Issues in Measuring the Fairness of Social Representation in Synthetic (Speech) Data

Arjun Subramonian*[1], Brooklyn Sheppard*[2], Levent Sagun*[1]
[1]Meta FAIR, [2]University of Calgary
{arjunsub, leventsagun}@meta.com brooklyn.sheppard1@ucalgary.ca

As data-hungry large generative models have proliferated in recent years, AI practitioners have used synthetic data to pretrain [7] and align [19] these models. Proponents of synthetic data argue that it can provide an additional high-quality learning signal for models and promote privacy [11]. However, concerns have been raised about how training on synthetic data can contribute to model collapse [5] and how synthetic data may not fairly represent marginalized social groups [20]. In this provocation, we discuss three main issues in measuring the fairness of social representation in synthetic data: (1) tensions between different conceptualizations of representation, (2) the association of synthetic data subjects with social categories, and (3) aligning measurements with social groups' desires and interests. Throughout this article, we ground our discussion with examples from speech processing for concreteness and in tenets of Intersectionality [3]. We provide intersectional guiding questions for practitioners looking to measure social representation in both synthetic and non-synthetic data.

Conceptualizations of Representation: There can exist tensions between different conceptualizations of fair representation [2]. [2, 10] describe two predominant conceptualizations in AI: (C1) proportions of social groups in data are consistent with a reference (e.g., "ground-truth") distribution of groups, and (C2) proportions of different social groups are equal. These conceptualizations are often informed by different aims, e.g., (A1) faithfully evaluating models with respect to a reference data distribution vs. (A2) faithfully evaluating performance disparities of models across groups [2]. Taking an example of representation in speech AI, recent work has demonstrated that there is a significant lack of investigation into how well these models work for gender diverse individuals [14]. That is, almost all speech datasets assume binary notions of gender, and thus gender bias evaluations are limited to performance gaps across the categories of "male" and "female." One exception to this is the Casual Conversations V2 dataset, which includes recordings from 80 speakers outside the cis binary - representing 1.44% of all speakers in the dataset [13]. According to (C1), this dataset could be seen as "over representing" this community by almost 1% in terms of their "true" distribution in the world [9]. In the context of (C2), however, this represents a significant lack of data for this community to meaningfully compare performance disparities across gender identities.

Across speech and other modalities, when measuring the fairness of social representation in synthetic data, (C1) entails computing the divergence of the synthetic data group distribution from a chosen "ground-truth" group distribution. However, ML practitioners may tend to choose "ground-truth" distributions based on hegemonic social contexts within the field, without reflecting on the power dynamics that give rise to this choice [12]. Moreover, the "ground-truth" distribution is often estimated using government-collected data (e.g., the U.S. census), which itself suffers from systemic data quality issues such as the undercounting of minoritized racial groups [18] and the erasure of entire social categories (e.g., non-binary people). In contrast to (C1), (C2) entails computing the divergence of the data group distribution from a uniform group distribution. However, both (C1) and (C2) require foregrounding certain social axes (e.g., race, gender over caste) and along these axes, categorizing individuals into a finite number of discrete groups. Such decisions are often informed by a U.S. context, and can be incompatible with fluid and intersectional nature of people's identities.

Association with Social Categories: Both (C1) and (C2) involve associating each synthetic data instance with a group membership label. In the context of speech, popular techniques for obtaining these labels include: (T1) using an auxiliary classifier to infer the social group membership of a speaker in a synthetic audio clip, and (T2) conditioning a generative model on a speaker's social group to produce a synthetic audio clip. (T1) suffers from numerous measurement validity issues

*equal contribution

(e.g., unquantifiable error rates, groups with poor construct validity) and ethical issues (e.g., unevenly distributed errors, support of surveillance infrastructure), as has been shown in other domains [6, 16]. On the other hand, (T2) can flatten and stereotype social groups. The need for vast amounts of synthetic data and consequently the usage of efficient methods to obtain group membership labels to measure representation, despite their pitfalls, are driven by "scale thinking" [8]. It is important for the AI community to explore methods to measure representation that need not occur at scale.

Alignment with Desired Outcomes: While not a design fix [15], participatory design is crucial to guard against the harms of predictive design and align synthetic data generation with the desires and interests of marginalized groups. Given that high quality and diverse speech data is difficult to collect, this modality might seem like an ideal one for data augmentation via synthetic speech. For marginalised communities, however, this approach risks the researcher shifting from a descriptive analysis to a predictive one. Representation in a descriptive analysis, such as evaluating model performance across demographic groups, may be a desirable form of representation to understand the social inequities represented in these systems. In contrast, predictive analyses, such as predicting a set of features associated with a particular group (or vice versa), introduces a significant risk of harm and surveillance [4]. In the case of synthetic speech data generation, not only does this fall in the category of predictive analysis, but it also risks homogenizing and stereotyping the attributes of entire demographics of speakers, as we have seen with generation in other domains [1, 17].

Participatory design inherently sparks contextual discussions about political goals, power, and inclusion. For example, the downstream use cases and sociopolitical impacts of models trained and evaluated on synthetic data will affect groups' conceptualizations of ideal representation. These conceptualization may diverge greatly from (C1) and (C2): a group may view representation through a reparative lens, considering synthetic data to be representative if it centers or exclusively captures the group's most vulnerable communities. Within groups, communities may negotiate distinct criteria for fair representation, developing processes for choosing representative individuals and data, as well navigating their relationality. Importantly, measurements of fair representation often sidestep whether marginalized groups would like to be included *at all* in synthetic data, for example, if they feel that such data would fundamentally misrepresent them or bolster systems that are designed to harm or surveil them [2]. Marginalized groups should have the power to refuse systems that generate or rely on synthetic representation.

Reflections and guiding questions: To help synthetic data practitioners engage in reflexivity when measuring the fairness of social representation, we offer the following guiding questions:

- Is your conceptualization of fair representation more aligned with (C1) or (C2)? What normative factors have informed this conceptualization?
- How do your social context and power dynamics affect how you choose a "ground-truth" group distribution? How you choose which axes and groups to consider?
- How does your estimate of the "ground-truth" distribution reflect and further entrench social inequality?
- What methodological and ethical issues arise from your approach to associating synthetic data instances with group membership labels?
- What harms arise from scale thinking when measuring social representation in data? How may we avoid them going forward?
- Where do your design and development methodologies fall along the continuum from descriptive to prescriptive? What elements of participatory design are you integrating?
- How may your work harm, homogenize, or stereotype marginalized social groups?
- How are you accommodating reparative conceptualizations of representation? How are you accommodating refusal?

With a grounding in speech processing, we identify three major issues in measuring the fairness of social representation in synthetic data: (1) differing conceptualizations of representation, (2) associating synthetic data instances with social groups, and (3) aligning measurements with groups' desires and interests. We advocate for synthetic data practitioners to consider our reflexive guiding questions and interrogate their work through an intersectional lens.

References

[1] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023.

[2] Kyla Chasalow and Karen Levy. Representativeness in statistics, politics, and machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 77–89, 2021.

[3] Patricia Hill Collins and Sirma Bilge. *Intersectionality*. John Wiley & Sons, 2020.

[4] Marco Crocco, Marco Cristani, Andrea Trucco, and Vittorio Murino. Audio surveillance: A systematic review. *ACM Comput. Surv.*, 48(4), February 2016.

[5] Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. *International Conference on Learning Representations*, 2024.

[6] Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes. Stop! in the name of flaws: Disentangling personal names and sociodemographic attributes in NLP. In Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 323–337, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[7] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[8] Alex Hanna and Tina M Park. Against scale: Provocations and resistances to scale thinking. *arXiv preprint arXiv:2010.08850*, 2020.

[9] Jody L Herman, Andrew R Flores, and Kathryn K O'Neill. How many adults and youth identify as transgender in the united states? 2022.

[10] Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models, 2024.

[11] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data. *arXiv preprint arXiv:2404.07503*, 2024.

[12] Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 496–511, New York, NY, USA, 2023. Association for Computing Machinery.

[13] Bilal Porgali, Vítor Albiero, Jordan Ryda, Cristian Canton Ferrer, and Caner Hazirbas. The casual conversations v2 dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10–17, 2023.

[14] Ariadna Sanchez, Alice Ross, and Nina Markl. Beyond the binary: Limitations and possibilities of gender-related speech technology research. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 526–532. IEEE, 2024.

[15] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA, 2022. Association for Computing Machinery.

[16] Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*, 2024.

[17] Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models should not replace human participants because they can misportray and flatten identity groups. *arXiv preprint arXiv:2402.01908*, 2024.

[18] Hansi Lo Wang. The 2020 census had big undercounts of black people, latinos and native americans, 2022.

[19] Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T. Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. Codeclm: Aligning language models with tailored synthetic data. In *NAACL-HLT*, 2024.

[20] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. Fairness feedback loops: Training on synthetic data amplifies bias. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 2113–2147, New York, NY, USA, 2024. Association for Computing Machinery.