Can LLM-generated synthetic data be *representative*?

The use of synthetic data poses fundamental questions and challenges to representations of perspectives. In this position statement, we propose the framework of *epistemic injustice* (Fricker, 2007) to add nuances to these questions. The guiding question for this work is: *Can synthetic data be representative?* To narrow the scope, we will focus on Large Language Models (LLMs) generated synthetic data about people, individuals, as well as groups.

The motivation for this is to invite more interdisciplinary conversations in the use and development of data-driven technologies. Here specifically, we suggest directing our attention to epistemic injustice (EI) as a lens for understanding how synthetic data practices may perpetuate or mitigate injustices in knowledge production and representation. EI occurs when individuals or groups are wronged in their capacity as *knowers* – harmed or disadvantaged through misjudgments of their ability to possess truth and being seen as less trustworthy than they actually are. This can happen when speakers are given less credibility due to mistaken judgments about their knowledge based on negative identity stereotypes, or when gaps in collective interpretive resources put someone at an unfair disadvantage in making sense of their social experiences. Central to this framework is the recognition that our epistemic practices are especially situated, embedded within power relations between participants. A critical aspect of EI is that when listeners base their credibility judgments on identity stereotypes of the speaker, they become insensitive to counter-evidence provided in the actual testimony. Fundamentally, EI results in the exclusion of groups or individuals from contributing to the shared pool of knowledge, thereby impoverishing our collective understanding and perpetuating systemic disadvantages.

To assess the potential for increasing or reducing EI of *representation* in synthetic data for and from natural language processing applications, we first disentangle what we mean by *representation*. In this work, we focus on representation in the sense of whether identities and voices[1] are accounted for in the data. Within this scope, we distinguish between two scenarios: (a) Representation **in data**, where parts of identities are properly accounted for (e.g, augmentation of data for a named entity recognition system, where we add all theoretically possible first and last name combinations to ensure proper *representation* of the system's outputs and performance (see e.g., Lassen et. al, 2023) (b) Representation **by data**, where identities themselves and their expression are being modeled (e.g., generating synthetic survey responses from different demographic groups, opinion generation based on political affiliations, or persona prompting with LLMs to simulate specific cultural perspectives (Hu and Collier, 2024; Giorgi et al., 2024; Hou et al., 2025, inter alia)).

While their potential broader impacts should be inspected for both scenarios, we focus on (b) in the following. Representation **by data** poses more pressing questions about EI because these approaches directly model people's identities and voices, yet their impacts remain relatively under-researched despite growing popularity. Moreover, risks are growing more salient with increasing amounts of AI-generated content populating the web (Brooks et al., 2024; Spennemann, 2025).

Representing people's identities and voices with synthetic, LLM-generated data, especially when it includes sociodemographic proxies, frequently gets framed as a way to

[1] We refer to people's voices as expressions of opinions, perspectives, and sets of values which *should* be representative of a certain group and its individual members. By silenced voices, we mean those voices belonging to people who are excluded or marginalized and not represented in the pool of knowledge, constituting EI.

reproduce the subjectivity that humans exhibit (Argyle et al., 2023; Hämäläinen et al., 2023, inter alia), and represent (sub)populations and real-life narratives (Moon et al., 2024). The common implicit assumption appears to be that, if prompted and adapted properly, LLMs can be used as a proxy that will reproduce people's voices appropriately. If this is true, it would allow for the amplification, or in the most extreme case, introduction of underrepresented or silenced voices. Moreover, it is implicitly assumed that, by making LLMs reproduce human tendencies *sufficiently* well (at least for some groups), this will automatically lead to *representative* synthetic data. Through the lens of EI, synthetic data can be seen as a way of including diverse perspectives in the collective pool of knowledge by potentially amplifying voices that have been excluded or given less credibility in earlier data collection processes.

However, these assumptions raise multiple questions. The first regards the adaptability of this data, both synchronically and diachronically. When synthetic data related to a certain socio-demographic group is generated, it necessarily only captures a snapshot at a given point in time that aims to portray the voices of this specific group of people. But does it reflect their voices *sufficiently* (c.f. Giorgi et al., 2024; Wang et al., 2025)? And if there is a shift in these people's opinions, how will this be captured? Furthermore, persona prompting approaches typically rely on sociodemographic markers and identity categories – precisely the kind of identity stereotypes that Fricker warns can lead to prejudicial credibility assessments. Just as EI occurs when listeners rely on identity stereotypes and fail to properly evaluate the actual content of testimony, persona-prompted synthetic data generation risks relying on assumptions about how certain groups think or speak, potentially suppressing the actual perspectives and voices within the given communities.

In sum: How do we ensure our *representation* of voices aligns with actual voices? How do we ensure it does not misportray or, maybe unwillingly, silence actual voices by assuming them or how representative subsets are, and neglecting their potential to change? To answer these questions, one would need to gather real-world silenced voices to compare the synthetic data with, which, if collected anyway, poses the question of why we can't simply use *real* empirical data to solve the issue of representativeness. Finally, what power are we as researchers exercising over people, groups, and their voices with our definitions of representativeness of them, especially via sociodemographic proxies, and with the ways we operationalize the measurement of representativeness?

As these questions illustrate, there is a need to critically engage with our assumptions about representation and claims about what our methods can offer. Given the lack of explicit mentioning and engagement with underlying assumptions – not only about what we mean by *representativeness*, but also in the role of researchers deciding which voices are considered in the data and how we assess whether they are considered appropriately – we invite researchers to examine their synthetic data practices. This includes, but is not limited to asking: Who is speaking? Who is being spoken for? Whose voice is forgotten, misrepresented, or silenced? And, finally, what are the consequences for those whose identities and voices are or are not approximated?

These are difficult and inherently ethical questions that cannot be resolved through technical metrics or model performance evaluations alone. The framework of EI reveals that questions of representation are fundamentally about power, knowledge, and whose voices are deemed credible. We therefore call for careful critical engagement with the assumptions underlying synthetic data practices, moving beyond treating representation as mere statistical presence toward deeper considerations of whose knowledge counts and on whose terms it is included in our collective knowledge.

Bibliography

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3), 337–351. doi:10.1017/pan.2023.2

Brooks, C., Eggert, S., & Peskoff, D. (2024). *The rise of AI-generated content in Wikipedia*. In L. Lucie-Aimée, A. Fan, T. Gwadabe, I. Johnson, F. Petroni, & D. van Strien (Eds.), *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia* (pp. 67–79). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wikinlp-1.12

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Giorgi, S., Liu, T., Aich, A., Isman, K. J., Sherman, G., Fried, Z., Sedoc, J., Ungar, L., & Curtis, B. (2024). *Modeling human subjectivity in LLMs using explicit and implicit human factors in personas*. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 7174–7188). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.420

Hämäläinen, P., Tavast, M., & Kunnari, A. (2023). *Evaluating large language models in generating synthetic HCI research data: A case study*. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Article 433, 19 pages). Association for Computing Machinery. https://doi.org/10.1145/3544548.3580688

Hou, Y., Daume, H. III., & Rudinger, R. (2025). Language models predict empathy gaps between social in-groups and out-groups. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 12288–12304). Association for Computational Linguistics. https://aclanthology.org/2025.naacl-long.611/

Hu, T., & Collier, N. (2024). Quantifying the persona effect in LLM simulations. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10289–10307). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.554

Lassen, I. M. S., Almasi, M., Enevoldsen, K., & Kristensen-McLachlan, R. D. (2023). *Detecting intersectionality in NER models: A data-driven approach*. In S. Degaetano-Ortlieb, A. Kazantseva, N. Reiter, & S. Szpakowicz (Eds.), *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (pp. 116–127). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.latechclfl-1.13

Moon, S., Abdulhai, M., Kang, M., Suh, J., Soedarmadji, W., Behar, E. K., & Chan, D. (2024). *Virtual personas for language models via an anthology of backstories*. In Y. Al-Onaizan, M.

Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 19864–19897). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.1110

Spennemann, D. H. R. (2025). *Delving into: The quantification of AI-generated content on the internet (synthetic data)*. arXiv. https://arxiv.org/abs/2504.08755

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence, 7*(3), 400–411. https://doi.org/10.1038/s42256-025-00986-z