立する必要がある。

また、現在の大手有力 AI は大規模プラットフォーマーが提供するものが主流となっているが、今後、AI 市場あるいは隣接市場（例えばプラットフォーム事業）において市場支配力が濫用される可能性について検討が必要である。

特にプラットフォーマーのような複数レイヤーで事業展開を行う垂直統合型の AI 開発者は、それ以外の AI 開発者と比して高い市場支配力を持ち、かつ隣接市場への市場支配力を行使する可能性が高いのではないかという懸念がある。その際、競争政策としてどのように対処すべきか検討する必要がある。

さらに、従来のクラウド型と分散型の AI が併存するハイブリッドなネットワーク型 AI が普及していくものと考えられるが、こうした AI の「集中と分散」についてどのように考えるべきか議論が必要である。

なお、EU の AI 法においては法律の域外適用の条項が盛り込まれているが、こうした域外適用が増加することで国外の規制が重畳的に適用されることとなるなど過度の規制をもたらす可能性についても検討が求められる。

(8) オープン性の確保

インターネットが爆発的に普及した主因の一つはそのオープン性にある。同様に、AI についてもクローズドな私権型 AI（proprietary AI）とオープン型の AI(open AI)の２つのアプローチが考えられるが、健全な市場の発展を促すとともに AI 関連サービスの品質を維持する観点からは、十分な競争環境を創出するオープン性の確保が不可欠である。同様のアプローチは欧米でも見られる[12]。

こうした観点から、オープンソースの活用、異なる AI 間の相互運用性の確保をどのように実現するのか、こうした環境を実現するための標準化の促進、オープン型の AI 開発を促すことを前提とした研究開発支援など、政策的な観点から検討すべき課題について、今後さらに議論を深める必要がある。

また、AI 関連の技術開発について日本はグローバル市場において既に遅れをとっている状況にある中、オープン型の AIを組み込んだソリューションの開発などを国が支援するなど、オープン型の AI に対して積極的な振興策を講じることを検討すべきである。特に AI 系のベンチャー支援のための取り組みを強化するための議論が必要である。

(9) 国際的コンセンサスの醸成

AI は国内に閉じて開発・利用されるものではなく、ネットワーク化されサイバー空間で広く利用されることが前提となる。その際、上記の論点については国際的に緩やかなコンセンサスを形成しながら、各国の法制度などのルールに反映し、必要な調和を図っていくことが求められる。その際、AI がグローバルサウスの抱える課題解決に貢献する可能性が大きいことを踏まえ、グローバルサウスの十分な参加を得た形で進めることが求められる。

また、こうした国際的コンセンサスの醸成の中で特に急務なのが、AI の軍事利用に係る規範の形成である。2023 年 2 月にハーグで開催された「軍事領域における責任ある AI に関する会議」(REALM Summit)において提案した「人工知能及び自律性の責任ある軍事利用に関する政治宣言」[13] にあるような AI 利用に関する自主的な

コミットメントを拡大していく必要がある。同時に国連の安全保障の枠組みの中で AI セキュリティ監査（査察）の仕組みを取り入れることも検討に値する。こうした AI と安全保障のあり方について、既に AI の軍事利用が現実化[14]していることを踏まえて議論を急ぐ必要がある。

（10）倫理的問題への対処

AI の急速な進歩に伴い、将来的に「自意識」を持つ AI の可能性も考慮に入れる必要がある。このため、生命科学分野と同様に、AI 研究に関する倫理的問題を検討し、具体的な研究倫理規定や研究承認プロセスを確立すべきである。例えば、「AI に自意識を持たせること」や「自己複製や改変能力をどこまで持たせるか」といった問題に対する倫理的指針を策定し、実装していく必要がある。

今後の作業計画

本文書は、日本における AI 関連法制の議論やグローバルに広がりを見せている AI ガバナンスの議論の方向性を示すことを目的としている。

本フォーラム（DPFJ）は本文書を基に AI の技術・政策・利活用という3つの観点から有識者のヒアリングを継続し、本文書の更新を継続的に行う。併せて、本文書の更新機会などを捉えてオープンフォーラムを開催するなど、広く AI ガバナンスの枠組み構築に向けた議論を深め、概ね本年末を目処に最終的な文書に取りまとめる予定である。その際、同様の議論を進めている他のフォーラム等との連携を積極的に進め、コンセンサスの醸成を図っていくこととしている。

以　上

---

[1] 2023 年 6 月、欧州議会は「AI 法」を採択。2024 年 5 月から段階的に施行されており、全面適用は 2026 年夏の予定。
https://artificialintelligenceact.eu/

[2] 2023 年 8 月、中国は「生成人工知能サービス管理のための規則」を施行。本規則（第4条）では、法律や行政規則で禁止されているコンテンツの作成を禁止しており、「社会主義の中核的価値観を遵守」する生成物のみが認められている。
（出典）原田雅史「中国「「生成人工知能サービス管理暫定弁法」の制定とその解説」企業法務ナビ（2023 年 7 月 21 日）
https://www.corporate-legal.jp/matomes/5362

[3] 2023 年 10 月、米国政府は AI ガバナンスに関する大統領令を公表。政府機関において取り組むべき施策として、脆弱性調査（Red Teaming）の基準策定、アルゴリズムによる差別禁止のための明確なガイダンスの策定、ヘルスケア、教育等の分野における AI 適正利用の支援などの内容を盛り込んでいる。
https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

[4] 例えば AI 戦略会議「「AI 制度に関する考え方」について」(2024 年 5 月)を参照。
https://www8.cao.go.jp/cstp/ai/ai_senryaku/9kai/shiryo2-1.pdf

[5] 「AI 技術の制御可能性」には技術の開発面と利用面(社会実装)の2つの局面がある。本文書では、便益とリスクのバランスを考慮するものである以上、実際の利用面における制御可能性(一定の制御を加えつつ積極的に利用する)を重視する立場をとる。ただし、後段の項目(10)においては利用面のみならず開発面まで含まれる議論となる。

[6] 欧州 AI 法は、AI のリスクを許容できないリスク(人の生命や基本的人権に対する直接的脅威を及ぼすものとして開発を禁止)、ハイリスク(事前の適合性評価、データベースへの登録等の義務)、限定リスク(AI とのやり取りであることを利用者に知らせる透明性確保の義務)、最小リスク(規制なし)の4種類に分類している。

[7] データ汚染(data poisoning)攻撃では、学習データに間違った出力を生じさせる汚染データを挿入し、モデルが悪意をもって機能するように修正することを試みる。また、データ回避(data evasion)攻撃では、人間の知覚できないノイズ等を学習データに混入させて AI の判定結果を誤らせる。

[8] 共同規制とは、国がルールの基本方針を示し、その趣旨に賛同した事業者が基本方針に基づくルールを運用して運用結果を国に報告、国はこれを評価して必要に応じて基本方針を修正するという方式。欧州においてはプラットフォーマーの偽情報対策などで採択されている。
なお、AI 分野における共同規制の例としては、大統領令(脚注3参照)の公表に先立つ 2023 年 7 月、大統領府と AI 関連7社(Amazon, Anthropic, Google, Inflection, Meta, Microsoft & OpenAI)との間で非拘束の合意(AI 開発における安全性、セキュリティ、信頼の確保という3つの項目から各社が取り組む内容を具体化)がなされた事例がある。なお、同年 9 月、上記7社に加えて8社(Adobe, Cohere, IBM, Nvidia, Palantir, Salesforce, Scale AI, Stability)が本合意に加わった。

[9] 共同規制のアプローチを採用する場合、プレーヤー間の発言力の格差、法規制によらないことによる透明性・公開性の低下等が懸念されることから、適切な公的サポートを運用方針に明確に盛り込むことが必要となる。

[10] I. Shumailov et al. "The Curse of Recursion: Training on Generated Data Makes Models Forget" arXiv (May 2023)
https://arxiv.org/abs/2305.17493

[11] OECD "Artificial Intelligence, Data and Competition" OECD Artificial Intelligence Papers No. 18 (May 2024)
https://www.oecd.org/daf/competition/artificial-intelligence-data-and-competition.htm

[12] 欧州では、2024 年 1 月に公表された招請文書"Competition in Virtual Worlds and Generative AI: Calls for Contribution"において、生成 AI と競争政策に関する論点リストが提示されている。
https://ec.europa.eu/commission/presscorner/detail/en/ip_24_85
また、米国では大統領令(脚注3参照)の中でイノベーションや競争を促す観点から、「公正・オープン・競争的なエコシステムの促進」を主要推進項目の一つに挙げている。

[13] 本提案(US DoS "Political Declaration on Responsible Use of Artificial Intelligence and Autonomy" (February 2023))では、軍事 AI が国際法(特に国際人道法)の義務に合致した形でのみ使用されることを前提とし、軍事 AI について設計・開発・配備・使用に関する原則の公表、意図しない偏りを最小化する対策の実施、監査可能な軍事 AI の開発、軍事 AI の安全性・セキュリティ・有効性についてライフサイクル全体にわたる厳格なテストと保証を行うこと等について国が自主的にコミットすることをその内容としており、現在、日本を含む51か国が賛同している。
https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/

[14] 2024 年 4 月のイスラエルのネットメディア「+972 マガジン」による調査報道によれば、イスラエル軍は生成 AI「ラベンダー」を用いてガザ地区の 3 万 7 千人を抽出して工作員リスト化し、これをもとに標的として攻撃する等の行為が行われている。
(出典)Yual Abraham "Lavender': The Ai machine directing Israel's bombing spree in Gaza" +972 Magazine (April 3, 2024)
https://www.972mag.com/lavender-ai-israeli-army-gaza/
上記調査の詳細については川上泰典「ガザの3万7千人を標的化:AI マシーン「ラベンダー」の存在明らかに」Yahoo!ニュース(2024 年 4 月 9 日)も参照。
https://news.yahoo.co.jp/expert/articles/c72d4cbc32aa5577eac494dfd75b43652a20555f