# Cheap science, real harm: the cost of replacing human participation with synthetic data

**Abeba Birhane**
School of Computer Science and Statistics, Trinity College Dublin, Ireland
she\her, birhanea@tcd.ie

## Abstract

Driven by the goals of augmenting diversity, increasing speed, reducing cost, the use of synthetic data as a replacement for human participants is gaining traction in AI research and product development. This talk critically examines the claim that synthetic data can "augment diversity," arguing that this notion is empirically unsubstantiated, conceptually flawed, and epistemically harmful. While speed and cost-efficiency may be achievable, they often come at the expense of rigour, insight, and robust science. Drawing on research from dataset audits, model evaluations, Black feminist scholarship, and complexity science, I argue that replacing human participants with synthetic data risks producing both real-world and epistemic harms at worst and superficial knowledge and cheap science at best.

## 1 Representing complex behaviour in computational models

In order to examine what it means to represent human behaviour using synthetic data, it is necessary to first look at what it means to represent people, behaviour, and phenomena in data and models. Models are formal representations (often mathematical) of phenomena, processes, or aspects of the world that simulate behaviour over time or under various conditions. Although the goal of modelling is presumed to be prediction, especially with the rise of machine learning over the past couple of decades, models can also serve the purpose of, for example, understanding, describing, or explaining a given phenomenon. It is common to think of models as maps that capture the territory. While the map is never the territory, a good model is one that represents the territory with as high fidelity as possible.

Human behaviour is never simple, static, unambiguous, discrete, or governed by some generalisable rules, meaning a universal, "accurate", and exhaustive description, specification, and definition is not possible [28]. Representing complex phenomena as accurately as possible then depends on one's objective. Human behaviour, like other complex phenomena, is unfinalisable, inexhaustible, and not compressible into mathematical formalisms and description [13]. This means that we can never capture any given behaviour in its entirety with models. We might, at best, capture a snapshot of a system through our models. A given complex phenomenon can be modelled in an infinite number of ways depending on the objective. For example, "a portrait of a person, a store mannequin, and a pig" can all be a good model of the human body depending on whether one wants to remember a deceased grandparent, to buy clothes, or to study anatomy [19, 10]. Complex systems are also inherently open-ended, meaning that there is no uncontestable way of telling whether what we have included in a model is crucial or what we have omitted as irrelevant is indeed so [13, 12]. Thus, who is telling the story, in what context, and for what purpose all have dire implications. Afro-feminist scholars have, for example, emphasised that concrete, lived experiences are of primary importance to understanding, knowing, and telling a story with accuracy [14, 3]. Ahmed [3] uses the metaphor of a comfortable chair to explain how heteronormative and marginalised bodies experience reality differently. A person who can seamlessly sink into a chair that has possessed their bodily impression from repeated use can

Preprint.

easily and confidently speak of the chair's comfort and accommodating surface. It might, however, be difficult for this person to comprehend how uncomfortable and unaccommodating the same chair can be for others whose bodies are "shaped" differently and cannot sink into it [3]. Nonetheless, bell hooks [24] reminds us that marginality is not "something one wants to lose or give up or surrender as part of moving into the centre". But rather, a site of resistance, a position that "offers a radical perspective from which to see and create, to imagine alternatives, new worlds" [24].

Alas, this does not mean we should abandon mathematical models altogether but rather approach them with great humility . The better informed we are about the complex nature of the phenomena we are trying to model, the more modest our claims about the extent of our models' fidelity to the "ground truth," the more useful our models are in the real world.

## 2 Synthetic data: stereotypes compressed

The use of synthetic data as a human proxy is gaining momentum in numerous domains. In medical research and clinical settings, synthetic data is used as a proxy for real clinical data in predictive modelling [17, 7, 30]. In social and behavioural sciences research, synthetic data are considered proxies for real humans where LLMs are used as stand-ins for human behaviour and decision-making [40, 2, 4]. In psychometrics, LLM-based agents are proposed as quantifiable, controllable, and accessible alternatives that "overcome the constraints of human subject studies" for social science inquiries [25, 32, 42]. In political science, [21] have proposed using LLMs as substitute for human experts in annotating political text. In AI safety and "value alignment" research, LLMs are increasingly used in place of human raters to "extract" human values and preferences that LLMs ought to be aligned with [39, 29, 15]. In the design space, Morris and Brubaker [37] have proposed building a "simulacra that can produce believable human behaviours, including capabilities such as memory and planning" using LLMs. In the humanitarian sector, the recent proposal from the United Nations University to use LLM-based agents to simulate refugees marks a prime example such trend [5]. The proposed agents are "designed to authentically represent "refugee living in Chad and eastern Sudan" for the purpose of enabling "rapid data collection in dangerous or time-sensitive situations, overcome language barriers and interpreter bias common in conventional surveys" [5].

Despite such expectations placed on LLMs to substitute humans in a plethora of fast growing use cases, a recurring theme within a growing body of work examining LLMs shows that these models tend to have a homogenising effect, reduce cultural nuances, flatten out complexity, and perpetuate Western values, closely resembling those observed in Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies [6, 11, 27]. Comparing 2,200 human written college admissions essays with those written by GPT-4 showed the homogenizing effect of LLMs on creative diversity at the individual and collective level, reducing creative diversity across groups of people [36]. Similarly, according to [16], generative AI–enabled stories are more similar to each other than stories by humans alone. LLMs' responses on cognitive psychological tasks most resemble those of people from WEIRD societies while similarity rapidly declines as we move away from these populations [6]; LLMs perpetuate linguistic discrimination toward speakers of non-"standard" English language varieties and perpetuate covert racism based on dialects [18, 22]; in response to culturally appropriate nuances of emotion, LLMs reflect Anglo-centric, Western norms, even when responding to prompts in other languages [20]; they consistently portray African, Asian, and Hispanic Americans as more homogeneous than White Americans, flattening out descriptions of racial minority groups with a narrow range [33]; and in response to queries regarding human rights, LLMs are shown to avoid fully answering with a singular yes/no output even though the models were fed clear prompts that elicit yes or no responses, with the greatest disparities involving conflict-associated identities such as Palestinians, Kashmiris, and Russians [26].

The multitude of issue emerging from LLMs are often attributed to training data, proxies for the real world. Most popular and state-of-the-art models including the GPT variants, DeepSeek, Claude, Gemini, Grok, and Llama operate under secrecy when it comes to training data. Critical information that has significant implication on understanding the "representativeness" of the data including what is in the training data, the sources of the data, and filtering policies used to remove toxic and low quality content remain a secret. However, audits on the open-source replicates of propitiatory datasets have shown that major public datasets suffer from numerous issues including low quality [31, 41]; duplicates [43, 9]; inclusion of problematic content such as NSFW images, hate-speech, and sexually explicit content [9, 23, 8, 35] as well as representational concerns. For example, over half of the

datasets used for performance benchmarking across more than 26,000 research papers came from just 12 elite institutions and tech companies in the US, Germany, and China [38], while an audit of nearly 4000 widely used publicly available multimodal datasets found that the cultural, geographical, and ideological representations overwhelmingly homogeneous, particularly concentrated in the US, China and Western Europe with little to no representation of Africa and Southern America [34].

Given such trends in the composition, quality, and homogeneity of training data that then often impacts model output, claims of synthetic data as a tool to advance diversity is empirically unsustainable, logically incoherent, and practically likely to be harmful to the marginalised groups such data is supposed to represent. Previous work has pointed out that attempts to substitute human participation with synthetic data as "diversity washing" [44], standing in an ultimate "conflict with foundational values of work with human participants: representation, inclusion, and understanding" [1], and empirically shown to struggle to accurately reflect real-world conditions [45]. Time and money might be gained from utilizing synthetic data as stand-in for human participation — at the cost of rigour, validity, and groundlessness — normalising cheap science.

Web sourced data of any modality (text, image, audio, video) not only flattens human complexity, also reduces it to homogenous Western values and perspectives. LLMs encode and exacerbate the various problems stemming from training data. Using LLMs as a substitute for human participation reduces human participation to a statistical average of underlying data distribution. Ultimately, to propose that synthetic data – a model of a model – that caricatures a homogeneous, oppressive, discriminatory, and negative stereotypical perspective can be used to advance "diversity" rests on intellectual dishonesty, exposes lack of familiarity with experiences and scholarship from the margins, and marks double erasure that is likely contribute to harms towards such group due to practical applications of LLMs.

## References


[1] W. Agnew, A. S. Bergman, J. Chien, M. Díaz, S. El-Sayed, J. Pittman, S. Mohamed, and K. R. McKee. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2024.

[2] G. V. Aher, R. I. Arriaga, and A. T. Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR, 23–29 Jul 2023.

[3] S. Ahmed. A phenomenology of whiteness. *Feminist theory*, 8(2):149–168, 2007.

[4] D. Albert and S. Billinger. Reproducing and extending experiments in behavioral strategy with large language models. *arXiv preprint arXiv:2410.06932*, 2024.

[5] E. Albrecht. Does the united nations need agents? testing the role of ai agent generated personas in humanitarian action, 2025.

[6] M. Atari, M. J. Xue, P. S. Park, D. Blasi, and J. Henrich. Which humans? 2023.

[7] Z. Azizi, C. Zheng, L. Mosquera, L. Pilote, and K. El Emam. Can synthetic data be a proxy for real clinical trial data? a validation study. *BMJ open*, 11(4):e043497, 2021.

[8] A. Birhane, S. Han, V. Boddeti, S. Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. *Advances in neural information processing systems*, 36:21268–21284, 2023.

[9] A. Birhane, V. U. Prabhu, and E. Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[10] P. Blanchard, R. L. Devaney, and G. R. Hall. *Differential equations*. Cengage Learning, 2012.

[11] Y. Cao, L. Zhou, S. Lee, L. Cabello, M. Chen, and D. Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. *arXiv preprint arXiv:2303.17466*, 2023.

[12] P. Cilliers. *Complexity and postmodernism: Understanding complex systems*. routledge, 2002.

[13] P. Cilliers. Why we cannot know complex things completely. *Critical complexity: Collected essays*, 6:97–105, 2016.

[14] P. H. Collins. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge, 2002.

[15] G. Cui, L. Yuan, N. Ding, G. Yao, B. He, W. Zhu, Y. Ni, G. Xie, R. Xie, Y. Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. In *International Conference on Machine Learning*, pages 9722–9744. PMLR, 2024.

[16] A. R. Doshi and O. P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024.

[17] J.-N. Eckardt, W. Hahn, C. Röllig, S. Stasik, U. Platzbecker, C. Müller-Tidow, H. Serve, C. D. Baldus, C. Schliemann, K. Schäfer-Eckart, et al. Mimicking clinical trials with synthetic acute myeloid leukemia patients using generative artificial intelligence. *NPJ digital medicine*, 7(1):76, 2024.

[18] E. Fleisig, G. Smith, M. Bossi, I. Rustagi, X. Yin, and D. Klein. Linguistic bias in chatgpt: Language models reinforce dialect discrimination. *arXiv preprint arXiv:2406.08818*, 2024.

[19] L. Gyllingberg, A. Birhane, and D. J. Sumpter. The lost art of mathematical modelling. *Mathematical biosciences*, 362:109033, 2023.

[20] S. Havaldar, S. Rai, B. Singhal, L. Liu, S. C. Guntuku, and L. Ungar. Multilingual language models are not multicultural: A case study in emotion. *arXiv preprint arXiv:2307.01370*, 2023.

[21] M. Heseltine and B. C. von Hohenberg. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239, 2024.

[22] V. Hofmann, P. R. Kalluri, D. Jurafsky, and S. King. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, 2024.

[23] R. Hong, W. Agnew, T. Kohno, and J. Morgenstern. Who's in and who's out? a case study of multimodal clip-filtering in datacomp. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–17, 2024.

[24] b. hooks. Choosing the margin as a space of radical openness. *Framework: The Journal of Cinema and Media*, (36):15–23, 1989.

[25] M. Huang, X. Zhang, C. Soto, and J. Evans. Designing llm-agents with personalities: A psychometric approach, 2024.

[26] R. Javed, J. Kay, D. Yanni, A. Zaini, A. Sheikh, M. Rauh, R. Comanescu, I. Gabriel, and L. Weidinger. Do llms exhibit demographic parity in responses to queries about human rights? *arXiv preprint arXiv:2502.19463*, 2025.

[27] R. L. Johnson, G. Pistilli, N. Menédez-González, L. D. D. Duran, E. Panai, J. Kalpokiene, and D. J. Bertulfo. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*, 2022.

[28] A. Juarrero. Dynamics in action: Intentional behavior as a complex system. *Emergence*, 2(2):24–57, 2000.

[29] D. Kim, K. Lee, J. Shin, and J. Kim. Spread preference annotation: Direct preference judgment for efficient llm alignment, 2025.

[30] T. Kokosi and K. Harron. Synthetic data in medical research. *BMJ medicine*, 1(1):e000167, 2022.

[31] J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramani, A. Sokolov, C. Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

[32] L. La Cava and A. Tagarelli. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1355–1363, 2025.

[33] M. H. Lee, J. M. Montgomery, and C. K. Lai. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1321–1340, 2024.

[34] S. Longpre, N. Singh, M. Cherep, K. Tiwary, J. Materzynska, W. Brannon, R. Mahari, N. Obeng-Marnu, M. Dey, M. Hamdy, et al. Bridging the data provenance gap across text, speech and video. *arXiv preprint arXiv:2412.17847*, 2024.

[35] A. Luccioni and J. Viviano. What's in the box? an analysis of undesirable content in the common crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, 2021.

[36] K. Moon, A. Green, and K. Kushlev. Homogenizing effect of large language model (llm) on creative diversity: An empirical comparison, 2024.

[37] M. R. Morris and J. R. Brubaker. Generative ghosts: Anticipating benefits and risks of ai afterlives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2025.

[38] Mozilla Foundation. Internet health report 2022: Who has power over ai?, 2022.

[39] I. Padhi, K. Natesan Ramamurthy, P. Sattigeri, M. Nagireddy, P. Dognin, and K. R. Varshney. Value alignment from unstructured text. In F. Dernoncourt, D. Preoţiuc-Pietro, and A. Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1083–1095, Miami, Florida, US, Nov. 2024. Association for Computational Linguistics.

[40] J. Tang, H. Gao, X. Pan, L. Wang, H. Tan, D. Gao, Y. Chen, X. Chen, Y. Lin, Y. Li, et al. Gensim: A general social simulation platform with large language model based agents. *CoRR*, 2024.

[41] R. Van Noord, M. Esplà-Gomis, M. Chichirău, G. Ramírez-Sánchez, and A. Toral. Quality beyond a glance: Revealing large quality differences between web-crawled parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1824–1838, 2025.

[42] Y. Wang, J. Zhao, D. S. Ones, L. He, and X. Xu. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15(1):519, 2025.

[43] R. Webster, J. Rabin, L. Simon, and F. Jurie. On the de-duplication of laion-2b. *arXiv preprint arXiv:2303.12733*, 2023.

[44] C. D. Whitney and J. Norman. Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1733–1744, 2024.

[45] X. Zhou, Z. Su, T. Eisape, H. Kim, and M. Sap. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714, 2024.