United Nations University

TECHNOLOGYBRIEF
No. 1, SEPTEMBER 2023

# 使用合成数据训练人工智能模型：可持续发展的机遇与风险

## 了解机器学习流水线中使用的合成数据的广泛影响

Tshilidzi Marwala, 联合国大学，日本东京
Eleonore Fournier-Tombs, 联合国大学政策研究中心，美国纽约
Serge Stinckwich, 联合国大学澳门研究所，中国澳门特别行政区

**建议采取的技术行动：**

- 在创建合成数据集时使用不同的数据源
- 使用不同类型的生成式人工智能模型创建合成数据集
- 公开或标记所有合成数据及其出处
- 计算并披露合成数据的质量指标
- 制定网络安全协议，保护合成数据及其来源
- 尽可能优先考虑非合成数据

**建议采取的政策行动：**

- 将合成数据与全球人工智能治理工作联系起来
- 认识到合成数据是全球数据治理中的一个关键和独特问题
- 制定全球质量标准和安全措施
- 促进有关安全和合乎道德地使用合成数据的全球研究网络
- 明确包括透明度的道德准则

**简介**

使用合成或人工生成的数据来训练人工智能算法是一种新兴的做法，具有巨大的潜力。它可以解决数据稀缺、隐秘和存在偏见等问题，同时也引起人们对数据质量、安全性和道德影响的关注。这一问题在全球南方更为严重，因为那里的数据稀缺程度比全球北方严重得多。因此，合成数据可以解决数据缺失的问题，在最好的情况下，合成数据可以更好地代表数据集中的样本，并带来更公平的结果。但是，我们不能认为合成数据比现实世界的实际数据更好，更不能将其等同。事实上，使用合成数据有很多风险，包括网络安全风险、偏见的宣传，以及单纯地增加模型误差。本政策简报提出了在人工智能训练中负责任地使用合成数据的建议，以及规范合成数据使用的相关准则。

本政策简报旨在探索合成数据的潜力，以通过全球南部地区的人工智能加速实现可持续发展目标，同时降低其主要风险。

## 政策背景

随着多边体系开始转向人工智能的全球治理，一个重要的挑战将是解决日益扩大的数字鸿沟。如今，在数字化加速发展的时代，数字鸿沟不仅表现在互联网连接方面，而且越来越多地表现在数据集的代表性和异质性方面。[1]

数据在人工智能开发的各个阶段都至关重要，尤其是在训练和测试阶段。在仅代表部分样本的数据集上训练的人工智能模型，对未代表的样本而言，风险率要高得多。事实上，近年来人工智能的许多风险都是由于国家内部和国家之间数据集的同质性造成的。

Buolamwini 和 Gebru 发现了一个有据可查的明显例子，他们发现美国使用的人脸识别系统是在一个非常狭窄的数据集上训练出来的，对有色人种女性的错误率极高 .[2] 出于类似原因，在人工智能系统中，尤其是在人力资源和金融应用领域，妇女被定型或歧视的例子也不胜枚举。

但在全球南部，人工智能数据的同质性问题更加突出。从当地语言的使用到服装，再到许多其他社会、文化和经济方面，数据集对当地的细微差别仍然没有很好的体现。这日益导致对可持续发展的影响，从人工智能系统在卫生、教育和政府服务方面的较差表现，到对和平与治理的长期影响。

一个有力的解决方案是使用合成数据。人工创建的合成数据可以填补同质数据集的空白，增加数据的多样性，从而减少人工智能系统对特定群体的歧视。这类技术被誉为重新平衡有偏见数据集的快速方法，可提高人工智能系统在全球南部地区的适用性。事实上，Gartner 认为，最快到 2024 年，人工智能系统使用的数据将有 60% 是合成的。[3]

## 什么是合成数据?

合成数据（SD）是由计算机模拟或算法创建的信息，可再现真实世界数据的某些结构和统计属性。这种“合成“过程产生的数据可以是图像、视频、文本或表格数据。合成数据一般由一个生成模型产生，该模型基于基本事实（领域知识、科学理论或收集的数据），将产生新的合成数据样本。

值得注意的是，大型语言模型（LLMs）的生成也是合成数据。因此，低资源语言（如齐文达语）的大型语言模型准确率低于高资源语言（如英语）。

合成数据越来越多地被用于训练人工智能算法，尤其是在真实数据敏感、稀缺或存在偏见的情况下。[4]

创建合成数据的方法多种多样。最早的合成数据生成依据缺失数据估算技术。[5, 6, 7] 1993 年 ,[8] Rubin 提出了在统计分析的基础上生成合成数据，同时又能保证微观数据的保密性的想法。随着人工智能的发展，基于神经网络的深度生成模型（DGM）已成为生成合成数据的首选技术。与对输入进行分类的神经网络不同，深度生成模型专门用于生成新的输出。它们学会生成与现有收集数据“相似“的新示例。这类模型可以生成高保真图像、音乐、自动驾驶汽车或机器人的感官数据、病人电子健康记录或城市中的人类流动模式。可以使用各种 DGMs 方法，如变异自动编码器 (VAE)、生成对抗网络和最近的大型语言模型 (LLM)。它们之间的选择取决于具体的使用案例和数据特征。

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.

## 为什么使用合成数据?

合成数据提供了许多机会，例如重新平衡有偏差的数据集、保护数据隐私和降低数据收集成本。下表总结了这些机会。

| 合成数据的使用 | 说明 |
|---|---|
| 数据可用性 | 合成数据可为人工智能系统 " 补充 " 训练数据集，从而解决数据可用性和代表性问题。 |
| 隐私保护 | 合成数据不代表真实的人，因此不包含任何可识别个人身份的信息，从而避免在信息泄露时对其造成伤害。 |
| 减少偏见 | 合成数据可以解决使用有偏见的训练数据集导致人工智能偏见的问题，如性别偏见或种族偏见。 |
| 合规性 | 当真实数据受到限制时，合成数据可用于训练人工智能模型，例如在医疗领域。 |
| 成本 | 使用合成数据而不是真实数据收集可以带来成本优势，尽管计算和环境成本仍然很重要。 |

数据可用性：合成数据可以克服与数据稀缺相关的限制，实现更稳健的人工智能训练和开发。使用大型语言模型生成的合成数据已被用于训练人工智能模型，以完成疾病诊断和开发新治疗方法等任务（可持续发展目标 3）。在金融服务业（可持续发展目标 8），合成数据已被用于训练人工智能模型，以完成经济预测、欺诈检测和风险评估等任务。[9] 在气候科学行业（可持续发展目标 13），合成数据被用来训练人工智能模型，用于天气预报和气候建模等，这对于制定新的气候变化减缓和适应战略至关重要。

隐私保护：合成数据不包含个人身份信息（PII），因此是遵守数据保护法规和保护用户隐私的重要工具 [10] 在医疗保健行业，在使用真实世界的医疗保健数据生成合成数据之前，PII 会被移除或去标识，从而允许人工智能模型在现实数据上进行训练，同时保护患者的隐私（可持续发展目标 3）。[11]

减少偏差：合成数据可以设计得平衡且具有代表性，有助于减少人工智能模型中的偏差。例如，合成数据已被用于确保在人工智能模型中最大限度地减少性别歧视，从而推进可持续发展目标 5 的实现。[12]

在下表中，我们展示了在理想情况下如何使用合成数据集来强化训练数据集，从而提高人工智能模型对特定子群（女性）的准确性。但需要注意的是，合成数据集的质量可能无法与真实数据集完全匹配，因此每个子群的错误率也不尽相同。所提供的数据仅供参考。

| | 男性 | 女性 | 男性准确率模型 | 女性准确率模型 |
|---|---|---|---|---|
| 无合成数据 | 20 | 12 | 97% | 65% |
| 添加合成数据 | 20（真实） | 12（真实）+ 8（合成） | 97% | 85% |

合规性：合成数据可用于遵守限制使用真实世界数据的法规。例如，合成数据可用于训练机器学习模型，而无需访问敏感数据。合成数据已被用于生成敏感的医学图像，用于培训医科学生，从而推进可持续发展目标 4 的实现。[13]

成本：生成合成数据比收集真实世界的数据更具成本效益。这对于临床试验和市场研究等需要高成本数据收集的应用来说意义重大。不过，在某些情况下，生成合成数据的计算成本和环境成本也很高。因此，在选择生成模型来合成数据以推进可持续发展目标 13 时，必须考虑到这一点。

## 合成数据的主要风险

然而，使用合成数据也存在许多风险，如数据质量、网络安全、滥用、偏差传播、知识产权侵权、数据污染和数据感染等。

**数据质量：**合成数据的质量和真实性对于有效的人工智能训练至关重要。生成的合成数据不佳会导致人工智能模型不准确、不可靠。根据所采用的方法，合成数据的完整性可能会有所不同。通常情况下，生成式对抗网络（GAN）生成的数据非常逼真，但要控制其分布却很困难。统计模型可以生成分布更均匀的数据，但生成的数据可能不太可信。一些变量会影响合成数据的完整性。首先，生成合成数据的方法会严重影响其质量。其次，当使用更多数据来训练模型时，合成数据的质量会得到提高。这是因为模型将有更多的数据来学习并生成更真实的合成数据。第三，如果用于训练模型的真实世界数据质量很高，那么合成数据的质量也会很高，因为模型可以从真实世界数据中学习并生成类似的合成数据。

**安全风险：**如果对合成数据进行逆向工程，就有可能泄露底层真实数据或生成过程的信息，从而带来安全风险。因此，重新识别是合成数据的真正风险，尤其是在使用的源数据与合成数据一起发布，或用于创建合成数据的模型“过度拟合”训练数据（即与原始数据集过于相似）的情况下。

**滥用：**在人工智能训练中使用合成数据会引发伦理问题，例如可能会被滥用于创建深度伪造数据或其他欺骗性人工智能技术。越来越多的人发现，合成数据也存在知识产权风险，

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.

尤其是在从艺术素材或人类拥有知识产权的其他来源生成图像时。

**偏差传播、数据污染或数据感染：**如果合成数据不平衡、错误地反映了某个人群或存在其他偏差，其偏差就会传播到整个训练过的模型中，甚至传播到其他合成数据集。合成数据是由数据集生成的，如果该数据集很窄，那么这种窄性就会投射到合成数据集中，正如西格尔等人在 2021 年以 X 射线图像为例所指出的那样。[14] 随着合成数据使用的广泛化和成本的降低，互联网上越来越多的数据将由深度生成模型生成，这些输入将再次用于训练人工智能系统，形成一种负反馈，这些模型可能会崩溃。[15] 到最后，要区分哪些是合成数据，哪些是真实数据将变得越来越困难。

## 技术建议

利用合成数据的技术建议：

1. 创建合成数据集时使用多种数据源：在生成合成数据时，关键是要利用各种数据源，以确保数据尽可能多样化并具有许多独立特征。这可能涉及使用收集的真实世界数据和其他来源的数据，如模拟、专家知识或公民参与数据。[16]
2. 使用不同类型的生成式人工智能模型创建合成数据集：必须选择适合当前任务的模型，并能生成真实且能代表现实世界的数据。
3. 公开或标记所有合成数据及其出处：必须披露所有合成数据的来源和生成方式，并遵守任何相关知识产权保护规定。[17]
4. 计算并披露合成数据的质量指标：合成数据生成后，必须对其质量进行评估，以确保其适合预期应用。这可能包括检查信息的准确性、完整性和多样性。[18]
5. 尽可能优先使用非合成数据：合成数据是一种强大的资源，但必须负责任地使用。这就要求了解合成数据的局限性，不误导或欺骗使用合成数据的用户。

## 政策建议

1. 将合成数据与全球人工智能治理工作联系起来：当前的全球人工智能治理工作，包括联合国人工智能多方利益相关者咨询机构所建议的工作均指出，必须建立一个合成数据工作组，以确保在全球范围内全面应对所概述的风险。
2. 认识到合成数据是全球数据治理中的一个关键和独特问题：将与合成数据相关的问题纳入全球数据治理的平行政策轨道，以便国际社会为未来几年合成数据使用的增加做好更好的准备。
3. 建立全球质量标准和安全措施：制定并实施在人工智能培训中生成和使用合成数据的标准，以确保其质量和可靠性。实施强有力的安全措施，防止合成数据的逆向工程，保护人工智能训练过程的完整性。
4. 促进有关安全、合乎伦理地使用合成数据的全球研究网络：确保有关合成数据的学术和政策研究获得充足的资金，确保全球南方和北方的研究人员能够有效合作并交流解决方案。
5. 明确道德准则，包括透明度：为在人工智能训练中使用合成数据提供明确的道德准则，解决同意、透明度和潜在滥用等问题。各组织在人工智能训练中使用合成数据时应保持透明。这可以建立信任、确保问责并促进负责任的创新。

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/

## 结论

去年，在全球范围内，各行各业和不同技能水平的人都在大量使用人工智能技术。这些技术使得合成数据集的创建比以前更加容易。然而，正如本简报所述，在人工智能系统中不适当地创建和使用合成数据集可能会对可持续发展产生重大不利影响，尤其是传播人工智能偏见。另一方面，在许多情况下，合理使用这类数据也能带来益处，如加强医学研究和减少人工智能模型的歧视性输出。因此，本简报通过概述软件开发者应采用的技术标准和对政策制定者的建议，在规范合成数据使用的道路上迈出了第一步。这些努力尤其旨在为目前在联合国和其他多边场合进行的人工智能全球治理对话提供参考。

## 编辑信息

**关于研究**
本技术简报是联合国大学系列简报的第一份，重点介绍全球技术治理中与全球南部和可持续发展相关的具体领域。

**作者简介**
Tshilidzi Marwala 教授是总部位于东京的联合国大学校长和联合国副秘书长。他曾任约翰内斯堡大学副校长兼校长。Marwala 在同行评审期刊和会议上发表了 300 多篇论文，出版了 27 本关于人工智能和相关主题的书籍，并拥有五项专利。他是美国艺术与科学院、世界科学院（TWAS）和非洲科学院的成员。

Eleonore Fournier-Tombs 博士是联合国大学政策研究中心预测行动与创新部主任，主要负责开发与联合国人工智能和数据有关的方法工具和政策建议。她还是渥太华大学法学院“负责任的人工智能与全球背景”的兼职教授，并经常为麦吉尔大学和蒙特利尔大学讲授新技术和网络安全。

Serge Stinckwich 博士是一名计算机科学家，也是澳门联合国大学研究所的研究负责人。澳门联合国大学研究所是联合国的一个智囊团，从以人为本的角度研究如何扩大数字技术对可持续发展的积极贡献并降低其风险。他的主要研究兴趣是复杂系统建模、社会模拟以及人工智能对可持续发展目标（SDGs）的影响。

**免责声明**
本文所表达的观点和意见并不一定反映联合国大学的官方政策或立场。

**译者**
吴博晋、潘天一，北京外国语大学国际组织学院

**引用格式**
Tshilidzi Marwala, Eleonore Fournier-Tombs, Serge Stinckwich, “The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development”, UNU Technology Brief 1 (Tokyo: United Nations University, 2023).

Copyright © 2023 United Nations University. All rights reserved.

ISBN 978-92-808-9150-8