Scholedge International Journal of *Business Policy & Governance*
ISSN 2394-3351 Vol. 10, Issue 09 (2023)
DOI: https://dx.doi.org/10.19085/sijbpg100901

The Dark Side of Synthetic Data: Reality Gaps, Model Collapse and Hallucinative Analytics

Prof Dr. MS S El Namaki
Dean, Victoria University, School of Management, Switzerland.
Dean (Retired) Maastricht School of Management, MSM, The Netherlands.

Abstract

Data lies at the heart of our artificial intelligence revolution. Massive volumes of data hold the key to the generative AI analytical processes that induce artificial intelligence outcomes. Data are drawn from a wide variety of sources and, as a result, provide an amorphous whole. They are, with a measure of simplification, neither homogeneous, generic nor malleable! Nor are they always available to sustain an argument or complete an algorithm. This could induce reality gaps, model collapse and hallucinations.

What are synthetic data and how do they emerge. And could they lead to what we label as dark side? This will be the focus on the following article.

The article is qualitative in approach. It starts with an identification of the trigger of the problem and the emergence of the need for synthetic data. It then proceeds to define synthetic data, classify it according to a set of criteria, analyze its framework and draw its possible impact on reality status, model collapse and hallucinative analytical outcomes.

What Is Synthetic Data?

The Concept

AI models need much more data and a greater variety than the real world can offer. Synthetic data provides a solution

Synthetic data is artificial data that is generated from original data. It could also take the shape of a model that is trained to reproduce the characteristics and structure of the original data. It is tantamount to the creation of artificial datasets simulating the original dataset's statistical characteristics. Synthetic data allows for the generation of highly diverse or even novel data sets. It is also possible to enhance a training data set by generating synthetic data that does not reproduce the characteristics of the original data set but instead exaggerates certain characteristics.

Synthetic data provides innovative solutions to problems of data scarcity and privacy as well as algorithmic biases commonly used in machine learning applications.

Synthetic data can also be deployed to validate mathematical models and to train machine learning models. Data generated by a computer simulation can be seen as synthetic data.

Cite this article:
El Namaki, MSS. (2023). The Dark Side of Synthetic Data: Reality Gaps, Model Collapse and Hallucinative Analytics. *Scholedge International Journal of Business Policy & Governance*, *10*(9), 87-91.
https://dx.doi.org/10.19085/sijbpg100901

Categories

Synthetic data could be fully synthetic, partly synthetic or hybrid synthetic.

Fully synthetic data is entirely synthetic and contains no information from the original data. The data generator here identifies the density function of features in the real data and estimates their parameters. Partially synthetic data, on the other hand, replaces only the values of selected sensitive features with synthetic values. Real values are replaced only if they contain a high risk of disclosure. Hybrid synthetic data, finally, combines both real and synthetic data. A close record in the synthetic data is chosen for each random record of real data, and then both are combined to form hybrid data.

Synthetic data could have different varieties or forms, also. They could take the form of text, as used in NLP to train language models. They could also be tabular resembling real-life logs or tables. And they, finally, could take a media shape i.e. generated images, videos, and sounds for computer vision applications. (Kajal Singh, 2021)

Generation

How is synthetic data generated? Synthetic data can be generated by resorting to a variety of methods. Those include the following three prime methods.

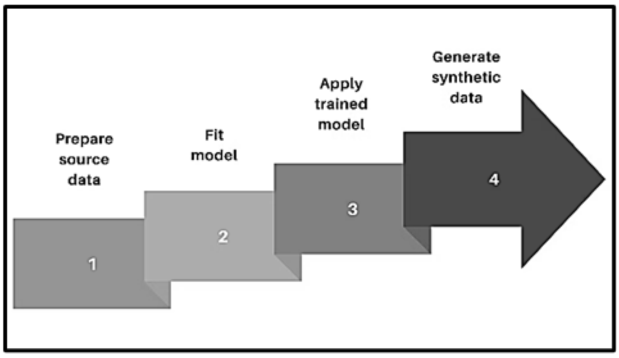

Figure 1: Synthetic data generation

Source: Trampert et al (2021)

- Generative Models where a discovery and learning process leads to the discovery of real data insights and patterns inducing new data sets matching the distribution of the real-world data used for the original training.
- Drawing numbers from a distribution or simply draw, or sample numbers from a distribution creating curve that is loosely based on real-world data
- Agent-Based Modeling (ABM) is a simulation technique where individual agents, people, cells, or even computer programs, are created that interact with each other,

Generative AI approaches synthetic data generation by examining the statistical distributions in a real dataset and creating a new, synthetic one to train other AI models. This synthetic 'pseudo' data is similar but not identical to the original, meaning it can also ensure privacy, skirt data regulations, and be freely shared or distributed.

The Dark Side

Reality Gaps

Risks of synthetic data could be many. A key risk is the "reality gap "or the subtle differences between the synthetic data and the real world. Another dimension is the benchmarking. While synthetic data can be used to benchmark data quality, data bias and algorithmic bias, synthetic data itself can also create (or even amplify) unintended biases

Inferior quality could lead to synthetics. Data quality could fall below the befitting standards of completeness, accuracy, reliability, relevance, and timeliness. Sources of this quality decline could include human error, incompleteness, duplication, inaccuracy, outdatedness and amorphous structure. (Steinhoff, J., & Hind, S. (2024). *Simulation and the Reality Gap: Moments in a Prehistory of Synthetic Data*. (pp. 1-23).

Model Collapse

Model collapse refers to the declining performance of generative AI models especially those relying on synthetic data. Generative AI models based on synthetic data drew attention in recent years for creating inaccurate and nonsensical outputs.

Search for interpretation lead to several explanations. The first is a statistical approximation resulting from the finite number of samples used in training. The second is functional expressivity resulting from limitations in the model's ability to approximate the true data distribution. And the third is functional approximation arising from limitations in the learning procedures themselves, such as the choice of optimization algorithms or loss functions. (Iriondo R, 2024)

It is worth referring here to another research that concluded that models trained on synthetic data, initially lost information from the tails, or extremes, of the true distribution of data, a phenomenon given the label "early model collapse." In later model iterations, the data distribution converged so much that it looked nothing like the original data leading to what was termed "late model collapse." (Shumailov et al, 2024)

The phenomenon of model collapse poses serious ramifications for AI development as it could lead to the forgetting accurate data distributions: the production of bland and generic outputs and difficulties in creativity and innovation. (https://www.ibm. com/think/topics/model-collapse).

Hallucinative Analytics

Put in generic terms, hallucination as a process and hallucinations as nouns have a great variety of definitions. The one that the author subscribes to is that where hallucination as a process is tantamount to ".... sensory perceptions that appear in the absence of stimuli" (kohut, 1978). And hallucinations as nouns are equated to a figment of imagination, an imaginary occurrence or a fictitious invention. ("International Journal of Education, Business and Economics Research ...")

Seen in Generative AI terms, hallucination is a phenomenon where a large language model (LLM)often a Generative AI chatbot or computer vision tool, perceives patterns, objects or algorithms that are nonexistent or imperceptible to an observer i.e. synthetics. Outcomes, then, are inaccurate, nonfactual and nonsensical statements or objects rendering analysis futile. LLM hallucinations could be input-conflicting, context-conflicting and fact-conflicting.

A Speculative View Of The Consequences: Synthetic Data Is A Dangerous Teacher

The web is becoming increasingly a dangerous place to look for your data," (Sina Alemohammad, 2023). A term "MAD" short for "Model Autophagy Disorder" was coined to describe the effects of AI self-consumption. AI models are data-hungry and resort to the net carries the threat that AI output s do not convey validity and authenticity of data inputs.

Generative AI will worsen the situation. Vast generative AI outputs will be used as training material for future generative AI models. As a result, an incredibly significant part of the training material for generative models will be synthetic data produced from generative models. Nvidia predicts that, by 2030, there will be more synthetic data than real data in AI models. And unfortunately, this will be the data applied to train generative models used in high-stake sectors including medicine, therapy, education, and law.

Gartner research indicates that synthetic data will become a dominant force in AI, projecting a situation where 60% of data used in AI and analytics will be synthetically generated by 2030. This shift is driven by the need for AI models to manage various scenarios, including those where real-world data is scarce, expensive to obtain, or regulated. (Gartner, 2022)

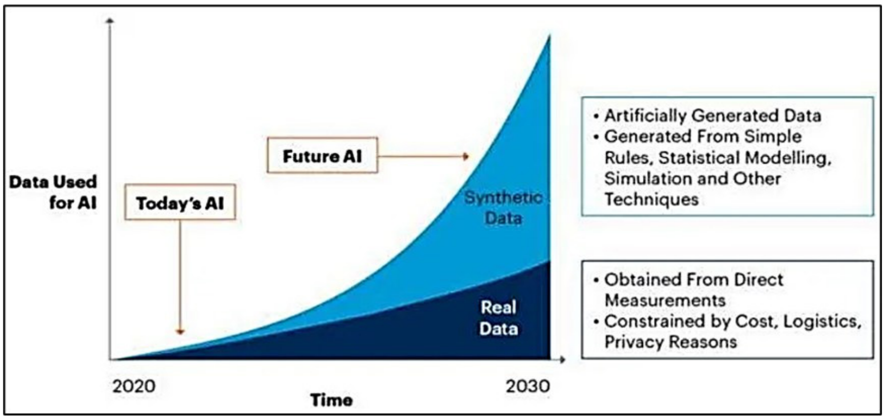

Figure 2. Synthetic data over time.

Gartner 7750175. Can synthetic data drive the future of AI? "AI BUSINESS, September 2, 2022)

Generally, mixing artificial intelligence with synthetic data is set to drive further innovation. AI-synthetic data powered tools are enabling the creation of more realistic training datasets and enhancing model performance. Examples include medical imaging where synthetic MRI scans to train AI models for diagnosis, financial modelling where synthetic financial transactions to assess risk and fraud and autonomous driving were simulation of driving scenarios trains self-driving cars.

Summary and Conclusions

Data lies at the heart of our artificial intelligence revolution. Massive volumes of data hold the key to the generative AI analytical processes that induce artificial intelligence outcomes. Data are drawn from a wide variety of sources and, as a result, provide an amorphous whole. They are, with a measure of simplification, neither homogeneous, generic nor malleable! Nor are they always available to sustain an argument or complete an algorithm. This could induce reality gaps, model collapse and hallucinations.

What are synthetic data and how do they emerge. And could they lead to what we label as dark side? This is the focus of this article.

The article is qualitative in approach. It defines synthetic data, classifies it according to a set of criteria, analyzes its generation and segments it according to specific criteria. It then draws its possible impact on reality status, model collapse and hallucinative analytical outcomes.

A careful conclusion is that AI-synthetic data powered tools are enabling the creation of more realistic training datasets and enhancing model performance. Also, that AI models are data-hungry and resort to the net for data carries the hazard that AI outputs do not convey a proof of the validity and authenticity of data inputs.

References

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., & Baraniuk, R. G. (2023, July 4). Self-Consuming Generative Models Go MAD. ArXiv.org. https://doi.org/10.48550/arXiv.2307.01850

Gartner: Can synthetic data drive the future of AI? | AI Business. (2022). Aibusiness.com. https://aibusiness.com/data/gartner-can-synthetic-data-drive-the-future-of-ai-

Iriondo, R. (2024, August 28). Understanding Model Collapse: A Hidden Threat in Generative AI - Generative AI Lab. Generative AI Lab. https://generativeailab.org/l/trends/understanding-model-collapse/1080/

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

Singh, K. (2021, May 12). Synthetic Data - key benefits, types, generation methods, and challenges! | Towards Data Science. Towards Data Science. https://towardsdatascience.com/synthetic-data-key-benefits-types-generation-methods-and-challenges-11b0ad304b55/

Steinhoff , J., & Hind, S. (2025a). OSF. Doi.org. https://doi.org/10.33767/osf.io/np3vb

The Search for the Self: Selected Writings of Heinz Kohut 1950–1978, Vol. 2 (1978). Edited by Paul Ornstein. International Universities Press, New York. ISBN 0-8236-6016-8

Trampert, P., Rubinstein, D., Boughorbel, F., Schlinkmann, C., Luschkova, M., Slusallek, P., Dahmen, T., & Sandfeld, S. (2021). Deep Neural Networks for Analysis of Microscopy Images—Synthetic Data Generation and Adaptive Sampling. Crystals, 11(3), 258. https://doi.org/10.3390/cryst11030258

Trampert, P., Rubinstein, D., Boughorbel, F., Schlinkmann, C., Luschkova, M., Slusallek, P., Dahmen, T., & Sandfeld, S. (2021b). Deep Neural Networks for Analysis of Microscopy Images—Synthetic Data Generation and Adaptive Sampling. Crystals, 11(3), 258. https://doi.org/10.3390/cryst11030258