Bansilal Ramnath Agarwal Charitable Trust's
Vishwakarma Institute of Technology ,Pune-411037
(An Autonomous Institute affiliated to Savitribai Phule University)
DEPARTMENT OF INFORMATION TECHNOLOGY
Synthetic Data: Training AI Without Real People



Welcome to the IT Bulletin – October 2026 Edition

This issue explores the growing role of synthetic data in artificial intelligence, where AI models are trained using artificially generated data instead of relying entirely on real people's information. As AI becomes increasingly integrated into education, healthcare, software development, research, and everyday life, protecting personal information has become more important than ever.

Synthetic data is artificially created to imitate the patterns and characteristics of real-world data without directly using actual personal records. For example, instead of training an AI model using real students' academic records, developers can generate fictional student profiles with different marks, attendance records, and learning patterns. This allows AI systems to learn and improve while reducing the exposure of sensitive personal information.

By using synthetic data, students, researchers, and developers can explore new possibilities in AI development while supporting privacy, security, and responsible innovation. It can also help create datasets for situations where real-world data is limited, expensive, difficult to access, or sensitive. However, synthetic data must be carefully designed and tested to avoid bias, unrealistic patterns, and the accidental reproduction of information from the original data.

Discover how synthetic data is shaping the future of artificial intelligence, making AI development more privacy-conscious, accessible, and ethical — and why training AI without directly relying on real people's data could be an important step towards building more trustworthy technology.

Did you know?

Synthetic data does not have to come from real people at all. It can be generated entirely through computer simulations, mathematical models, or AI systems.

AI can learn from fake people! Synthetic data creates fictional people, profiles, and scenarios that AI models can use for training without directly relying on real individuals.

Fake data can help save real lives!

Artificially generated medical scans and patient records can help researchers develop and test diagnostic AI while reducing reliance on sensitive patient information.

SYNTHETIC DATA : WHAT IS IT?



Why it matters?

Real data is often scarce, expensive, biased, or locked behind privacy rules. Synthetic data offers a way around these limits, letting teams build and test systems faster and at lower cost, while keeping sensitive personal details out of the picture.

Key use cases

It is widely used to train machine learning models, test software, and share datasets in healthcare, finance, and retail without exposing customer identities. It also helps fill gaps by creating rare edge cases, such as fraud patterns or unusual medical conditions, that real datasets seldom capture well.

What it is?

Synthetic data is artificially generated information that mimics the statistical patterns and structure of real-world data, without being collected from actual events or people. It can take the form of tables, text, images, audio, or sensor readings, created by algorithms, simulations, or generative AI models.

How it's made

Techniques range from simple rule-based generation and statistical sampling to advanced methods like GANs, variational autoencoders, and large language models. Simulation engines are also common, such as virtual environments used to generate driving scenarios for autonomous vehicles.

Did you know?

- It's older than AI hype. The idea dates back to the early 1990s, when statistician Donald Rubin proposed generating fake datasets that preserve the patterns of census-style data while protecting people's identities.
- Faces that never lived. Websites like "This Person Does Not Exist" show lifelike portraits generated entirely by AI. No camera, no subject, no real person involved.
- Xbox taught itself to see you. Microsoft's Kinect motion sensor was trained in part on millions of computer-generated images of human bodies in different poses, since photographing every possible pose in real life would have been impractical.

HOW AI LEARNS WITHOUT REAL PEOPLE



Did you know?

Robots that learned to solve a Rubik's Cube. OpenAI trained a robot hand to manipulate a cube entirely in simulation, randomizing things like friction, lighting, and cube size so it could cope with surprises. It then worked on a real hand with no extra real-world training.

Rendered patients. Researchers have generated synthetic chest X-rays and brain scans to train diagnostic models for rare conditions, where real scans are hard to find and hospitals can't easily share them

LEARNING FROM PATTERNS, NOT PEOPLE

Generally, artificial intelligence requires large quantities of actual human information in order to obtain knowledge. However, artificial intelligence's only prerequisite lies in the presence of existing situations. This is accomplished using synthetic information, in which we generate records that imitate real-life instances. The AI model does not have to use information coming from non-synthetic individuals; it is capable of learning using synthetic representatives of the data.

Synthetic information is produced in three distinct manners. The first way is simulation, which generates synthetic information. The second way is statistical modeling where actual patterns are used to generate new records. Finally, generative AI (GANs, VAEs, LLMs) produces such products as pictures or tables, which resemble actual information.

REAL EXAMPLES

- Software testing. Developers generate fake users, names, addresses, and orders to test apps and databases. A shopping site can simulate 100,000 customers placing orders at once, without touching real customer data.
- Insurance and credit scoring. Insurers create artificial claim records, including rare large losses, to stress-test pricing models and fraud checks without exposing policyholders.
- Cybersecurity drills. Security teams generate fake network traffic containing simulated attacks, so intrusion-detection systems can learn what malicious behavior looks like before a real breach occurs.

THE SCIENCE BEHIND SYNTHETIC DATA GENERATION



BEYOND REAL DATA

Despite its advantages, synthetic data can contain biases, inaccuracies, or unrealistic patterns inherited from the original dataset. Proper validation is necessary to ensure that generated data is reliable, diverse, and useful for real-world applications.

FAKE IT TO MAKE IT: THE SCIENCE BEHIND SYNTHETIC DATA

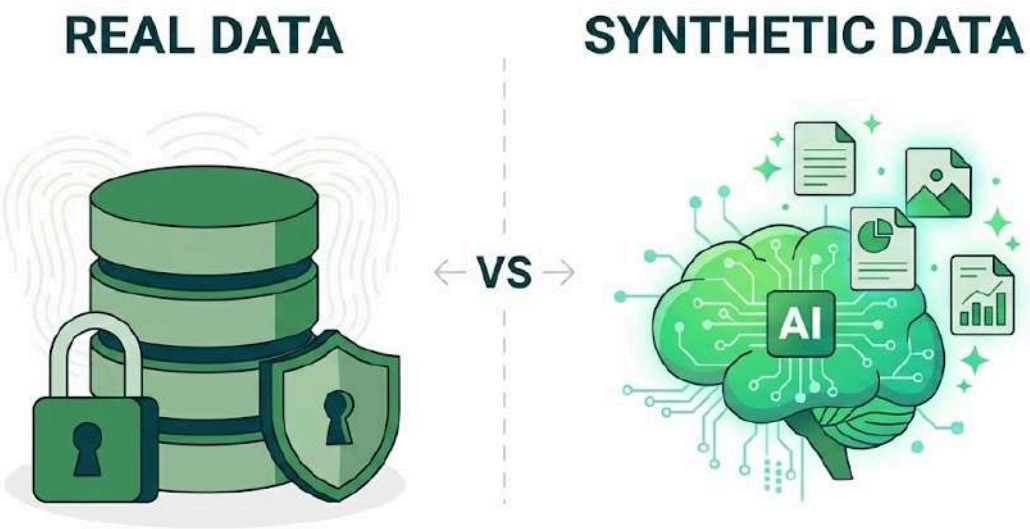
Synthetic data generation is an innovative technique that uses mathematics, statistics, and artificial intelligence to create artificial datasets that resemble real-world data. Unlike traditional data collection, synthetic data is generated using algorithms that learn patterns and relationships from existing information. Technologies such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), diffusion models, and Large Language Models (LLMs) enable computers to generate realistic images, text, and structured datasets.

Synthetic data is widely used in healthcare, autonomous driving, finance, and cybersecurity, where collecting real-world data can be expensive, time-consuming, or restricted by privacy concerns. It helps train AI models, simulate rare scenarios, and improve machine learning performance. However, challenges such as data bias, accuracy, and privacy risks must be carefully addressed. As artificial intelligence continues to evolve, synthetic data generation is becoming an important tool for building more efficient, reliable, and intelligent systems.

ARTIFICIAL. INTELLIGENT. INNOVATIVE.

Generative Adversarial Networks (GANs) use two neural networks, a generator and a discriminator, that compete to produce realistic data. Variational Autoencoders (VAEs) learn compressed representations of data and generate new samples from a latent space. Diffusion models generate data by gradually removing noise, while Large Language Models can produce synthetic text, conversations, and code.

REAL DATA VS. SYNTHETIC DATA



BRIDGING REALITY AND GENERATIVE INNOVATION

Real datasets are real human information, but are often subject to strict privacy and data protection laws on a large scale. In addition, real-world data often does not have enough examples of rare situations, such as unusual medical conditions or particular fraud patterns.

Synthetic data utilizes algorithms, statistical modeling, or generative AI (e.g. GANs, VAEs, LLMs) to replicate real-world distributions, thereby overcoming these bottlenecks. This allows developers to simulate dangerous edge cases (rare accidents in autonomous driving, for example) without putting human safety at risk. This is particularly useful when training systems in fields like healthcare, finance and cybersecurity as it means you do not have to rely on sensitive personal records. The future of AI development will not be to eliminate empirical data entirely, but to embrace a hybrid approach that uses synthetic information to scale up the volume, and oversight by humans to maintain ethical fairness.

THE CORE TRADEOFF

In many cases real world data is limited, costly or subject to tight data privacy regulations. Synthetic data can be used to improve speed and reduce costs when building and testing systems whilst keeping personal information private. Synthetic data needs to be checked carefully to ensure it doesn't contain errors or any bias from the original dataset.

STRATEGIC DEPLOYMENT

In scenarios where actual human input is needed or a baseline accuracy is to be established, actual data is still important. Synthetic data is ideal for cases where original datasets are scarce, to simulate extreme edge cases, or when a strict "Privacy by Design" approach requires data minimization to protect human identity.

PRIVACY BY DESIGN: PROTECTING HUMAN IDENTITY



Privacy by Design is an approach that integrates privacy and data protection into technology from the very beginning. Instead of treating privacy as an afterthought, systems are designed to minimize data collection, protect sensitive information, and ensure that personal identities remain secure. Techniques such as data anonymization, encryption, and differential privacy help reduce the risk of exposing individuals' identities.

As artificial intelligence continues to evolve, Privacy by Design becomes increasingly important in protecting human identity. Synthetic data generation can support this approach by creating artificial datasets that reduce dependence on real personal information. However, synthetic data must be carefully evaluated to prevent the possibility of revealing sensitive details. By making privacy a fundamental part of technological development, we can build AI systems that are both innovative and trustworthy.

INNOVATION WITHOUT EXPOSURE

One important principle of Privacy by Design is data minimization, which means collecting only the information necessary for a specific purpose. Techniques such as encryption, anonymization, and access control help protect sensitive data and reduce the chances of privacy breaches.

DID YOU KNOW?

Techniques such as encryption, anonymization, and differential privacy help protect individuals from identity theft and data misuse

Privacy by Design makes data protection a core part of technology, helping build secure, trustworthy AI systems that respect human identity while enabling innovation.

GENERATIVE AI: CREATING ARTIFICIAL DATASETS



WHY IS IT USEFUL?

- Privacy: Decreases reliance on revealing sensitive personal data.
- Availability: Useful where real-world data is hard to come by.
- Customization: Creates data tailored to particular situations.
- Creativity: Aids in AI development and testing.

WHAT IS GENERATIVE AI?

Generative AI model is a type of artificial intelligence that learns patterns, structures, and relationships within the inputted data to output new data. This kind of algorithm can generate artificial datasets, pictures, text, and other samples that are similar to real data.

In case the AI model generates a dataset, the process is referred to as synthetic data generation. The algorithms learn how the real data looks like and then create new data based on the learned patterns.

The generated data can be used for machine learning purposes, simulating a particular situation, and conducting different kinds of studies.It is very useful to generate data if it is impossible or too expensive to obtain the real data because of some limitations. However, the generated data should be verified first.By creating realistic artificial datasets, Generative AI opens new possibilities for experimentation and innovation across industries. It helps build smarter AI systems while highlighting the importance of responsible and reliable data generation.

BENEFITS & CHALLENGES

BENEFITS

Synthetic data creates additional data points when the original data is scarce or difficult to obtain. This helps in AI training, research, and testing without being exposed to the actual personal information directly.

CHALLENGES

Synthetic data could be biased or inaccurate, resulting in low-quality AI operations. Proper validation is key to ensuring the quality of data.

REAL-WORLD APPLICATIONS OF SYNTHETIC DATA-



FROM SIMULATION TO SOLUTION: SYNTHETIC DATA IN THE REAL WORLD

Every industry runs into the same obstacle sooner or later: the data it needs is private, scarce, or too costly to collect. Synthetic data, information generated by algorithms to reflect the statistical behavior of real datasets, offers a way around that obstacle, and its uses are expanding quickly.

Medical teams can now build and test diagnostic tools using computer-generated patient histories that contain no real person's details. A hospital in one country can share such records with researchers in another without breaching confidentiality rules. For diseases that affect only a handful of people, artificially produced scans and lab results give algorithms enough examples to learn from, something real-world collection alone could never provide.

Synthetic Data: Lesser-Known Facts
Disaster response teams train AI on disasters that haven't happened.
Generated satellite and drone imagery of floods, wildfires, and collapsed buildings helps damage-assessment models prepare for events with little real footage.
Space agencies simulate alien terrain.

OCR learned to read from fake text.
Text-recognition models are often trained on rendered words in thousands of fonts, distortions, and backgrounds, because real labeled text images are hard to gather at scale.

Rover navigation systems are tested on computer-generated Martian and lunar landscapes, since you can't collect millions of labeled photos on another planet.

Drug discovery uses invented molecules.
Generative models propose synthetic molecular structures, helping researchers explore chemical possibilities no lab has made yet.

BIAS IN AI:CAN SYNTHETIC DATA MAKE IT FAIRER?



Did you know?

An AI trained on biased data can learn to discriminate even when no one intended it to! Synthetic data can introduce carefully designed examples of underrepresented groups, helping AI make fairer decisions. But there's a twist: artificial data can inherit real-world prejudice unless it is designed and tested carefully.

AI can perform well overall while making more mistakes for particular communities. Synthetic data can fill gaps in training examples, but a generator that copies historical data may also copy its discrimination

ENGINEERING FAIRER DATA

HOW IT WORKS

CTGAN (Conditional Tabular Generative Adversarial Network) learns relationships between columns and generates artificial records. Targeted sampling can broaden coverage of underrepresented groups and outcomes, but CTGAN does not guarantee fairness

DECAF models causal relationships and removes selected unfair pathways during generation. Its guarantees depend on the causal assumptions and fairness definition.

WHEN BIAS RETURNS

Equal group counts cannot repair biased labels. A generator that learns from discriminatory hiring records may reproduce their patterns.

Model collapse is another risk. A 2024 Nature study showed that repeatedly training on earlier models' outputs can erase rare patterns, raising concerns for underrepresented groups. This concerns recursive training; it does not mean all synthetic-data augmentation fails.

RESEARCH SPOTLIGHT

A 2024 study evaluated five synthetic data generation methods across four datasets. In one test using the Adult dataset, TVAE (Tabular Variational Autoencoder) reduced the equalized-odds gap from 0.122 to 0.079, while prediction accuracy decreased slightly from 86.8% to 85.7%.

What does this mean?

The smaller gap suggests more balanced true-positive and false-positive rates across groups, but it does not guarantee overall fairness.

Key takeaway: Synthetic data can help reduce algorithmic bias, but models should be tested on held-out real-world data and across intersecting identities, such as gender and ethnicity, to ensure fairness improvements generalize beyond the training data.

BENEFITS, RISKS AND ETHICAL CHALLENGES



A POWERFUL TOOL, USED CAREFULLY

Artificial data is revolutionizing the development process for A.I. by providing safe and quick methods to train A.I. models. This innovation carries a duty and should be evaluated and validated before its acceptance.

Synthetic data is A.I. generated information that copies real data without belonging to any real person. Its biggest benefit is privacy, as hospitals, banks and schools can train A.I. without exposing confidential details. It is also faster and cheaper than collecting and labelling real data, which can take months. Most importantly, it helps A.I. learn from rare situations such as fraud or uncommon diseases, where real examples are very limited.

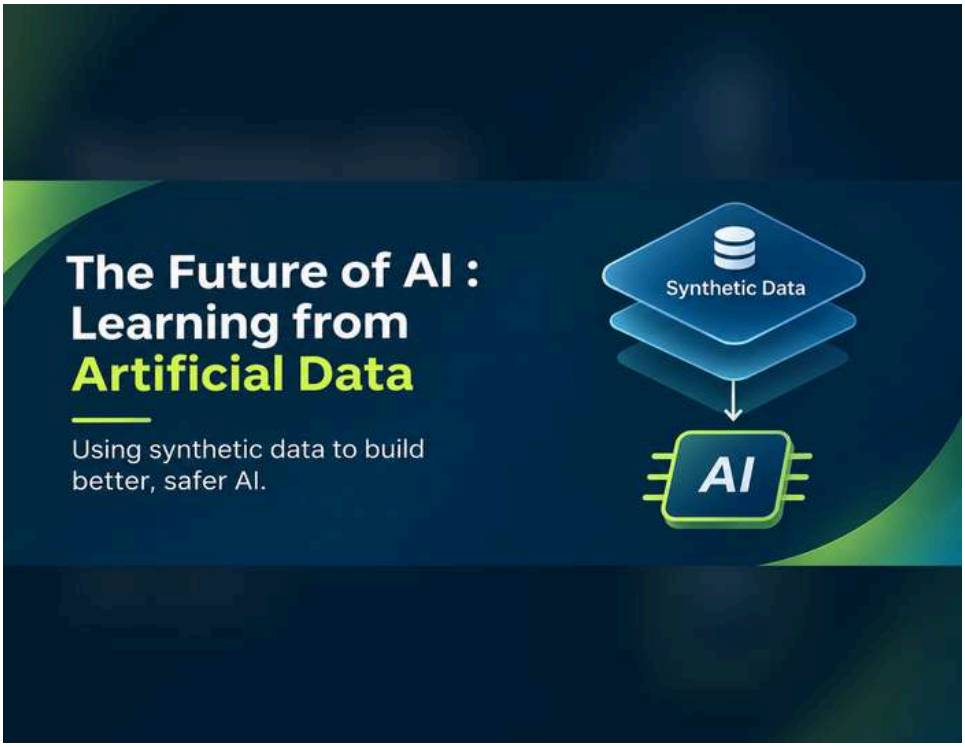
However, the synthetic data has got some issues of its own. First, for example, if the synthetic data is very simple, then the AI may perform well in tests but may not perform well in real-time situations. Any bias in the original dataset can be propagated and amplified through the synthesis process, while any carelessness in generating the data might lead to information leakage about real people.

WHO IS RESPONSIBLE?

In the case where an artificial intelligence built using artificial data comes up with a wrong decision, especially within the health and finance sector, who takes responsibility?

Should humans also know how an artificial intelligence was built? Ethical artificial intelligence requires transparency, robust validation, and proper regulation to ensure that artificial data benefits humans and not harm them.

THE FUTURE OF AI :LEARNING FROM ARTIFICIAL DATA



LEARNING FROM ARTIFICIAL DATA

Genuine information is scanty, expensive, and secured by laws regarding privacy. However, synthetic data can be produced according to requirements and in large quantities.

A SMART MIX OF REAL AND SYNTHETIC

The direction in which AI is advancing indicates that artificial data will indeed be of substantial importance there. It will enable the creation of digital twins (let it be machines, urban areas, and human patients), higher safety of health-care development procedures as well as self-driving cars that have been trained digitally. It also means it will be easier to create AI solutions for students or small businesses without the need for huge datasets.

Nevertheless, there are some issues connected with synthetic data. In case the training process is based only on artificial data, the resulting product may lose its quality during time. It can also happen that some biases will only be repeated and amplified instead of being detected.

WHAT COMES NEXT



AI training AI, digital twins, and personalized healthcare research without sharing patient data.

Humans keep it accurate, fair, and ethical. Thank you!

REFERENCES

1. Nobani, N., Officioso, G., Pallucchini, F., Sperlì, G., & Mercorio, F. (2026). “Synthetic Data Generation: A Tertiary Study.” Information Processing & Management, 63(6), Article 104715.
2. Ma, Q., & Rajtmajer, S. (2026). “Private Seeds, Public LLMs: Realistic and Privacy-Preserving Synthetic Data Generation.” Findings of the Association for Computational Linguistics: ACL 2026, 189–210.
3. Ponomareva, N., Xu, Z., McMahan, H. B., Kairouz, P., Rosenblatt, L., Cohen-Addad, V., et al. (2026). “How to DP-Fy Your Data: A Practical Guide to Generating Synthetic Data With Differential Privacy.” Journal of Artificial Intelligence Research, 86.
4. Zuo, M., Kang, I., Patterson, S., & Seneviratne, O. (2026). “Measuring Privacy Risks and Tradeoffs in Financial Synthetic Data Generation.” Companion Proceedings of the ACM Web Conference 2026, 1339–1349.
5. “CLOVER: A Framework for Benchmarking Synthetic Data Generation Methods Balancing Utility and Privacy in Healthcare.” (2026). Artificial Intelligence in the Life Sciences, 9, Article 100155.
6. Abdalla, H. B., Kumar, Y., Li, J. J., Kruger, D., & Ehimwenma, K. (2026). “Beyond Realism: A Utility-Fidelity-Privacy Framework for Trustworthy Synthetic Data.” Journal of Data and Information Science, 11(3), 217–235.
7. Liu, Y., & Reiter, J. P. (2026). “Synthetic Data: A Tool for Privacy Protection and Model Empowerment.” Annual Review of Statistics and Its Application, 14.
8. Kaabachi, B., et al. (2025). “A Scoping Review of Privacy and Utility Metrics in Medical Synthetic Data.” npj Digital Medicine, 8, Article 60. https://www.nature.com/articles/s41746-024-01359-3
9. Kiran, A., Rubini, P., & Saravana Kumar, S. (2025). “Comprehensive Review of Privacy, Utility, and Fairness Offered by Synthetic Data.” IEEE Access, 13, 15795–15811. https://doi.org/10.1109/ACCESS.2025.3532128
10. Liu, Y., Huang, J., Li, Y., Wang, D., & Xiao, B. (2025). “Generative AI Model Privacy: A Survey.” Artificial Intelligence Review, 58, Article 33. https://doi.org/10.1007/s10462-024-11024-6

STUDENT EDITORS



Avantika Chatterjee
TY IT-A



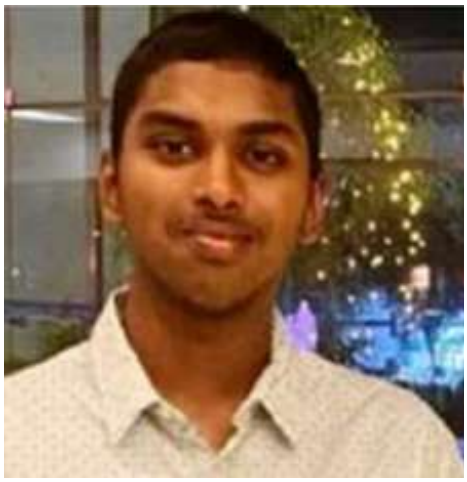
Bhavesh Patil
SY IT-D



Priyanka Kalawadiya
TY IT-C



Srushti Mule
SY IT-D



Kushal Chunduru
TY IT-B



Arya Bhaiya
SY IT-A



Shivam Gunjal
SY IT-D