United Nations University

TECHNOLOGY BRIEF
No. 1, SEPTEMBER 2023

AI モデルのトレーニングにおける合成データの利用：持続可能な開発に向けたチャンスとリスク

機械学習の工程で利用される合成データの広範な影響の理解

Tshilidzi Marwala, United Nations University, Tokyo, Japan
Eleonore Fournier-Tombs, UNU Centre for Policy Research, New York, USA
Serge Stinckwich, UNU Macau, Macau SAR, China

技術的行動への提言：

- 合成データセットの生成時に多様なデータソースを使用する
- 合成データセットの生成時に異なる種類の生成AIモデルを使用する
- すべての合成データとその出自を開示、またはウォーターマーク（識別用情報）で示す
- 合成データの品質指標を計算および開示する
- 合成データとそのソースを保護するためのサイバーセキュリティのルールを構築する
- 可能な限り非合成データを優先する

政策提言：

- グローバルAIガバナンスの取り組みに合成データを結び付ける
- 合成データをグローバルなデータガバナンスの不可欠かつ固有の問題として認識する
- グローバルな品質基準とセキュリティ対策を確立する
- 合成データの安全かつ倫理的な利用に関するグローバルな研究ネットワークを推進する
- 透明性の確保を含む倫理指針を明示する

はじめに

合成データ、すなわち人工的に生成されたデータを用いたAIアルゴリズムのトレーニングは、大きな潜在的可能性を秘めた急成長の手法である。この手法は、データの不足、プライバシー、バイアスといった問題の解決につながる可能性がある一方で、データの品質、セキュリティ、倫理的意味といった面で懸念を生じさせる可能性もある。この問題はグローバル・ノースに比べてデータ不足がはるかに深刻なグローバル・サウスにおいて、とくに顕著である。合成データは、欠測データの問題を解決し、うまくいけばデータセットにおける母集団の代表性を高め、より公平な結果をもたらす。しかしそうであるからといって、合成データを現実世界の実際のデータより優れている、または同等であるとみなすことはできない。実際、合成データの利用は、サイバーセキュリティリスク、バイアス伝播、およびモデル誤差の単純な増加など、多くのリスクをともなう。この技術ブリーフは、AIトレーニングにおける合成データの責任ある利用のための提言と、合成データの利用を規制するための関連指針を示すものである。

この技術ブリーフの目的は、AIを通じてグローバル・サウスにおけるSDGs達成を加速するために、合成データの重要なリスクを軽減しつつ、その潜在的可能性を模索することである。

政策の背景

多国間システムが人工知能のグローバルガバナンスへとかじを切り始めるなか、重要な課題となるのは拡大するデジタル・ディバイド（デジタル上の分断）への対処である。デジタル化が加速する現在、デジタル・ディバイドは、インターネット接続の問題のみならず、データセットの代表性と不均一性の問題としても表面化しつつある[1]。

データは、人工知能開発のすべての段階、とくにトレーニングとテストの段階においてきわめて重要である。母集団の中の1つのセグメントのみを代表するデータセットでトレーニングされたAIモデルは、代表されていないセグメントに対してリスクが大幅に大きくなる。実際に、近年明らかになった人工知能のリスクの多くは、国内および国家間のデータセットの均一性に起因している。

十分に実証された明白な事例が、ブオラムウィニとゲブルにより明らかにされた。彼らは、アメリカで使用されている顔認識システムがとても小さなデータセットでトレーニングされており、有色人種の女性についてはエラー率が圧倒的に高いことを発見した[2]。同様の理由で、AIシステムにおいて女性がステレオタイプ（固定観念の型）にはめられていたり差別されていたりする例は多く、とくに人事や金融分野においては顕著である。

しかし、AIデータの均一性の問題は、グローバル・サウスでさらに一際目立っている。現地語の使用から、服装、およびその他のさまざまな社会的・文化的・経済的要素に至るまで、その土地ならではの微細な特徴が、依然としてデータセットに十分に代表されていない。このことが、医療、教育、行政サービスにおけるAIシステムのパフォーマンス低下から、より長期的には平和やガバナンスに至るまで、持続可能な開発にますます影響を及ぼすようになりつつある。

この問題に対して提案された強力なソリューション（解決策）が、合成データの利用である。人工的に生成された合成データは、AIシステムにおける特定のグループに対する差別を減らすために、均一なデータセットのギャップを埋め、データの多様性を高めることができる。このような技術は、偏ったデータセットのバランスを取り戻し、グローバル・サウスにおけるAIシステムの妥当性を高めるための近道として、もてはやされている。事実、ガートナーは、2024年にはAIシステムで使用されるデータの60%が合成データになると主張している[3]。

合成データとは何か

合成データとは、実世界におけるデータの一部の構造的・統計的特性を複製するコンピューターシミュレーション、またはアルゴリズムによって生成された情報である。この「合成」過程により生成されたデータには、画像、動画、テキスト、表形式データがある。合成データは一般的に、グラウンドトゥルース（ドメイン知識、科学的理論、または収集データ）を基盤とする生成モデルによって生成され、これが新しい合成データのサンプルを生成する。

ここで重要なのは、大規模言語モデル（LLM）も合成データだという点である。したがって、ヴェンダ語などの低リソース言語における大規模言語モデルの精度は、英語などの高リソース言語よりも低くなる。

合成データは、とくに実世界のデータの扱いに慎重さが求められる場合や、データ数が限られている場合、または偏っている場合において、AIアルゴリズムのトレーニングにますます利用されるようになりつつある[4]。

合成データの生成にはさまざまな方法がある。合成データ生成の最も古い形態が、欠測値補完技術である[5,6,7]。1993年[8]にルービンが、マイクロデータの機密性を守りつつ、統計分析に基づいて合成データを生成するというアイデアを着想した。

人工知能の進化にともない、ニューラルネットワークに基づく深層生成モデル（DGM）が合成データ生成の最も好適な技術となった。インプットを分類するニューラルネットワークとは異なり、DGMは新たなアウトプットを生成することに特化している。DGMは既存の収集データに「類似した」新たなサンプルを生成することを学習する。このようなモデルは、自動走行車やロボット、患者の電子カルテ、または都市における人の移動パターンに利用できる高忠実度の画像や音楽、官能データを生成することができる。変分オートエンコーダ（VAE）、敵対的生成ネットワーク、およびより最近の大規模言語モデルなど、さまざまなDGMアプローチが利用可能である。これらのうちのどれを使用するかは、個々の用途やデータの特性によって決まる。

1 Fournier-Tombs, E. (2023). Local transplantation, adaptation and creation of AI models for public health policy. *Frontiers in Artificial Intelligence*, *6*, 1085671.

2 Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

3 White, A. (2021). By 2024, 60% of the data used for the development of AI and analytics projects will be synthetically generated. Gartner Blog. Accessed at: https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/

4 Marwala, T. (2023). *Artificial Intelligence, Game Theory and Mechanism Design in Politics*. Springer Nature.

5 Marwala, T. (Ed.). (2009). *Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques: Knowledge Optimization Techniques*. IGI Global.

6 Leke, C. A., & Marwala, T. (2019). *Deep learning and missing data in engineering systems* (p. 179). Berlin, Germany: Springer International Publishing.

7 Mbuvha, R., Adounkpe, J. Y., Houngnibo, M. C., Mongwe, W. T., Nikraftar, Z., Marwala, T., & Newlands, N. K. (2023). A novel workflow for streamflow prediction in the presence of missing gauge observations. *Environmental Data Science, 2*, e23.

8 Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics, 9* (2), 461-468.

なぜ合成データを利用するのか

合成データは、偏ったデータセットのバランス調整、データプライバシーの保護、およびデータ収集コストの削減など、数多くのメリットを提供する。下の表は、これらのメリットをまとめたものである。

合成データの用途	詳細
データ可用性	合成データは、AI システムのトレーニング用のデータセットを補完することにより、データの可用性と代表性に関する懸念に対処することができる。
プライバシー保護	合成データは実在する人を代表してはいないため、侵害が発生した場合に人に損害をもたらしうる個人情報が含まれていない。
バイアス低減	合成データは、ジェンダーバイアスや民族バイアスのような、AI バイアスにつながる不均衡なトレーニング・データセットの問題に対処することができる。
コンプライアンス	合成データは、医療分野など、実世界のデータが限られている場合に、AI モデルのトレーニングに利用することができる。
コスト	計算コストと環境コストは依然として重要ではあるものの、実世界のデータを収集する代わりに合成データを用いることで費用便益が得られる可能性がある。

データ可用性：合成データは、データ不足にともなう限界を克服し、より頑健な AI のトレーニングと開発を可能にする。大規模言語モデルを使用して生成された合成データは、病気の診断や新たな治療法の開発（SDGs の目標 3「すべての人に健康と福祉を」）といったタスクのための AI モデルのトレーニングに利用されてきた。金融サービス産業（SDGs の目標 8「働きがいも経済成長も」）では、景気予測、不正検出、リスク評価などのタスクのための AI モデルのトレーニングに合成データが利用されてきた [9]。気候科学産業（SDGs の目標 13「気候変動に具体的な対策を」）では、気象予報や気候モデリングなどの用途のための AI モデルのトレーニングに合成データが利用されてきた。これは、気候変動に対する新たな緩和・適応戦略を策定するうえで不可欠である。

プライバシー保護：合成データには個人情報（PII）が含まれていないため、データ保護規則の遵守やユーザーのプライバシー保護に対する有益な手段となる [10]。医療産業では、合成データを生成する際、実際の医療データを使用する前に個人情報を削除または匿名化することにより、患者のプライバシーを保護しながら、現実に近いデータを用いて AI モデルのトレーニングを行っている（SDGs の目標 3）[11]。

バイアス低減：合成データは均衡のとれた代表的な内容となるよう設計することができるため、AI モデルのバイアス低減に寄与することができる。例えば、人工知能モデルにおけるジェンダー差別を最小化するために合成データが利用されており、SDGs の目標 5（「ジェンダー平等を実現しよう」）に貢献している [12]。

下の表は、合成データセットを用いてトレーニング・データセットを強化し、特定のサブグループ（女性）に対する AI モデルの精度を高める方法を示したものである。ただし、合成データセットの品質が実世界のデータセットの品質と完全に一致することはないため、各サブグループのエラー率は同じではないという点に留意する必要がある。表示されたデータは例示のみを目的としている。

	男性	女性	男性のモデル精度	女性のモデル精度
合成データなし	20	12	97%	65%
合成データあり	20（リアルデータ）	12（リアルデータ）＋8（合成データ）	97%	85%

コンプライアンス：実世界のデータの使用を制限する規則を遵守するために、合成データを利用することができる。例えば、合成データを利用することによって、扱いに慎重さが求められるデータにアクセスすることなく、機械学習モデルのトレーニングを行うことができる。医学生の訓練に使用される慎重に扱う必要のある医療用画像を生成するために合成データが利用されており、SDGs の目標 4（「質の高い教育をみんなに」）に貢献している [13]。

コスト：合成データの生成は、実世界のデータの収集よりも費用効率が高い可能性がある。臨床試験や市場調査など、費用のかさむデータ収集を要する場面において、これは大きな意味を持つ。しかし、合成データの生成にはかなりの計算コストや環境コストがかかる場合もある。したがって、SDGs の目標 13（「気候変動に具体的な対策を」）の進展に向けてデータを合成するための生成モデルを選択する際には、この点に留意することが不可欠である。

合成データの主なリスク

ただそれでもなお、合成データの利用には、データの品質、サイバーセキュリティ、不正使用、バイアス伝播、知的財産

9 Sidogi, T., Mongwe, W. T., Mbuvha, R., & 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). IEEE.

10 Savage, N. (2023). Synthetic data could be better than real data. *Nature*. https://www.nature.com/articles/d41586-023-01445-8

11 Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493, 28-45.

12 Sharma, S., Zhang, Y., Ríos Aliaga, J. M., Bouneffouf, D., Muthusamy, V., & Varshney, K. R. (2020, February). Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 358-364).

13 Costello, J. P., Olivieri, L. J., Krieger, A., Thabit, O., Marshall, M. B., Yoo, S. J., Kim P.C., Jonas R. A., & Nath, D. S. (2014). Utilizing three-dimensional printing technology to assess the feasibility of high-fidelity synthetic ventricular septal defect models for simulation in medical education. *World Journal for Pediatric and Congenital Heart Surgery, 5* (3), 421-426.

権の侵害、データ汚染およびデータコンタミネーションなど、さまざまなリスクがともなう。

データの品質：合成データの品質と現実性は、効果的な AI トレーニングにとってきわめて重要である。不適切な方法で生成された合成データは、不正確で信頼できない AI モデルを生む可能性がある。採用された手法によって、合成データのインテグリティ（完全性・正確性）は左右される。一般的に、敵対的生成ネットワーク（GAN）によって生成されたデータは現実性が高い一方、データ分布の管理が困難となる可能性がある。統計モデルはより均等に分布したデータを生成することができるが、生成されたデータの妥当性は低くなる可能性がある。いくつかの変数が合成データのインテグリティに影響を及ぼす。第一に、合成データを生成するために用いられる手法がデータの品質に大きな影響を及ぼす可能性がある。第二に、モデルのトレーニングに使用されるデータが多いほど、合成データの品質は高まる。なぜなら、より多くのデータから学習したモデルは、より現実的な合成データを生成することができるからである。第三に、モデルのトレーニングに使用される実世界のデータの品質が高いほど、そのモデルは実世界のデータから学習して類似した合成データを生成することができるため、合成データの品質も高くなる。

セキュリティリスク：合成データがリバースエンジニアリング（逆行分析）されると、基礎となる実世界のデータやデータ生成に用いられたプロセスについての情報が暴露される可能性があり、セキュリティリスクが生じる。したがって、とくに使用されたソースデータが合成データとともに公表されている場合や、合成データを生成するために用いられたモデルがトレーニング・データに「オーバーフィット」、つまり、オリジナルのデータセットに似すぎている場合には、合成データの再識別は現実的なリスクとなる。

不正使用：AI トレーニングにおける合成データの利用は、ディープフェイク作成のための不正使用の可能性やその他の欺瞞的な AI テクノロジーなど、倫理的な問題を提起する。また、合成データには知的財産リスクがともなうことも明らかになりつつあり、とくに、芸術的なソース資料や、人間が知的所有権を有するその他のソースから画像を生成する場合には、注意を要する。

バイアス伝播、データ汚染またはデータコンタミネーション：合成データが不均衡であったり、母集団を正しく代表していなかったり、その他の形で偏ったりしている場合、そのバイアスはトレーニングされたモデル全体や、さらには他の合成データセットにまで伝播する可能性がある。合成データはデータセットから生成されるため、そのデータセットが小さければ、2021 年にセガール他が X 線画像の事例で明らかにしたように、その小ささは合成データセットにも投影される [14]。合成データの利用が一般的になり、より安価になるにつれ、インターネット上で入手できるより多くのデータが深層生成モデルによって生成されるようになり、これらのインプットがある種のネガティブフィードバックによって AI システムのトレーニングに再び使用され、その結果として深層生成モデルが崩壊する可能性がある [15]。こうして最終的に、合成データと実世界のデータを区別することがますます困難になる。

技術的提言

以下に、合成データの利用における技術的提言を示す。

1. **合成データセットの生成時には多様なデータソースを使用する：**合成データを生成する時には、さまざまなデータソースを使用して、実行可能な限り多様で多くの独立した特性を持つデータとなるようにすることが極めて重要である。これには、収集された実世界のデータと、シミュレーションや、専門家の知識、市民から提供された参加型データといった他のソースから得られたデータとの両方の使用が含まれる [16]。

2. **合成データセットの生成に異なる種類の生成 AI モデルを使用する：**目の前のタスクにふさわしい、現実的で実際の世界を代表したデータを生成するモデルを選択することが不可欠である。

3. **すべての合成データとその出自を開示する、またはウォーターマークで示す：**すべての合成データがどこから来て、どのように生成されたのかを開示し [17]、あらゆる知的財産保護規定を遵守することが不可欠である。

4. **合成データの品質指標を計算および開示する：**合成データが生成されたら、それが意図された適用場面にふさわしいものとなるように、その品質を評価しなければならない [18]。これは、正確性、網羅性、多様性についての情報の検証をともなう可能性がある。

5. **可能な限り、非合成データを優先する：**合成データは強力なリソースとなりうるが、責任ある方法で利用されなければならない。そのためには、合成データの限界を認識し、合成データのユーザーを誤解させたり、だましたりすることのないよう利用しなければならない。

14 Segal, B., Rubin, D. M., Rubin, G., & Pantanowitz, A. (2021). Evaluating the Clinical Realism of Synthetic Chest X-Rays Generated Using Progressively Growing GANs. *SN Computer Science, 2* (4), 321.

15 Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The Curse of Recursion: Training on Generated Data Makes Models Forget. *arXiv preprint arxiv:2305.17493.*

16 Tan, Y. R., Agrawal, A., Matsoso, M. P., Katz, R., Davis, S. L., Winkler, A. S., Huber A., Joshi A., El-Mohandes A., Mellado B., Mubaira C. A., Canlas F. C., Asiki G., Khosa H., Lazarus J. V., Choisy M., Recamonde-Mendoza M., Keiser O., Okwen P., English R., Stinckwich S., Kiwuwa-Muyingo S., Kutadza T., Sethi T., Mathaha T., Nguyen V.K, Gill A. & Yap, P. (2022). A call for citizen science in pandemic preparedness and response: beyond data collection. *BMJ Global Health, 7* (6), e009389.

17 Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64* (12), 86-92.

18 https://www.vanderschaar-lab.com/generating-and-evaluating-synthetic-data-a-two-sided-research-agenda/

政策提言

1. **グローバル AI ガバナンスの取り組みに合成データを結び付ける：**国連の人工知能に関するマルチステークホルダー諮問機関の勧告によるものなど、現在進められているグローバル AI ガバナンスの取り組みにおいて、合成データに関する作業部会を設立し、提示されたリスクがグローバルかつ包括的に対処されるようにすることが不可欠である。

2. **合成データをグローバルなデータガバナンスの不可欠かつ固有の問題として認識する：**国際社会が今後数年間における合成データの利用増加に備えられるよう、合成データに関する問題をグローバルなデータガバナンスの並行政策軌道に含める。

3. **グローバルな品質基準とセキュリティ対策を確立する：**合成データの生成と AI トレーニングへのその利用に関する基準を策定・実施し、合成データの品質と信頼性を確保する。合成データのリバースエンジニアリングを防止し、AI トレーニングプロセスのインテグリティを保護するための頑健なセキュリティ対策を実施する。

4. **合成データの安全かつ倫理的な利用に関するグローバルな研究ネットワークを推進する：**合成データに関する学術・政策研究に十分な資金が提供され、グローバル・サウスとグローバル・ノースの研究者が効果的に協力し、ソリューションを共有できるようにする。

5. **透明性を含む倫理指針を明示する：**AI トレーニングにおける合成データの利用について明確な倫理指針を示し、同意、透明性、不正使用の可能性などの問題に対処する。組織は、AI トレーニングにおける合成データの利用について透明性を確保しなければならない。これにより、信頼を構築し、アカウンタビリティ（説明責任）を確保し、責任あるイノベーションを促進することができる。

結論

昨年は、世界的にすべてのセクターやスキルレベルで生成 AI の使用が著しく増加した。これらのテクノロジーによって、合成データセットの生成がより一層身近なものとなった。しかし、本稿で述べた通り、合成データセットの不適切な生成と人工知能システムへの利用は、とくに AI バイアスの伝播によって、持続可能な開発に多大な悪影響を及ぼす可能性がある。その一方で、この種のデータが正しく利用されることにより、医学研究の強化や人工知能モデルの差別的アウトプットの削減などのメリットが得られるケースも数多くある。したがって本稿では、ソフトウェア開発者の採用すべき技術的基準と、政策立案者に向けた提言の両方を示すことにより、合成データ利用の標準化に向けた第一歩を提案する。こうした取り組みは、国連およびその他の多国間機関で現在行われている人工知能のグローバルガバナンスに関する対話に反映されることをとくに目的としている。

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 1 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。

エレノア・フルニエ＝トムズ博士は、国連大学政策研究センター（UNU-CPR）の予期的人道行動とイノベーションの部門長で、国連における AI とデータに関する方法論的ツールの開発と政策提言に注力している。同博士はオタワ大学法学部で “Accountable AI and a Global Context” の非常勤教授を務めるほか、マギル大学とモントリオール大学で新技術とサイバーセキュリティについて定期的に講師を務めている。

サージ・スタンクウィッチ博士はコンピュータ科学者であり、国連大学在マカオ研究所（UNU Macau）の研究部長を務めている。人間を中心とした視点から持続可能な開発のためにデジタル技術がもたらし得るプラスの貢献を拡大し、リスクを軽減する方法を模索している。博士の主な研究テーマは、複雑系のモデリング、社会シミュレーション、持続可能な開発目標（SDGs）に対する人工知能の影響などである。

免責事項

引用の際の表記
Tshilidzi Marwala, Eleonore Fournier-Tombs, Serge Stinckwich, “The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development”, UNU Policy Brief (Tokyo: United Nations University, 2023).

ISBN 978-92-808-9146-1