United Nations University

TECHNOLOGY BRIEF
No. 4, FEBRUARY 2024

アルゴリズム・バイアス：回避可能なものと不可能なもの

Tshilidzi Marwala, United Nations University, Tokyo, Japan

政策提言

回避可能なアルゴリズム・バイアスへの対策

1. 多様な人口集団を正確に代表するため、AI システムの訓練データの多様性を最大限に高める。
2. 意思決定プロセスの検証を可能にするため、アルゴリズムの透明性を最大限に高める。
3. バイアスを検出してこれに対処するため、AI システムの定期監査を行い、その結果を公開して説明責任を確保する。
4. AI システムの開発時には、包摂的な AI 開発チームによって多様な視点が組み込まれるようにする。

回避不可能なアルゴリズム・バイアスへの対策

5. 存在する不可避なバイアスとそれを軽減する手段など、AI システムの内在的な限界について、ステークホルダーが十分な情報を得られるようにする。
6. 不可避なバイアスが存在する場合に許容可能なトレードオフと意思決定基準を説明する政策的枠組みを確立する。
7. 重大な状況において AI システムの評価と承認を行う独立した倫理監視委員会を設置し、不可避なバイアスに対して倫理的な対処がなされるようにする。
8. 変化し続ける文化規範と整合させるため、AI アルゴリズムの継続的なモニタリングと改善を行う。

はじめに

アルゴリズム・バイアスは、人工知能（AI）と機械学習という急成長分野において大きな課題となっている[1,2]。意図していなくてもアルゴリズム・バイアスがさまざまな形で現れて、特定の個人や集団に対して不当に不利益を与える差別的な結果を生む可能性がある。

テクノロジーと倫理規範との関係が進化し続けるなか、回避可能なアルゴリズム・バイアスと不可避なアルゴリズム・バイアスの違いを理解することは、政策立案者、開発者、消費者にとってますます重要となる。興味深いことに、調査によれば人々は、現時点では人間のバイアス（偏見）ほどにはアルゴリズム・バイアスに悩まされていないようだが、こうした考えは時間とともに変わっていくかもしれない[3]。

回避可能なアルゴリズム・バイアスは、きめ細かなシステムデザイン、データ分析、包摂的な開発を通じて削減することが可能な事象である[4]。回避可能なバイアスは、偏ったデータセット、不完全なアルゴリズム設計、開発プロセスにおける不注意に由来する。これらのバイアスに対処するためには、多様性、開放性、および AI システムの展開における継続的な監視を重視する積極的な戦略が必要となる。

これに対し、回避不可能なアルゴリズム・バイアスは、相反する公平性の原則、データの複雑さ、または既存の AI 技術の限界により、解消することが困難なバイアスを指す。これらのバイアスは倫理面および実務面で複雑な問題を生み、対立する利害の間で慎重に均衡をとること、ならびに、真摯な取り組みをしてもある程度のバイアスは残存する可能性があると認めることが必要になる。

回避の可不可に関わらず、アルゴリズム・バイアスへの対処は、技術的な問題であると同時に、社会的な義務でもある。政策立案者、エンジニア、倫理学者、および社会は、AI システムの知性と効率性と公平性を担保するために協力しなければならない。科学技術の進歩におけるこの重要な瞬間に私たちが下す政策決定は、AI が社会に対して及ぼす長期的な影響を左右する可能性がある。

回避可能なアルゴリズム・バイアスの原因

意思決定プロセスの自動化が進むにつれて、アルゴリズム・バイアスの影響はますます重大になる[5,6]。これらのバイアスは、限られたデータ収集、アルゴリズム設計時の仮定、および多様な集団がアルゴリズムによってどのように扱われるかということへの配慮の不足により生じる[7]。

例えば、過去の差別的な慣行を反映した雇用データを用いて訓練された AI システムは、特定の人口集団を有利に扱い、他の集団を差別することによって、同様の偏見を永続させる可能性がある。社会のさまざまな分野で AI の普及が進むなか、自動化システムの信頼性、公正性、平等性を高めるために、回避可能なアルゴリズム・バイアスを特定して除去することがきわめて重要である。

バイアスを回避するための戦略

回避可能なアルゴリズム・バイアスに対処するためには、透明性、多様性、継続的なモニタリングに重点を置いた積極的な戦略が必要である[8]。何よりもまず、AI システムの訓練に使用されるデータを、包括的かつ多様で、関連する人口区分をすべて含んだものにすることが不可欠である。また、広範な視点を提供して潜在的なバイアスの早期検出に貢献できる倫理学者、社会学者、各領域の専門家からなる多角的な設計チームを構成しなければならない。

AI アルゴリズムの透明性を確保することもきわめて重要である。AI システムの意思決定プロセスをアクセス可能で説明可能なものにすることによって、バイアスを特定し、解消することが可能になる。しかし、正確性と透明性のジレンマを解決するには技術的なブレイクスルー (突破口となる技術革新) が必要かもしれない。現在の AI システムにおいては、アルゴリズムの正確性が高まるにつれて透明性は低下し、その逆もまた同様である[9]。

バイアスや偏見をつねに監視するためには、AI システムの定期的な監査と更新も必須である。AI の開発と導入に関し明確な規範や原則を確立する、強固な法的および倫理的枠組みを持つことが不可欠である。説明責任と包摂性を優先する文化を育むことによって、回避可能なアルゴリズム・バイアスの潜在的リスクを大幅に軽減しながら、AI 技術のメリットを効果的に活用することができる。

取り除けないバイアスもある

データ、AI 技術、および人間の行動の相互作用にはある程度のアルゴリズム・バイアスが内在しており、それらを完全になくすことは不可能で、せいぜい最小限に抑えることしかできない[10]。一部のバイアスの永続的性質は過去のデータに現れた社会的偏見に起因している場合があり、AI システムはこうしたデータから知識を得るため、以前から存在するバイアスを意図せずに強めてしまう。上述したように、データ・キュレーション（データの選択・編集・管理）、透明性のあるアルゴリズム設計、継続的なモニタリングによって大きな進展を図ることはできるが、現実的に目指すべきはバイアスの解消ではなく削減である。

このような制約を認めることは、AI システムに対するより責任ある良心的なアプローチを推進するうえで不可欠である。完全にバイアスのないアルゴリズムは原則として実現不可能だということを認めたうえで、継続的な改善とバイアスの地道な削減に重点を置くべきである。

不可避なアルゴリズム・バイアスへの対処

一部のアルゴリズム・バイアスは回避可能だが、相反する道徳的枠組みや文化的および技術的システムに深く根付いた構造によって回避不可能なバイアスがある程度残存する。これらの内在的なバイアスは、社会現象の複雑さ、公正さに対する多様な概念、絶えず変化する社会規範に由来する。人類の誕生時から存在する公平性の概念は、本質として主観的なものである。何をもって公平とするかは、集団や個人によって異なる。アルゴリズムは、公正性を求める過程でしばしば相反する概念に遭遇する。ある公平性に最適化することは、別の公平性に対するバイアスに意図せずつながる可能性があり、完全な平等を追求することは不可能と思われる。

さらに、社会規範は絶えず進化する[11, 12]。現時点では公平性を表しているアルゴリズムも、社会規範の変化によってやがてバイアスのあるアルゴリズムになってしまうかもしれない。こうした絶えず変化する情勢によって、公平性の追求は最終目標ではなく終わりのない探求となる。すなわち、一度きりの達成ではなく持続的な進歩なのである。

内在的なバイアスに取り組むためには、アルゴリズムの公平性を担保するためのアプローチを根本的に変える必要がある。一度で解決するのではなく、柔軟で継続的な手順として考えるべきである。変化する文化規範や公平性の理解に合わせて、絶えずアルゴリズムのモニタリングと改善を続けなければならない。

製薬業界の副作用対処に倣ったアルゴリズム・バイアスへのアプローチ

製薬産業の副作用への対処戦略を検証することで、アルゴリズム・バイアスへの取り組み方についての有益な視点が得られる[13]。製薬業界では、副作用の全くない薬はないということが広く認識されている。この認識は、バイアスの全くないアルゴリズムはないという考えと類似している。薬と同様に AI システムも、リリース前に潜在的に有害な影響を検出、測定、軽減するため、厳密な試験段階を経なければならない。AI システムは、バイアスを発見して削減するために、包括的な検査と継続的な監視を必要とする。

医薬品におけるインフォームドコンセントの原則には、起こり得る悪影響について患者に伝えることが含まれるが、AI の分野にもアルゴリズムの透明性という同様の原則がある[14]。これは、AI システムの意思決定プロセスとそのシステムに内包する可能性のある潜在的なバイアスについての情報を、ユーザーに提供することを求めるものである。

さらに、上市後の医薬品の実際の効能を継続的に監視するファーマコビジランス（医薬品安全性監視）のように、AI システムも継続的に知識を獲得し、調整し、能力を強化するための同様のプロセスを必要とする[15]。これは、生じる可能性のあるバイアスを検出、修正するために不可欠な過程である。このアプローチは、アルゴリズム・バイアスへの対処において積極的で透明性のある倫理に基づいた枠組みが重要だということを強調している。

アルゴリズム・バイアスの取り扱い手順

アルゴリズム・バイアスへの対処手順の案を**図 1** に示した。最初の質問は、「訓練された AI モデルにバイアスがあるか」というものである。答えが「いいえ」なら、AI モデルは展開される。答えが「はい」なら、次の質問は、「そのアルゴリズム・バイアスは回避可能か」というものである。

答えが「はい」なら、アルゴリズム・バイアスを最小化するためのメカニズムが導入され、残存するバイアスが定量化されて展開時にユーザーに報告される。答えが「いいえ」なら、バイアスの程度が定量化され、その結果をユーザーに報告したうえで AI モデルが展開される。

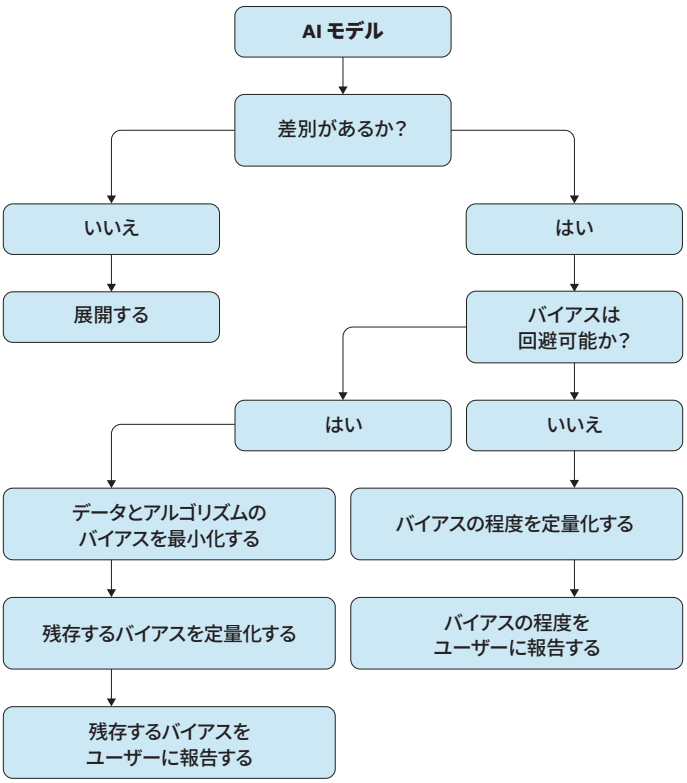

図 1 回避可能および不可避なアルゴリズム・バイアスの取り扱い手順

結論

AI が社会に大きな変化をもたらすにつれて、回避可能なアルゴリズム・バイアスと不可避なアルゴリズム・バイアスの両方に対処することがますます重要となる。

これら 2 種類のバイアスを区別し、具体的な戦略を用いてそれぞれのバイアスに対処することにより、政策立案者は知性的で効率的かつ公正な AI システムを構築することができる。AI に対する人々の信頼を維持し、社会の進歩に資するその潜在力を最大限に活用するためには、バランスの取れたアプローチが不可欠である。

ENDNOTES

1 T. Marwala (2023) Algorithm Bias — Synthetic Data Should Be Option of Last Resort When Training AI Systems. Daily Maverick.
https://unu.edu/article/algorithm-bias-synthetic-data-should-be-option-last-resort-when-training-ai-systems

2 T. Marwala (2024) The Dual Faces of Algorithmic Bias — Avoidable and Unavoidable Discrimination. Daily Maverick.
https://unu.edu/article/dual-faces-algorithmic-bias-avoidable-and-unavoidable-discrimination

3 Bigman, Y.E., Wilson, D., Arnestad, M.N., Waytz, A. and Gray, K., 2023. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), p.4.

4 Kordzadeh, N. and Ghasemaghaei, M., 2022. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems, 31*(3), pp.388-409.

5 Breidbach, C.F., 2024. Responsible algorithmic decision-making. *Organizational Dynamics*, p.101031.

6 Obermeyer, Z., Nissan, R., Stern, M., Eaneff, S., Bembeneck, E.J. and Mullainathan, S., 2021. Algorithmic bias playbook. *Center for Applied AI at Chicago Booth*.

7 Marwala, T., 2014. Missing Data Approaches for Rational Decision Making: Application to Antenatal Data. Artificial Intelligence Techniques for Rational Decision Making, pp.55-71.

8 Hastings, J., 2024. Preventing harm from non-conscious bias in medical generative AI. *The Lancet Digital Health, 6*(1), pp.e2-e3

9 von Eschenbach, W.J., 2021. Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology, 34*(4), pp.1607-1622.

10 Aysolmaz, B., Iren, D. and Dau, N., 2020. Preventing algorithmic Bias in the development of algorithmic decision-making systems: A Delphi study.

11 Acemoglu, D. and Jackson, M.O., 2015. History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies, 82*(2), pp.423-456.

12 Young, H.P., 2015. The evolution of social norms. *economics, 7*(1), pp.359-387.

13 Belenguer, L., 2022. AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics, 2*(4), pp.771-787.

14 Nijhawan, L.P., Janodia, M.D., Muddukrishna, B.S., Bhat, K.M., Bairy, K.L., Udupa, N. and Musmade, P.B., 2013. Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research, 4*(3), p.134.

15 Trenque, A., Rabiaza, A., Fedrizzi, S., Chretien, B., Sassier, M., Morello, R., Alexandre, J. and Humbert, X., 2024. Evaluation of a simplified pharmacovigilance tool for general practitioners: 5 years of insight. *Scientific Reports, 14*(1), p.1766.

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 4 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, *“Avoidable and Unavoidable Algorithmic Bias”*, UNU Technology Brief 4 (Tokyo: United Nations University, 2024).

ISBN 978-92-808-9153-9