Synthetic Data:
The New Data Frontier

WORLD ECONOMIC FORUM

BRIEFING PAPER

SEPTEMBER 2025

Images: Getty Images

Contents

Introduction

Data-driven decision-making shapes every aspect of our lives. Yet access to high-quality, representative data remains a persistent challenge. Real-world datasets are often incomplete, biased, restricted by privacy concerns, or simply unavailable – hindering innovation and reinforcing systemic inequalities. Synthetic data offers a solution.

Synthetic data is data artificially generated to mimic the statistical properties, structure and distribution of real-world data. It can fill data gaps, protect privacy and enable the testing of new scenarios, providing a scalable and cost-effective alternative when real-world data is limited or sensitive.

However, synthetic data introduces new governance and ethical risks. If not carefully generated and managed, it can perpetuate biases in the original datasets, mislead decision-makers, leak sensitive information, or be weaponized for malicious purposes (such as through the creation of deepfakes). Ensuring the accuracy, traceability and clear labelling of synthetic data is essential to mitigate risks, preserve model performance and maintain public trust.

Recognizing synthetic data's transformative potential, the World Economic Forum's **Global Future Council on Data Frontiers**[1] has developed this executive primer to explain its main types, use cases and governance considerations. This strategic brief seeks to empower leaders across public, private, academic and civil society sectors to harness synthetic data for innovation – while upholding standards of accuracy, equity and privacy.

1 Synthetic data: Core definitions

Broadly speaking, data[2] can be defined as a collection of facts, numbers, words or observations,[3] either structured or unstructured, and generated through a variety of interactions or processes (e.g. commercial transactions, creative or industrial processes, human behaviour or environmental measurements). It may be collected directly via human interaction or indirectly from devices, systems or networks, and is stored through technological means. Data processing and data analysis then allow for the transformation of raw datasets into insights for decision-making. But not all data is created equal, making it essential for stakeholders to understand the origin, purpose, authenticity and reliability of the data that feeds their decision-making and artificial intelligence (AI) systems.

Organic data refers to data that is generated naturally and collected directly from real sources, through authentic and unaltered interactions, behaviours or phenomena, such as user actions on a website, sensor readings of an environment or transactions within a system.[4] While organic data may be filtered, aggregated or organized for easier analysis, its core features and statistical properties remain faithful to real-world events.

Synthetic data refers to data that is generated by artificial means, such as statistical algorithmic methods or by AI. Synthetic data is created rather than collected from real-world sources, with the aim of addressing varied challenges of data unavailability, scarcity, privacy or representativeness. For specific use cases, synthetic data reproduces the key statistical characteristics, structure or distribution of organic data and is generated through methods like statistical modelling, machine learning (ML) algorithms, simulations or hybrid approaches.[5] While initially developed to enhance privacy,[6] it has many other uses as a substitute or complement when real-world data is unavailable, impractical or suboptimal.[7]

Like data in general, synthetic data can take many forms, each defined by its underlying generation method and the challenge it seeks to solve:

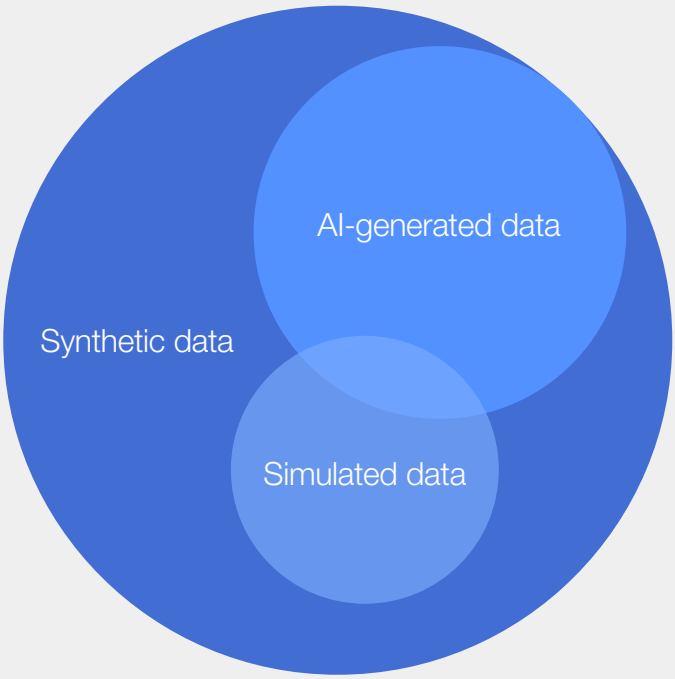
- **AI-generated data** is a type of synthetic data produced by AI models, including ML methods and generative models (e.g. generative adversarial networks or GANs; large language models or LLMs).[8] It is often created to replicate real-world data for specific tasks, such as image generation, text synthesis or content creation, but can also be produced purely for the enjoyment of a user. It can support data augmentation, model training and the creation of datasets for AI systems.[9] AI-generated data is distinct because it comes from AI models, which allow for more creativity and reasoning in their outputs, but introduces unique risks like lack of transparency or hallucinations.

- **Simulated data** is a type of synthetic data that is generated through traditional modelling, simulation techniques and AI methods. It aims to accurately represent the characteristics and behaviour of real-world phenomena (e.g. physical systems, economic models or digital twins). While the line between simulated data and other forms of synthetic data is increasingly blurring, a distinctive feature of simulated data is its focus on representing real-world dynamics and behaviours, often for testing, scenario analysis or risk assessment.

Hybrid datasets, meanwhile, combine organic and synthetic data, often with the goal of creating more robust datasets, for example to fill gaps where real data is missing or under-represented in clinical trial data.[10] Conversely, they may contain synthetic data refined through human-in-the-loop editing to enhance its accuracy or align it more closely with the real world.

Although the origin of synthetic data use was in research and development to address deficiencies of real-world data, its applicability and utility are far broader, as described in the following section.

FIGURE 1 **Main types of synthetic data**



Source: World Economic Forum

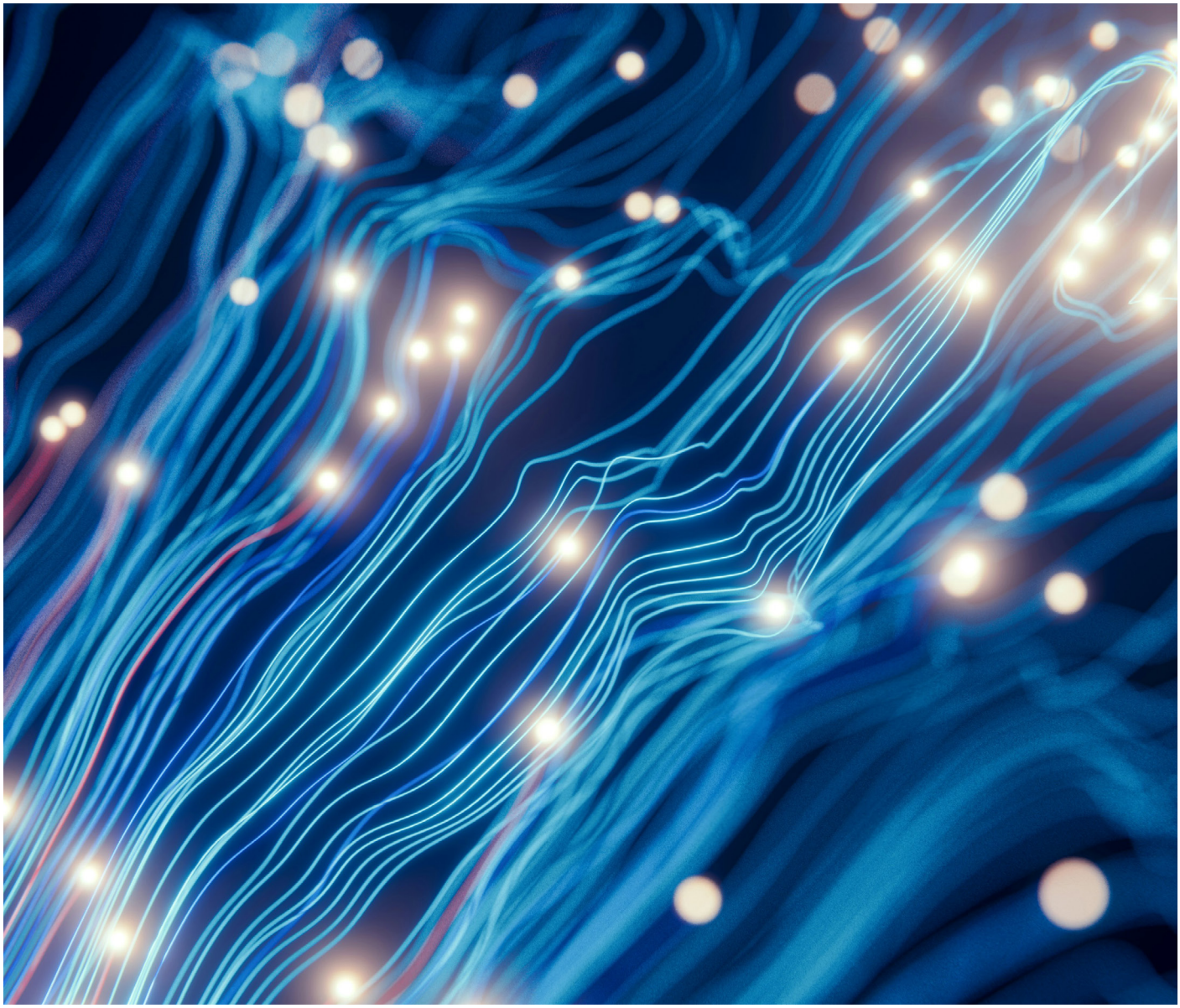
2 Taxonomy and use cases

This primer presents a taxonomy that categorizes the main uses of synthetic data by intended purpose, based on use cases. While the categories and examples below are not exhaustive or mutually exclusive, they seek to guide organizations and decision-makers to better understand how to leverage synthetic data in their sector.

TABLE 1 **Synthetic data uses**

Purpose: *Why might synthetic data be used?*	Example use cases
To enhance privacy To safeguard privacy and confidentiality by generating statistically equivalent datasets that maintain analytical utility while eliminating the risk of exposing sensitive personal information.	**Official statistics:** When confidentiality requirements prevent statistical offices from sharing granular microdata from official statistics, the use of synthetic data enables access to detailed data while maintaining privacy.[11]
To bridge data deficiencies To fill critical data gaps caused by scarcity or bias in organic data.	**Healthcare and demographics:** Organic datasets often exclude vulnerable populations – under-representing low-resource languages,[12] rare genetic conditions and entire demographic groups in clinical research. Synthetic data can improve model fairness and health equity by offering representative examples[13] and simulating under-represented groups in clinical trials.[14] **Child behaviour modelling:** Synthetic data can replicate children's behaviour patterns for safer research and development without the direct involvement of minors.[15] **Financial inclusion:** Synthetic data can correct lending practices driven by gender-biased data to provide fairer access to financial services for women in emerging markets.[16] **Criminal justice**: Synthetic data can ensure diverse demographic representation that reduces racial bias in predictive policing models.
To improve model performance To provide realistic, diverse inputs during model training for accuracy improvements.	**Diversification of training data:** AI model developers are often looking to diversify training datasets to deliver more accurate or representative outputs. For example, ByteDance reportedly uses synthetic data to augment training datasets for its LLMs.[17] **Add noise to training data:** For some use cases, adding synthetic noise during training can improve model robustness,[18] prevent overfitting[19] or help assess the impact of existing noise on model performance.[20]
For scenario modelling and forecasting To generate realistic future scenarios or to create precise virtual replicas of physical systems to simulate, analyse and optimize real-world performance (digital twins).	**Crisis modelling:** To simulate catastrophic events – prolonged blackouts, supply chain collapses, pandemic outbreaks – policy-makers can stress-test responses and build resilience across climate adaptation, urban planning, epidemiology and critical infrastructure. **Robotics and autonomous vehicles:** Synthetic modelling in robotics offers safe alternatives to plan for scenarios where it is impossible to obtain sufficient training data, such as real-world testing to detect near-misses or failures.[21] For example, Waymo created a digital twin of San Francisco's streetscapes to test autonomous vehicle behaviour in diverse driving conditions, including those not easily encountered – or safe to replicate – in real life.[22]

For cost-effectiveness Providing an affordable alternative to expensive, high-quality datasets needed for robust data analysis, AI model training and system testing.	**Testing, training and development:** In cases where data is sensitive, scarce or expensive to collect – such as financial records, medical records and satellite imagery – synthetic datasets enable more distributed access to quality data for testing, training and development.
For personalized AI and ML To create personalized datasets for specific use cases in AI training, where individual-specific data cannot be easily collected due to privacy or availability constraints.	**Healthcare:** Generating synthetic patient data based on common medical conditions can personalize treatment plans without needing real patient data. **E-commerce:** Creating synthetic user profiles for training recommendation engines can personalize shopping experiences.



3 Strategy and governance

As synthetic data becomes increasingly prevalent in the broader data ecosystem, it can also be misused or exploited to generate false representations and undermine public trust. Additionally, the strategic value of synthetic data can only be leveraged if its implementation is carried out responsibly. Dimensions like privacy, accuracy and representation that synthetic data is designed to benefit can instead be worsened if its risks are not addressed. The most critical issues include:[23]

- **Representativeness and bias.** If synthetic data is generated from biased or non-inclusive sources, or without the involvement of under-represented groups, models built from such data can perpetuate or even amplify existing inequalities. For example, biased healthcare training datasets may result in misdiagnosis or unequal care.
- **Accuracy and utility.** High-quality source data and robust generative processes are critical; otherwise, errors can be introduced in downstream systems. For example, synthetic images meant to train computer vision systems must reflect realistic lighting, motion and occlusion characteristics.
- **Model collapse.** Overreliance on synthetic data in generative AI model training can lead to deterioration in model performance, as it may not capture real-world complexity – a risk known as "model collapse" or "model autophagy." Clear identification, documentation and traceability of synthetic data are vital to maintain transparency and model integrity.
- **Provenance and traceability.** Without reliable metadata or provenance tracking, users cannot assess data origin or understand whether a dataset is synthetic, AI-generated or a mix of both to make informed decisions. Poor traceability also increases the risk of model collapse if synthetic data is unknowingly integrated into the AI training set.
- **Privacy and confidentiality.** While synthetic data is often seen as privacy-preserving, the level of protection depends on the generative method. Poorly anonymized datasets can leak sensitive information about individuals or groups or be vulnerable to deanonymization, especially when linked to external datasets in unsecured environments.
- **Misuse and deception.** A well-known risk with AI-generated synthetic media is the potential for misuse to create deepfakes or other convincing but deceptive material. Without clear legal frameworks for attribution and labelling, malicious actors can exploit synthetic or AI-generated data to impersonate individuals, spread disinformation or violate consent.
- **Erosion of public trust.** As synthetic content proliferates, public scepticism about data authenticity grows – even for organic data. This "liar's dividend" threatens societal trust and is worsened by poor disclosure or traceability practices.

Recommendations

Given the outlined risks and challenges, key decision-makers must carefully consider the trade-offs when deciding whether and how to leverage synthetic data. Typically, in organizations, the use of synthetic data is shaped by a collaboration between developers and adopters (e.g. researchers and data scientists) with regulators and policy-makers (e.g. legal advisors, executive leadership and policy teams).

For developers and adopters:

- **Prioritize model quality:** Implement quality assessment protocols to ensure that synthetic data generation models reflect relevant dimensions while also preserving real-world data distributions and meeting privacy and fairness standards.
- **Invest in robust traceability and provenance:** Implement robust systems to track data origins and transformations, using metadata to identify synthetic elements and their sources. Upfront investment is essential, as retroactive tracing can be costly or impossible.
- **Ensure transparency:** Make data generation processes transparent to distinguish synthetic data from organic data.
- **Implement technical safeguards:** Techniques like watermarking, cryptographic provenance or dataset "nutrition labels" build trust and should be combined with human oversight for high-risk applications.
- **Diversify stakeholder engagement:** Involve diverse communities in governance to identify risks, enhance legitimacy and assess for biases, especially with marginalized groups.
- **Mitigate model collapse:** Avoid relying solely on synthetic data for training AI models. Use hybrid approaches that combine synthetic and organic data,[24] and incorporate self-correction mechanisms based on organic data distributions.[25]

For regulators and policy-makers:

- **Tailor governance:** Not all synthetic data is created equal. Governance frameworks must distinguish between synthetic data intended to replicate real-world distributions and AI-generated data created for entertainment, expression or model training. Regulations should specify intended use, impact and safeguards for each category.
- **Develop context-aware standards:** Support the creation of sector-specific standards and related benchmarks, such as those for responsible AI, which safeguard an organization's long-term ability to innovate responsibly. It would be useful to build on efforts by privacy regulators (e.g. the European Commission,[26] Personal Data Protection Commission Singapore[27] and the Information Commissioner's Office in the United Kingdom[28]) and international organizations (e.g. the United Nations), especially for confidentiality-protecting public data releases.
- **Promote education and capacity-building:** Provide guidance for developers, regulators and decision-makers on when and how to use synthetic data responsibly, by using tools like impact assessments, provenance checklists and red-teaming exercises.

Conclusion

Synthetic data is becoming an essential tool in the data ecosystem, addressing many of the limitations of real-world data from data scarcity to privacy. To unlock its full potential and ensure responsible use, collaboration among developers, data scientists, policy-makers and organizational leaders is essential. This requires strong governance, a commitment to generating inclusive and high-quality data, and transparent practices that benefit all stakeholders, particularly historically marginalized groups.

The recommendations in this document offer a starting point for harnessing synthetic data in a responsible, equitable and ethical way. Future work by the GFC on Data Frontiers will build on these foundations, aiming to realize the transformative potential of synthetic data while safeguarding public trust and advancing the common good.

Contributors

Global Future Council on Data Frontiers 2025-2026

Lead authors

Justine Gauthier
General Counsel; Head, Artificial Intelligence (AI) Governance, MILA – Quebec Artificial Intelligence Institute

Francesca Rossi
IBM Fellow; Global Leader, Responsible AI and AI Governance, IBM

Arun Sundararajan
Harold Price Professor, Entrepreneurship and Technology, Stern School of Business, New York University; GFC co-chair

Lauren Woodman
Chief Executive Officer, DataKind; GFC co-chair

Members

Rachel Adams
Founder; CEO, Global Center on AI Governance

Kathy Baxter
Principal Architect, Responsible AI and Technology, Salesforce

Alberto Giovanni Busetto
Chief Artificial Intelligence Officer, HealthAI

Nighat Dad
Founder; Executive Director, Digital Rights Foundation

Sandhya Devanathan
Head; Vice-President, Meta India, Meta Platforms

Sara Hooker
Vice-President, Research, Cohere

Maui Hudson
Associate Professor; Director, Te Kotahi Research Institute, University of Waikato

Kathrin Gabriele Kind
Chief Data Scientist, Cognizant Technology Solutions

Jae Lee
Co-Founder; Chief Executive Officer, Twelve Labs

Zhang Linghan
Professor, Institute of Data Rule of Law, China University of Political Science and Law (CUPL)

Vukosi Marivate
Chair, Data Science and Professor of Computer Science, University of Pretoria; Co-founder, Lelapa AI

Graciela Márquez Colín
President, National Institute of Statistics and Geography, Mexico

Angela Oduor Lungati
Executive Director, Ushahidi

Fawad Qureshi
Global Field Chief Technology Officer, Snowflake

Crystal Rugege
Managing Director, Centre for the Fourth Industrial Revolution, Rwanda

Anna Tumadóttir
Chief Executive Officer, Creative Commons

World Economic Forum

Council managers

Casey Price
Specialist, AI/Data Governance and Community, Centre for AI Excellence

Karla Yee Amezaga
Initiatives Lead, AI and Data Governance, Centre for AI Excellence

Acknowledgements

Henry Ajder
Founder, Latent Space Advisory

Maria Basso
Head, AI Applications and Impact, Centre for AI Excellence, World Economic Forum

Daniel Dobrygowski
Head, Governance and Trust, Centre for AI Excellence, World Economic Forum

Audrey Duet
Head, Data and AI Innovation, Centre for AI Excellence, World Economic Forum

Khaled El Emam
Professor, School of Epidemiology and Public Health, University of Ottawa; Senior Scientist, Children's Hospital of Eastern Ontario Research Institute

Cathy Li
Head, Centre for AI Excellence; Member of the Executive Committee, World Economic Forum

Stefaan G. Verhulst
Co-Founder; Chief Research and Development Officer; Director, Data Program, GovLab, New York University

Elio Atenógenes Villaseñor García
Director, Data Science Laboratory, National Institute of Statistics and Geography, Mexico

Production

Michela Liberale Dorbolò
Designer, World Economic Forum

Madhur Singh
Editor, World Economic Forum

Endnotes

1. The World Economic Forum's network of Global Future Councils (GFCs) is the world's foremost multistakeholder and interdisciplinary knowledge network dedicated to promoting innovative thinking to shape a more resilient, inclusive and sustainable future. Learn more about the GFC on Data Frontiers here: https://initiatives.weforum.org/global-future-council-on-data-frontiers/home.
2. For the purposes of this publication, data is considered at the level of datasets or collections, rather than individual elements such as pixels, images and documents.
3. IBM. (2024, October 1). *What is data?* https://www.ibm.com/think/topics/data.
4. Johnston, P. (2025, April 1). *The Role of Organic and Synthetic Data in AI Safety and Security*. ActiveFence. https://www.activefence.com/the-role-of-organic-and-synthetic-data-in-ai-safety-and-security.
5. Synthetic data may be created through a variety of techniques, including statistical models (e.g. generative adversarial networks or GANs, diffusion models, and generative AI models such as large language models or LLMs); mathematical models (e.g. physical simulations and differential equations); rule-based systems and hybrid approaches. See https://aws.amazon.com/what-is/synthetic-data/ for discussion of various generation techniques.
6. Rubin, D. (1993). Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics*. 9: 461–468.
7. European Data Protection Supervisor. (n.d.). *Synthetic Data*. https://www.edps.europa.eu/press-publications/publications/techsonar/synthetic-data_en.
8. Cao, Y., et. al. (2023, March 7). *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*. arXiv:2303.04226v1.
9. Goyal, M., & Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. *Electronics*, 13(17), 3509.
10. IBM. (2023, January 31). *What is synthetic data?* https://www.ibm.com/think/topics/synthetic-data.
11. United Nations. Economic Commission for Europe. (2023). *Synthetic Data for Official Statistics: A Starter Guide*. https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide.
12. Peppin, A., et. al. (2025, May 27). *The Multilingual Divide and its Impact on Global AI Safety*. arXiv:2505.21344v1.
13. Peng, C., et. al. (2023, May 22). *A Study of Generative Large Language Model for Medical Research and Healthcare*. arXiv:2305.13523v1; Juwara, L., et. Al. (2024, April 12). *An evaluation of synthetic data augmentation for mitigating covariate bias in health data*. doi: 10.1016/100946.
14. El Kababji, S., et. al. (2025, March 5). *Augmenting Insufficiently Accruing Oncology Clinical Trials Using Generative Models: Validation Study*. doi: 10.2196/66821.
15. Terblanche, C., et. al. (2024, July 11). The development of synthetic child speech in three South African languages. doi: 10.1080/2374312.
16. World Economic Forum. (2024, September). *Advancing Data Equity: An Action-Oriented Framework*. https://www3.weforum.org/docs/WEF_Advancing_Data_Equity_2024.pdf.
17. ByteDance. (2025, May 12). *Bytedance's Seed1.5-Embedding Model Achieves Sota in Retrieval: Training Details Unveiled*. https://seed.bytedance.com/en/blog/bytedance-s-seed1-5-embedding-model-achieves-sota-in-retrieval-training-details-unveiled.
18. Karpukhin, et. al (2019, February 5). *Training on Synthetic Noise Improves Robustness to Natural Noise in Machine Translation*. arXiv:1902.01509v1.
19. Shuryak I. (2017). *Advantages of Synthetic Noise and Machine Learning for Analyzing Radioecological Data Sets. PloS one*, 12(1), e0170007. https://doi.org/10.1371/journal.pone.0170007.
20. De Vries & Thierens (2024, September 23). Generating the Ground Truth: Synthetic Data for Soft Label and Label Noise Research. arXiv:2309.04318v2.
21. Stanford Online. (2025 April 30). *Stanford Seminar - Towards Open World Robot Safety* [Video]. YouTube. https://www.youtube.com/watch?app=desktop&v=zRq_3f4qrcU.
22. Tancik, M., et. al. (2022, February 10). *Block-NeRF: Scalable Large Scene Neural View Synthesis*. arXiv:2202.05263v1.
23. De Wilde, P., et. al. (2024, February 29). *Recommendations on the Use of Synthetic Data to Train AI Models.* United Nations University.
24. Kazdan, J., et. al. (2025, March 17). *Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World*. arXiv:2410.16713v4.
25. Gillman, N., et. al. (2024, June 10). *Self-Correcting Self-Consuming Loops for Generative Model Training*. arXiv:2402.07087v3.
26. European Commission. (n.d.). *Data protection*. https://commission.europa.eu/law/law-topic/data-protection_en.
27. Personal Data Protection Commission Singapore. (n.d.). *Privacy Enhancing Technology (PET): Proposed Guide on Synthetic Data Generation*. https://www.pdpc.gov.sg/help-and-resources/2024/07/proposed-guide-on-synthetic-data-generation.
28. United Kingdom Information Commissioner's Office (ICO). (n.d.) *Guidance on Privacy-Enhancing Technologies.* https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/.

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744
contact@weforum.org
www.weforum.org