Junk food or a balanced diet? The picky politics of synthetic data for AI

Michael Strange, Malmö University

Data quantity and quality are often presented as the two defining ingredients driving AI innovation. To supply those two aspects, firms and governments have invested huge sums, overridden environmental regulations, ignored community concerns, and denied the rights of authors and artists to protect their works. Likewise, more and better data are presented as the solution for ameliorating two commonly experienced faults within large language models – bias, and so-called 'hallucinations'. Yet, in the rushed deployment of generative models, the internet – a core original source of training data – is said to have become polluted such that further AI training must find something else on which to feed.

Governments are actively legislating to allow greater data access, including in the USA with the DOGE project as well as in the European Union with a series of data sharing 'spaces', with varying degrees of controversy. The EU has chosen to build the first of these data spaces within healthcare, seen as an area where it is potentially easiest to argue the societal benefits of building larger systems for data sharing if it can be shown to help support, for example the development of new medications. The European Health Data Space (EHDS) is, however, nevertheless also a high-risk project where there remain significant legal restrictions to overcome as well as public concerns over privacy. There is also the risk that as data becomes shared at ever-greater distances from its point of collection, it becomes harder to ameliorate mistakes, false information, bias, as well as missing data points. Whilst ever-advanced bio-scans and wearables enable increasingly precise data, that quantity also opens new challenges, e.g. simple problems like faulty transmitters that mean missing values in pulse monitors, or erroneous data that suggests non-existent abnormalities leading to excessive and intrusive procedures.

For these reasons there is a now booming market led by USA-based firms focused on the production of synthetic data. The term 'synthetic' covers a broad spectrum from 'fully simulated' or 'slightly ameliorated', with the latter end mirroring long-standing practices of data control. For AI models to train, synthetic data can provide complete datasets such that they have greater utility, but also allow removal of datapoints that might otherwise lead to privacy breaches, unintended biases (e.g. such as considering how individuals move their cursor whilst submitting information on digital forms – a characteristic usually see as irrelevant but nevertheless potentially collected within training datasets and therefore made relevant to the algorithm), and support compliance within emerging regulations on AI.

Regulators are divided as to if synthetic data can be trusted with variation between legal jurisdictions as to whether AI developers can point to such data when applying for certification of their models – for example, and following broader patterns in AI regulation, the US is more open, but the EU remains opposed. Significant resources are being made to close the 'domain gap' – the level of divergence between real world data and synthetic data. For example, synthetic data is an important part of training for autonomous driving (Mullick et al, 2023), and the domain gap is closing even in some healthcare models (Pezoulas et al, 2024). Influential critical voices within AI, like Gary Marcus, warn though that synthetic data should only be trusted in closed environments and is unverifiable in more complex real-world environments at scale (Marcus, 2025). For this reason, synthetic data is

often analogised as facing many of the same problems as 'junk food' – that is, whilst offering convenience and a sense of immediate satisfaction, its level of artificiality exposes us to ingredients, many of which might not be well understood or easily identified, that may cause adverse effects. There is also concern that synthetic data only further replicates, or even worsens through making it harder to identify, biases and exclusions in AI models that risk societal harm (Lee et al, 2025).

Seeing synthetic data in the context of food nutrition and diet is worth further exploring. In direct correspondence, Marcus has clarified that synthetic data has a place within training models, but only alongside real-world human data that is licensed. By 'licensed', Marcus speaks to a much wider discussion on both transparency over training datasets, certified attention to data quality, as well as respect for intellectual property rights of creators whose work is used to feed the models. Within a nutrition paradigm, current models of data scraping picture a scenario in which AI models are allowed to eat everything and anything, with a total lack of discernment. The opposite scenario – one in which AI models were fed on an intentionally selected and transparently described balanced diet – leads to an alternative way of understanding and reconfiguring the politics of AI that I would like to develop in the workshop. Key aspects of that approach include how we think through regulation, societal participation, as well as current political-economic systems structuring the AI production line. In exploring this I will consider the case of AI in healthcare.

References:

Lee, F., Hajisharif, S., & Johnson, E. (2025). The ontological politics of synthetic data: Normalities, outliers, and intersectional hallucinations. *Big Data & Society*, *12*(2). https://doi.org/10.1177/20539517251318289.

Marcus, Gary (2025) *The "AI 2027" Scenario: How realistic is it?* Online: https://garymarcus.substack.com/p/the-ai-2027-scenario-how-realistic.

Mullick, K., Jain, H., Gupta, S., and Kale, A. (2023) 'Domain Adaptation of Synthetic Driving Datasets for Real-World Autonomous Driving', https://arxiv.org/abs/2302.04149.

Pezoulas, V., Zaridis, D., Mylona, E., Androutsos, C., Apostolidis, K., Tachos, N., and Fotiadis, D. (2024) 'Synthetic data generation methods in healthcare: A review on open-source tools and methods', *Computational and Structural Biotechnology Journal*, 23, pp. 2892-2910. https://doi.org/10.1016/j.csbj.2024.07.005.