United Nations University

TECHNOLOGYBRIEF
No. 5, APRIL 2024

AI ガバナンスのための枠組み

Tshilidzi Marwala, United Nations University, Tokyo, Japan

提言

- 人権と国連憲章の原則に沿ってAIの価値観を定義する。
- 行動科学を用いることにより、AIの可能性を最大限に引き出しながら、関連リスクを抑える文化を醸成する。
- メカニズムデザイン分野の知見に基づく戦略を実施することにより、AIの潜在的利益を最大限に引き出しながら、関連リスクを抑える文化を育成する。
- AIを規制するための枠組みと制度的ガバナンス組織を導入する。
- AIを管理するための政策と規則を確立する。
- AIに関する標準を確立する。
- AIに関する法規を策定する。
- データ、アルゴリズム、コンピューティングシステム、AIの活用を管理する。

はじめに

AI 技術の急速な進歩と導入は、大きな経済的、社会的、倫理的影響をもたらしている[1]。効率的なガバナンスは AI の恩恵を最大限に引き出しながら、そのリスクを最小限に抑えるために不可欠である[2,3]。本技術ブリーフでは、政策立案者に向けて重要なガバナンスの問題について簡潔にまとめ、AI 技術を適切に監視するための戦略的方法論を提案し、AI ガバナンスのための枠組みを提示する。

AI の価値観

AI のガバナンスは健全な価値観に基づかなければならない。本セクションではそうした価値観の一部を取り上げる。

透明性：AI の透明性は信頼と説明責任と公平性を促進する[4]。透明性のある AI システムは、ステークホルダーが意思決定プロセスを理解し評価することを可能にし、バイアスを特定して修正し、倫理と法律の遵守を確保できるようにする。透明性のある AI はその活用における安全性と信頼性

を高めるが、国民の安全のために透明性のある AI の行動が不可欠な医療と交通輸送の分野において、とりわけ重要である。AI システムの透明性を高めると、開発者とユーザーの協力が促され、技術の進歩につながる。こうした公開性は、ユーザーエンゲージメントと規制の遵守を改善し、AI に関し正しい情報に基づく生産的な議論を促す。

真実性：信頼されるためには、AI は真実を伝えなければならない[5,6]。AI が正確でバイアスのないデータを生成すると、メディア、教育、科学において信頼できる意思決定ツールとなる。信頼できる AI は誤情報を防止し、世論において批判的思考を促す。AI が正確なアウトプットと結びつくと、倫理規範が満たされ、損害や操作される危険性を減らせる。このように真実性を追求することで、AI 技術は広範に受容され、有益な形で日常生活に統合されるために必要な AI システムの正当性や社会的信頼が強化される。

安全性とセキュリティ：システムはさまざまな形で人間と相互に作用し、重大な影響を及ぼすため、その安全性がきわめて重要になる[7]。安全な AI は人々と社会を守り、信頼と受容を築く。安全性を優先することで、AI を活用している医療や交通輸送、金融などの産業において、事故やエラーを減らし、命と財産を守ることができる。安全な AI 技術は倫理と法律の遵守も促し、社会の安定性を高め、テクノロジーの悪用を防ぐ。AI の開発と展開における安全性は、社会がこの技術から確実に恩恵を受けられるようにする。

倫理：社会に恩恵をもたらし、損害を最小限に抑える AI 技術を設計、展開するには、倫理が必要である[8]。AI の倫理には正義、透明性、説明責任、プライバシー、人権が含まれる。倫理的価値を組み込むことで、差別を防止し、AI システムを適切に利用できるだろう。倫理的な AI はユーザーとステークホルダーの信頼、規制の遵守、持続可能なイノベーションを強化する。AI の開発者と運用者は倫理的行動を重視することで、テクノロジーが尊厳と自由を損なうことなく、人間の能力を高める公正な社会の確立を後押しできる。

プライバシー：個人データを保護してテクノロジーに対する信頼を維持するために、AI はプライバシーを尊重しなければならない[9]。AI システムは慎重な取り扱いが必要な個人情報を含め、膨大な量のデータを処理することから、プライバシーの侵害と搾取につながる可能性がある。厳格なプライバシー規則を用いることで、AI は不正アクセスと搾取から個人データを保護し、AI 技術への信頼を高める。

ガバナンスモデル

図 1 は本ブリーフで提唱する AI ガバナンスの実行階層である。図が示すように、AI ガバナンスは前セクションで説明した AI の価値を基礎としなければならない。これらの価値の上に人間の行動、インセンティブとディスインセンティブのためのメカニズム、制度的ガバナンス組織がくる[10]。ガバナンスの役割を持つ組織の例として、気候変動に関する政府間パネル（IPCC）や国際原子力機関（IAEA）があげられる[11]。その上にくるのは政策と規則で、企業レベルのものや医療分野などの専門家団体内のもの、政府レベルや国際レベルのものが考えられる。標準は国家および国際レベルで策定され、策定に当たって分野特有の専門性が求められる標準や、政策に関する専門性が必要な標準もある。その上にくるのは法律である。たとえば、人権に影響を与える AI ガバナンスは、規則ではなく法律によって統治されなければならない[12]。この AI ガバナンスの実行階層は、**図 2** に示す AI ガバナンスモデルにつながる。

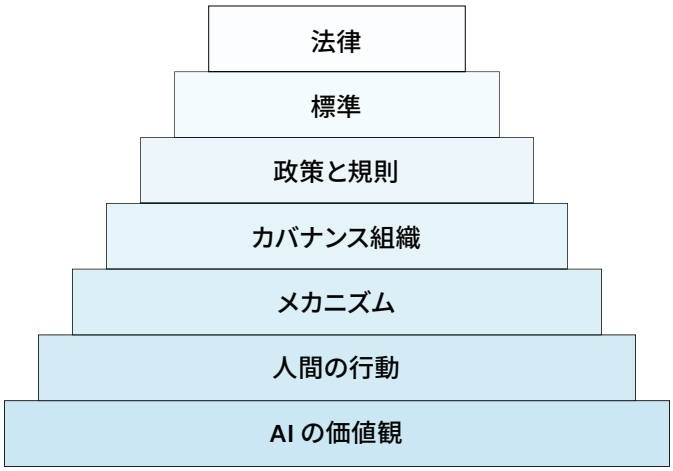

図 1 — AI ガバナンスの実行階層

AI ガバナンスの分野

AI ガバナンスモデルは AI ガバナンスの実行階層に対応しており、構造としてはデータがアルゴリズムに、アルゴリズムがコンピューティングに、コンピューティングが AI の活用に供給される。

データ：AI データは倫理的で、正確で、安全な利用のために管理されなければならない。頑健なガバナンスシステムは、データの収集、保存、分析（合成データなど）[13,14]、共有（越境データ流通など）[15] を監視しながら、機密データを保護して悪用を防止することを可能にする。優れたデータガバナンスがあれば、AI システムは高品質でバイアスのないデータを利用して、公平で生産的な判断を下せるようになる。優れた

データガバナンスは、規制的基準が満たされるようにし、AI技術への公共の信頼を維持することにも役立つ。データガバナンスは公開性と説明責任を促すことで、ステークホルダーがAIの判断を理解し、AIの判断について議論できるようにする。十分に管理されたデータは、AI活用の信頼性と説明責任と確実性を高める。

AIアルゴリズムのガバナンス：AIアルゴリズムは倫理的行動と偏りのない結果を確保するために管理されなければならない[16]。AIによる判断は医療、銀行業務、法執行などにますます取り入れられるようになっており、個人と社会に大きな影響を及ぼしうる[17, 18]。効果的な制度的ガバナンス体制は、倫理的で透明性のある責任ある行動を確保できるよう、AIアルゴリズムの開発、展開、モニタリングを後押ししうる。アルゴリズムの選定、設計、訓練、検証を管理する必要がある。こうしたガバナンスは、アルゴリズムのバイアスや差別を防ぎ、ユーザーの信頼を高める。アルゴリズムのガバナンスはまた、AI技術を文化的理念や法的義務に結び付けることで、テクノロジーの公平性と正義を促す。

コンピューティングのガバナンス：AI技術が安全で倫理的で有益なものになるように、コンピューティングを管理する必要がある[19]。コンピューター技術が個人データ管理から重要インフラまで、生活のあらゆる分野に浸透している中、明確なガバナンスの枠組みは悪用を防ぎ、リスクを緩和し、倫理と法律の遵守を確保する。ガバナンスは、プライバシーを尊重し、セキュリティを改善し、被害を防ぐコンピューティング技術の設計と利用に役立つ。明瞭なガバナンスメカニズムは、社会的信頼を構築し、テクノロジーへのより平等なアクセスを促す。半導体チップ、エッジコンピューティング、クラウドコンピューティング、アンビエントコンピューティング、量子コンピューティング、そしてコンピューティングの際に使用される電力や水量も管理する必要がある。

AI活用のガバナンス：AIによる膨大な影響と混乱を起こす危険性に対処するために、AIは社会、経済、政治の分野全体で管理されなければならない[20]。AI技術は労働市場、経済発展、政治的意思決定、公共政策に影響を与える。それゆえ、強固なガバナンスメカニズムが必要である。こうした管理により、AIは社会の健全性、経済的公正性、民主的プロセスを損なうことなく、それらを改善できる。効果的なガバナンスは悪用を防ぎ、予期せぬ影響を減らし、利益の公平な分配を保証する。ガバナンスはまた、AIの活用を倫理や規制要件に沿わせることで、公共の信頼を維持し、日常生活へのAIの持続可能な統合を促進する。

AIガバナンスの分野を**図2**に示す。

図2—AIガバナンスモデル

効果的なAIガバナンスのための提言

人権と国連憲章の原則に沿ってAIの価値観を定義する：AIの価値観を定義する際、AI開発が人間の尊厳と平等と自由を確実に支持するように、人権と国連憲章に沿わせる必要がある。AIに公平性と透明性と説明責任を組み込むことで、差別とプライバシーリスクを最小限に抑えることができる。さらに、国連憲章が重視する平和と正義と協力を取り入れることで、AIに世界的な課題への対処を促し、倫理的な行動と、保護的な標準づくりに対する国際的な合意形成が進む。

行動科学を用いることにより、AIの可能性を最大限に引き出しながら、それに伴うリスクを抑える文化を醸成する：AIに関して行動科学を用いることは、AIの恩恵を最大限に引き出しながら、リスクを低減する文化を形成するための知見を利用することを意味する。組織は、ナッジ（理想的な行動をそっと促す仕組み）と構造化されたインセンティブを導入することで、倫理的なAIの利用を促し、有害な行動を妨げることができる。このアプローチでは、プライバシー保護措置とバイアス緩和策の導入を促すことで、AI開発が安全で生産的な人間の相互作用に合致するようにする。

メカニズムデザイン分野の知見に基づく戦略を実施することにより、AIの潜在的利益を最大限に引き出しながら、それに伴うリスクを抑える文化を育成する：AIにおいてメカニズムデザイン戦略を実施すれば、個々のインセンティブを社会の目標に整合させることによって、恩恵を最大限に引き出し、リスクを最小限に抑えることが可能になる。経済学とゲーム理論に

おけるメカニズムデザインという分野は、ステークホルダーが透明性と公平性と説明責任を支持するように促す、そんなシステムの創出に役立つ。たとえば、バイアスを防ぐために、アルゴリズムが正確なデータや倫理的行動を報奨することが考えられる。このアプローチでは、セーフガードのある枠組み内で AI を稼働させることにより、AI システムの信頼性と有効性を高める。

AI を規制するための枠組みと制度的ガバナンス体制を導入する：AI を効果的に規制するには、IPCC や IAEA に類する組織を確立することが望ましい。こうした組織は世界的な AI 標準を策定し、その遵守を監視し、国際協調を促進する。また、倫理指針、技術基準、安全手順に取り組み、優れた慣行を共有して透明性を促進するためのプラットフォームを提供する。そうした組織は、AI の開発と展開を安全で、倫理的で、世界的に恩恵をもたらすものにするだろう。

AI を管理するための政策と規則を確立する：AI を管理するための政策と規則を確立することは、倫理的な AI の利用を確保するための法的枠組みの策定を意味する。そうした法的枠組みは透明性と説明責任と公平性に重点を置きつつプライバシーを保護して差別を防止しなければならない。重要な施策としては、AI に関する監査の義務化や厳格なデータ保護要件の確立が考えられる。政府は AI が社会に恩恵をもたらし、倫理基準を守るよう、AI に関する行動を監視し、規則を遵守させるための専門組織を創設してもよいだろう。

AI に関する標準を確立する：AI に関する標準の確立は、AI の設計、開発、展開に関する統一的な指針を策定することを意味し、その際は技術的な品質、倫理的に考慮すべき事項、互換性に焦点を合わせる必要がある。こうした標準ではデータプライバシー、アルゴリズムの透明性、セキュリティ、バイアス防止を扱わなければならない。こうした指針は産業界、学界、規制当局が共同で策定して、AI 技術の世界的な安全性、有効性、倫理の遵守を確保するために、定期的に更新していかなければならない。

AI に関する法規を策定する：AI に関する法規の策定では、AI の開発、使用、影響を管理し、倫理的な運用、プライバシー保護、差別防止を確保するための法律を起草する。立法者は専門家や国民と協力して、柔軟性があり、正しい情報に基づく規定を策定すべきである。そうした法律によって、AI 技術の浸透に合わせて個人の権利と社会の福祉を保護するために、法律の執行と紛争の処理を扱う監視機関も設立されるとよいだろう。

データ、アルゴリズム、コンピューティングシステム、AI 活用を管理する：データ、アルゴリズム、コンピューティングシステム、そして AI の活用には、責任ある倫理的な利用を確保するための強固な枠組みが必要である。このガバナンスには、データ保護措置、バイアスを防ぐためのアルゴリズムの透明性、コンピューティングシステムのための強固なセキュリティ基準、医療や金融などの重要分野における AI 活用に関する明確な規定が含まれなければならない。そうした包括的なアプローチは、公共の信頼を築き、技術革新が社会に貢献しながらリスクを最小限に抑えられるようにする。

結論

AI ガバナンスの枠組みは、技術の進歩に対応できる機動的で柔軟なものでなければならない。政策立案者は技術者、産業界、学界、市民社会と協力して、AI を公共の利益のために確実に機能させるための包括的なガバナンス戦略を作成する必要がある。継続的な対話と反復的な政策開発は、変化する AI 技術の環境と社会への影響に対応するために不可欠である。

ENDNOTES

1 Marwala, T., 2022. *Closing the gap: The fourth industrial revolution in Africa*. Pan Macmillan South Africa.
2 Marwala, T., 2023. *Artificial Intelligence, Game Theory and Mechanism Design in Politics* (pp. 41-58). Singapore: Springer Nature Singapore.
3 Roberts, H., Hine, E., Taddeo, M. and Floridi, L., 2024. Global AI governance: barriers and pathways forward. *International Affairs*, p.iiae073.
4 Ali, A.E., Venkatraj, K.P., Morosoli, S., Naudts, L., Helberger, N. and Cesar, P., 2024. Transparent AI Disclosure Obligations: Who, What, When, Where, Why, How. *arXiv preprint arXiv:2403.06823*.
5 Hurwitz, E. and Marwala, T., 2007, October. Learning to bluff. In *2007 IEEE International Conference on Systems, Man and Cybernetics* (pp. 1188-1193).
6 Markowitz, D.M. and Hancock, J.T., 2024. Generative AI are more truth-biased than humans: A replication and extension of core truth-default theory principles. *Journal of Language and Social Psychology*, 43(2), pp.261-267.
7 Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C. and Jirotka, M., 2021. Governing AI safety through independent audits. *Nature Machine Intelligence*, *3*(7), pp.566-571.
8 Coeckelbergh, M., 2020. *AI ethics*. MIT Press.
9 Elliott, D. and Soifer, E., 2022. AI technologies, privacy, and security. *Frontiers in Artificial Intelligence*, *5*, p.826737.
10 Marwala, T., 2024. *Mechanism Design, Behavioral Science, and Artificial Intelligence in International Relations*. Morgan Kaufmann.
11 Verbruggen, A. and Laes, E., 2015. Sustainability assessment of nuclear power: Discourse analysis of IAEA and IPCC frameworks. *Environmental Science & Policy, 51*, pp.170-180.
12 Marwala, T. and Mpedi, L.G., 2024. *Artificial intelligence and the law*. Palgrave Mcmillan.
13 Marwala, T., Fournier-Tombs, E. and Stinckwich, S., 2023. The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development. *arXiv preprint arXiv:2309.00652*.
14 Sidogi, T., Mongwe, W.T., Mbuvha, R. and Marwala, T., 2022, December. Creating synthetic volatility surfaces using generative adversarial networks with static arbitrage loss conditions. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1423-1429).
15 Tshilidzi Marwala, Eleonore Fournier-Tombs, Serge Stinckwich, "Regulating Cross-Border Data Flows: Harnessing Safe Data Sharing for Global and Inclusive Artificial Intelligence", UNU Technology Brief 3 (Tokyo: United Nations University, 2023).
16 Ebers, M. and Gamito, M.C., 2021. *Algorithmic Governance and Governance of Algorithms*. Springer.
17 Marwala, T., 2014. *Artificial intelligence techniques for rational decision making*. Springer.
18 Marwala, T. and Hurwitz, E., 2017. *Artificial intelligence and economic theory: skynet in the market (Vol. 1)*. Cham: Springer International Publishing.
19 Sasikala, P., 2012. Cloud computing and e-governance: Advances, opportunities and challenges. *International Journal of Cloud Applications and Computing (IJCAC), 2*(4), pp.32-52.
20 Marwala, T., 2023. *Intelligence, Game Theory and Mechanism Design in Politics* (pp. 135-155). Singapore: Springer Nature Singapore.

本稿について

本研究について
この技術ブリーフは、グローバル・サウスと持続可能な開発に関連したグローバルなテクノロジーガバナンスの特定領域に焦点を当てる国連大学の一連の技術ブリーフの第 5 弾である。

著者情報
チリツィ・マルワラ教授は東京に本部を置く国連大学の第 7 代学長であり、国連事務次長を務めている。人工知能（AI）の専門家であり、前職はヨハネスブルグ大学（南ア）の副学長である。マルワラ教授はこれまで 300 件以上もの雑誌記事や新聞寄稿を提供し、27 冊にのぼる AI とその関連分野の書籍を著し、5 つの特許を共同取得している。博士はまた、アメリカ芸術科学アカデミーに所属しているほか、世界科学アカデミー（TWAS）、南アフリカ科学アカデミーなどにフェローとして所属している。

引用の際の表記
Tshilidzi Marwala, "*Framework for the Governance of Artificial Intelligence*", UNU Technology Brief 5 (Tokyo: United Nations University, 2024).

ISBN 978-92-808-9158-4