

To cite this article: Prof Dr m s s el namaki (2025). AI Data syndrome: could synthetic data lead to hallucinative AI analytics outcomes? International Journal of Education, Business and Economics Research (IJEBER) 5 (2): 144-150

AI DATA SYNDROME: COULD SYNTHETIC DATA LEAD TO HALLUCINATIVE AI ANALYTICS OUTCOMES?

Prof Dr m s s el namaki

Retired Dean Maastricht School of Management, the Netherlands

<https://doi.org/10.59822/IJEBER.2025.5211>

ABSTRACT

Data lies at the heart of our artificial intelligence revolution. Massive volumes of data hold the key to the analytical processes that induce artificial intelligence outcomes. Data is drawn from a wide variety of sources. Yet data are neither homogeneous, generic nor mailable! Data are amorphous whole and carry with it a myriad of problems. This could and does induce "data malfunctions" or data syndrome and the derived issue of data analytics outcome reliability. A hallucinative outcome could emerge. The following article addresses this issue.

The article is qualitative in approach. It identified several triggers of data malignance including, as a prime trigger, synthetic data. Hallucinative outcomes of synthetic data application are projected as an outcome of resort to synthetic data in, among others, LLM software.

The article tries to place this observed data malignance within a synthetic data framework.

The conclusion suggests a set of hypotheses that could be the subject of future research.

KEYWORDS:- Data syndrome, Synthetic data, data analytics, Hallucinative analytics outcomes.

© The Authors 2025

Published Online: April 2025

Published by International Journal of Education, Business and Economics Research (IJEBER) (<https://ijeber.com/>) This article is published under the Creative Commons Attribution (CC BY 4.0) license. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this license may be seen at: <http://creativecommons.org/licences/by/4.0/legalcode>

1.0 AI DATA: THE ESSENCE

Data constitutes a prime input into AI systems.

Data is essentially "information in raw or unorganized forms (such as alphabets, numbers, or symbols) that refer to, or represent, conditions, ideas, or objects.... Data is limitless and present everywhere in the universe. To computers, data include symbols or signals that are input, stored,

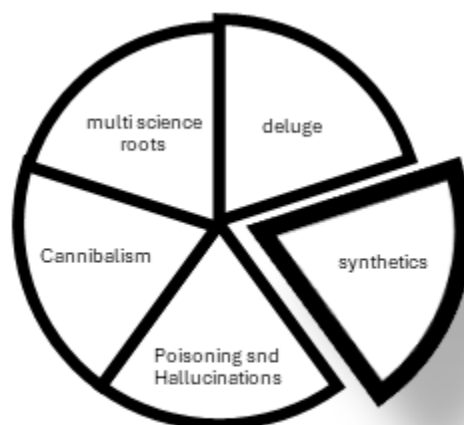
and processed by a computer, for output as usable information. (El Namaki, A Systems Approach to the Artificial Intelligence Concept, 2019).

Data segmentation follows different parameters. Most data can be categorized into four basic types from a machine learning perspective: numerical data, categorical data, time series data, and text data. "Data could also be segmented according to different criteria: useless, nominal, binary, ordinal, and count." (El Namaki, 2019) Time, interval, image, video, audio, and text (Jeff Hale, 2018).

Data could be logged or collected. Data logging is the process of collecting and storing data over a period to analyse specific trends or record data-based events/actions of a system, network, or IT environment. A data logger is an electronic device designed to measure and store data values, often independently of a PC.

Data collection consists of data acquisition, data labelling, and improvement of existing data or models. As AI technology advances it will develop an autonomous ability to seek, classify and validate data.

Figure: The multiple triggers of the data syndrome



2.0 DATA SYNDROME

Data holds the key to AI performance, outcomes, and system flows. Data analytics lead to key to those variables. (El namaki, 2019)

AI is either data centric or data driven.

Data-centric artificial intelligence (AI) is a methodology that prioritizes the management, quality, and relevance of data. The primary emphasis of a Data-Centric AI framework is to guarantee that the data utilized for training and modelling purposes is exhaustive, of superior quality, and in accordance with the goals of the AI system. Data-Driven AI is distinguished by its dependence on predictions, insights, and patterns that are extracted directly from data. Data-centric AI establishes

the foundation for efficient modelling. Conversely, Data-Driven AI leverages this meticulously curated data to facilitate ongoing learning, adjustment, and enhancement. (El namaki, 2023).

The problem here is that data centric and data driven AI rely on a wide variety of data sources. Those could suffer from many flaws and could induce many illnesses.

Five sources could trigger this data malaise: data synthetics, data deluge, multi science roots of data, data poisoning and data cannibalism (see figure). One of those constitutes, in the author's view, the prime trigger: synthetic data. It could lead to deviant analytics outcome and hallucinative outcomes. Those will be the focus of the following analysis.

3.0 SYNTHETIC DATA, POSSIBLE PRIME SOURCE OF DATA SYNDROME

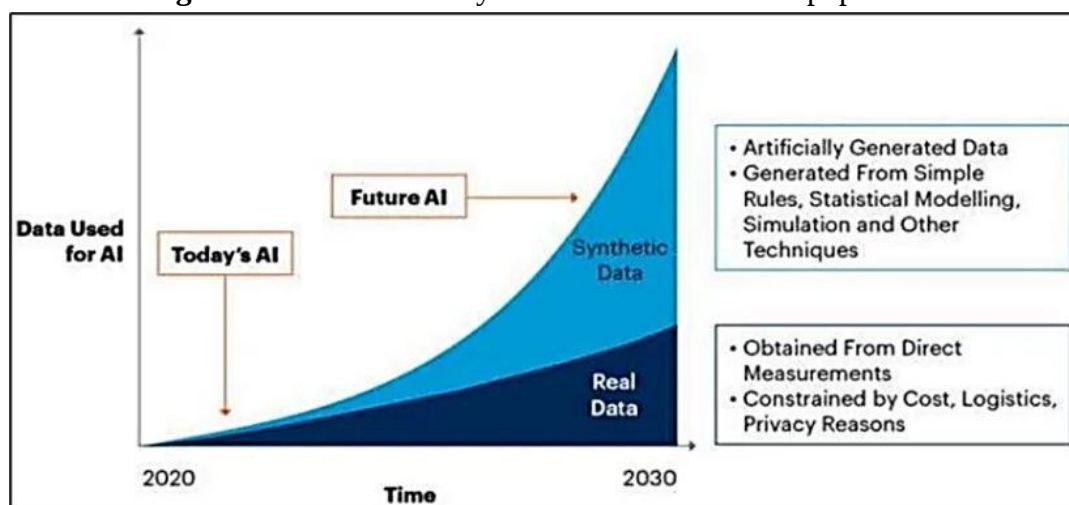
AI models need much more data and a greater variety than the real world can offer. Synthetic data provides a solution.

Synthetic data is artificial data that is generated from original data. ("Synthetic Data - European Data Protection Supervisor") 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. ("for Artificial Intelligence - United Nations Institute for Disarmament ...") 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 provide innovative solutions to problems of data scarcity and privacy as well as algorithmic biases commonly used in machine learning applications.

Synthetic data can be deployed to validate mathematical models and to train machine learning models. Data generated by a computer simulation can be seen as synthetic data. (Deng 2023).

Figure: Contribution of synthetic data to total data population



Source: Gartner 750175

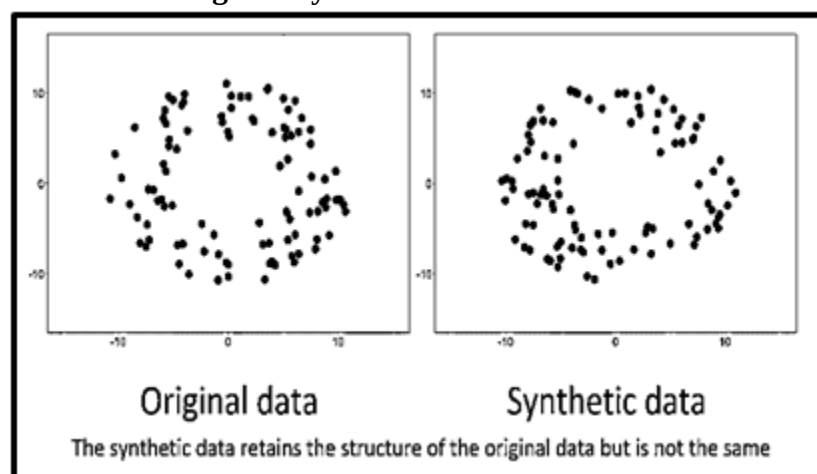
Generative AI approaches synthetic data by examining the statistical distributions in a real dataset and creating a new, synthetic one to train other AI models. This synthetic ‘pseudo’ data is similar but not identical to the original, meaning it can also ensure privacy, skirt data regulations, and be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

How is synthetic data generated? Synthetic data can be generated using a variety of methods. Prime among those are the Generative Models with unsupervised learning task where a discovery and learning process leads to the discovery of the insights and patterns in real data in such a way that the model can be used to output new examples that match the same distribution as the real-world data it was trained on.

Prime uses of synthetic data include privacy-by-design, safe data retention, testing software products and services, training ML and AI models, sharing data within organizations or without the organization, creating synthetic marketplaces and data exchanges and scaling out developments in the cloud. (Alex Watson, 2022), Data sensitive industries also turn to synthetic data to ensure privacy and regulatory compliance. Businesses are also leveraging synthetic data in domains where data availability is limited, or privacy is a concern.

Synthetic data have their flaws. And those could lead to what we term the synthetic data syndrome. A key flaw is the “reality gap” or the subtle differences between the synthetic data and the real world. Another flaw is the benchmarking as synthetic data itself can also create (or even amplify) unintended biases. And yet another flaw could arise from inferior quality of base data. Data quality could fall below the befitting standards of completeness, accuracy, reliability, relevance, and timeliness and could lead to synthetics that are equally inferior.

Figure: Synthetic data vs real data



Source: Synthetic data vs real data: which is a better option? Sumit Singh, Labeler, Nov 15, 2022

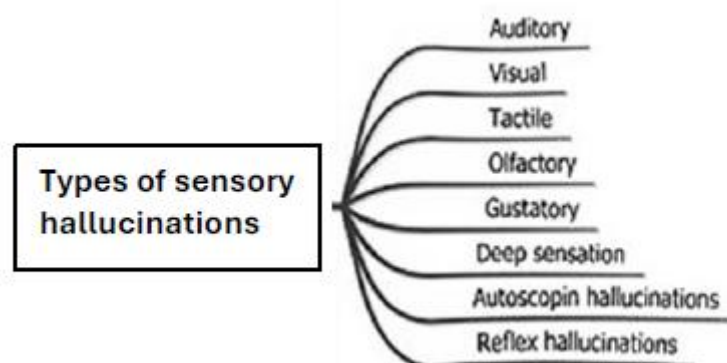
4.0 HALLUCINATION THE POSSIBLE OUTCOME OF SYNTHETIC DATA SYNDROME.

4.1 What is hallucination?

Hallucination as a process and hallucinations as nouns have a vast 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” (Cohut, 2020). And hallucinations as nouns are equated to a figment of imagination, an imaginary occurrence, or a fictitious invention. It could also mean a false belief or impression; illusion; delusion. Or a sensory experience of something that does not exist outside the mind, caused by various physical and mental disorders, or by reaction to certain toxic substances, and manifested as visual or auditory.

Hallucinations could also be seen from an individual or business angel. They can, at individual level, be visual (sight hallucinations), auditory (sound hallucinations), olfactory (smell hallucinations), gustatory (taste hallucinations), or tactile (touch hallucinations). (Cohut, 2024). They can, at business level, as we said earlier, take the form of factual errors or inaccuracies, fabricated information, harmful misinformation and or weird answers. (<https://builtin.com/artificial-intelligence/ai-hallucination>)

Figure: Types of sensory generic hallucinations



Source: <https://builtin.com/artificial-intelligence/ai-hallucination>.

4.2 Synthetic data hallucinative analytics outcomes.

Synthetic data inputs could induce hallucinative data analytic outcomes due, primarily, to the workings of Large Language Models (LLMs).

Large language models are trained on massive volumes of data, like articles, books, codes and social media posts and they are particularly good at generating text that is like whatever they” saw “during training. But those large language model (LLM) could also perceive patterns, objects or algorithms that are non-existent or imperceptible to an observer i.e. synthetic data. Outcomes, then, are inaccurate, nonfactual, and implausible statements or objects rendering analysis futile.

These LLM induced hallucinations could be input-conflicting, context-conflicting, and fact-conflicting. (Dealing with Hallucinations in LLMs: A Deep Dive, Oct 22, 2023, Sourabh <https://blog.uptrain.ai/author/sourabhb5820a3c9b/>). In input-conflicting hallucination LLMs

generate content that contradicts or deviates from the source input provided by users. In context-conflicting hallucination LLMs generate content that conflicts with information previously generated by the model itself. Finally, LLMs in fact-conflicting hallucinations produce content that does not align with established, factual world knowledge (Zac Amos, 2025).

Biased and skewed data training could lead to hallucinative outcomes. But so are also false and wrongly labelled data, factual data mistakes, biases and contractions in data, insufficient programming to interpret data correctly, complex data models and vague user provided data context. All of these could only enhance the hallucinative impact. (Mukherjee et al, 2023).

Synthetic data induced miscarriages mentioned above could undermine large language model outcomes and open the door for hallucinations. ChatGPT did cite sources for information that are either not correct or do not exist, for example. A study conducted in the Cureus Journal of Medical Science showed that out of 178 total references cited by GPT-3, 69 returned an incorrect or non-existent Digital Object Identifier (DOI) (Bhattacharyya et al, 2023).

5.0 IMPACT

The ultimate impact of synthetic data hallucinative outcomes could spread over a wide front. There is first the spread of misinformation and fake news, the erosion of trust in AI-generated content, legal and ethical consequences, and decision-making aberrations (Understanding Hallucination in LLMs: Causes, Consequences, and Mitigation Strategies, Gianluca Centulani · Medium, Apr 11, 2024). The first category, spread of misinformation could have far reaching consequences as it could in healthcare, for example, lead to incorrect medical information and dire consequences. It could also in fiancé, lead to faulty investment or assessment decisions. And it could lead in education to faulty infrastructure and learning decisions.

LLM hallucinations can also erode trust in AI systems. Inaccurate or misleading information may lead to questions regarding the very reliability of the system. This can lead to loss of confidence in AI technology.

6.0 SUGGESTED HYPOTHESIS

Analysis cited above could lead to several hypotheses. The author considers the following worth of further future analysis.

- Data integrity could be undermined by synthetic data.
- Synthetic data could induce significant data analytics deviations.
- An approach to synthetic data viability and context relevance should be developed.
- A measure of hallucination may infect the analytical outcomes of AI analytics as an outcome of synthetic data infiltration.

Those hypotheses could provide a base for further future analysis.

7.0 SUMMARY AND CONCLUSIONS

"Hallucinations are common in large language models (LLMs) like ChatGPT and Google Gemini." ("AI 'hallucinations' can lead to catastrophic mistakes, but a new ...") They could stem from

synthetic data. These hallucinations pose significant challenges for the development and deployment of these technologies across a wide range of applications. Synthetic data produced during the process of LLM operations could account for these functional miscarriages.

Several hypotheses are suggested for future research. The first: Data integrity could be undermined by synthetic data. The second: A measure of hallucination may infect the analytical outcomes of AI analytics as an outcome of synthetic data infiltration.

REFERENCES

- “Cohut M, “Hallucinations: What explains these tricks of the mind? Medial news today, August 17, 2020
- Bhattacharyya, Mehul & Miller, Valerie & Bhattacharyya, Debjani & Miller, Larry (2023). "High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content." ("High Rates of Fabricated and Inaccurate References in ChatGPT-Generated ...") Cureus. 15. 10.7759/cureus 39238.
- Dealing with Hallucinations in LLMs: A Deep Dive, Oct 22, 2023, Sourabh <https://blog.uptrain.ai/author/sourabhb5820a3c9b/>
- Deng H “Exploring Synthetic Data for Artificial Intelligence and Autonomous Systems: A Primer, Geneva, Switzerland”: UNIDIR, (2023.)
- El Namaki, M.S.S. (2023). Data-Driven Strategic Thinking. In: Neo Strategic Management. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-37208-7_1923.
- Gartner 750175
- Gianluca Cantalanin, “Understanding Hallucination in LLMs: Causes, Consequences, and Mitigation Strategies,” · Medium, Apr 11, 2024)
- [Httpd://www.cnet.com/tech/hallucinations-why-ai-makes-stuff-up-and-whats-being-done-about-it/](http://www.cnet.com/tech/hallucinations-why-ai-makes-stuff-up-and-whats-being-done-about-it/)
- <https://builtin.com/artificial-intelligence/ai-hallucination>).
- [https://en.wikipedia.org/wiki/Hallucination_\(artificial_intelligence\)](https://en.wikipedia.org/wiki/Hallucination_(artificial_intelligence))
- Mukherjee, Anirban, Chang, Hannah H. (2023). Managing the Creative Frontier of Generative AI: The Novelty-Usefulness Trade-off” (California Management Review.) (“Managing the Creative Frontier of Generative AI: The Novelty ... - SSRN”)
- Singh S “Synthetic data vs real data: which is a better option? “, Labeler, Nov 15, 2022
- Understanding AI Hallucinations: Causes and Consequences, data Scientist, 17 Apr 2024).
- Zac Amos Z, “How Does Synthetic Data Impact AI Hallucinations? “Unite.AI, February 8, 2025,
- What is AI Hallucination? Examples, Causes & How to Spot them (technopedia.com).
- Zapier. “What are AI hallucinations Here’s how to encourage AI to stop hallucinating,” July 10, 2024,
- "What is synthetic data? - Definition from WhatIs.com". Search CIO. Retrieved 2022-09-08)