Fake It Till You Make It: Synthetic Data and Algorithmic Bias

SOOK-LIN TOH
JIWON PARK
University of Southern California, USA

This article interrogates the claim that synthetic data is a risk-free and ethical solution to algorithmic bias. Synthetic data refers to artificial intelligence (AI)-generated datasets that substitute real-life data to train machine learning (ML) models. We examine how bias is supposedly corrected by synthetic data through the three key methods: (1) rebalancing the dataset, (2) the de- and reconstruction of data, and (3) classification. We argue that proponents of synthetic data presume that algorithmic bias resides in the lack of diversity in training data—rather than in sociopolitical inequalities—and mobilize prescriptions for fairness, bias correction, and equitable representations of gender and race. Practices in synthetic data assume that an unbiased, neutral state exists and can be achieved artificially. Analyzing this technical "solution" shows how mainstream ML discourse understands bias as a sociotechnical error. Building on existing literature on algorithmic bias, this article shows how synthetic data, despite promises to the contrary, actually exacerbates and creates new forms of inequality.

Keywords: synthetic data, algorithmic bias, generative AI

Synthetic data, or artificial intelligence (AI)-generated datasets, have emerged as a supposedly "risk-free" solution to concerns about data privacy and computational bias. Machine learning (ML) models like generative adversarial networks (GANs) use a sample set of real-life data to produce synthetic datasets that "reproduce the salient attributes or overall statistical distribution of real-world data" (Jacobsen, 2023, p. 2). Large companies like IBM claim that synthetic data can be procured "on-demand" and customized at an unlimited scale to adjust for bias and data augmentation (Martineau, 2023, para. 4). The ethical promise is usually focused on privacy, arguing that synthetic data can mask and prevent the exposure of sensitive personal information. This article focuses on its often secondary promise of fixing social biases by correcting the underrepresentation of minority groups.

Google Brain founder and Stanford professor Andrew Ng (2022) has championed synthetic data, arguing that to optimize AI performance, system builders can "systematically engineer" (para. 1) iterations of data, rather than tweaking ML models. This approach sees the dataset as the key source of system bias. Although researchers locate sites of bias across data, models, and evaluation of models, *dataset* bias often "attracts much more critical attention than the methods and the underlying algorithms themselves" (Dobson, 2023, p. 156). Synthetic data is a particularly powerful example of this data-centric approach for its claims that synthetically generated, *debiased* datasets offer technical corrections to the human biases of existing, real-world datasets.

Copyright © 2025 (Sook-Lin Toh, stoh@usc.edu, and Jiwon Park, parkjiwo@usc.edu). Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at https://ijoc.org.

This article examines how synthetic data is readily adopted as the technical solution to the error of bias. For proponents of synthetic data, bias is understood as human errors in data collection and processing—the lack of diverse, proportional datasets and the costly, inconsistent human labor involved in labeling. This notion of error narrows down systemic and complex sociopolitical configurations into a matter of representation within a dataset. Bias, and how it should be corrected, is assessed by the fairness of a model's output. This definition of error produces and valorizes utopic, performance-oriented imaginaries of a mythical, "balanced" dataset. In response, synthetic data is quickly emerging as an AI hotbed and a site of development that offers generative insight into how tech companies envision error and compute mitigations of bias. Black feminist science and technology studies (STS) scholars Ruha Benjamin (2019) and Meredith Broussard (2023) maintain that bias is "more than a glitch," arguing that racism and sexism are encoded in, and fundamental to, ML language and infrastructure. Building on extant literature on algorithmic bias, this article outlines three key elements involved in correcting bias in synthetic data: (1) rebalancing the dataset, (2) deconstructing real-life data to reconstruct artificial counterparts, and (3) the classification systems that these processes rely on.

Rebalancing the Dataset

The ability to rebalance a dataset is synthetic data's main claim to redressing bias. The rationale is that the biased outcomes of an AI model are largely due to the underrepresentation of minority groups—such as darker skin tones in facial recognition technology (Buolamwini & Gebru, 2018)—hence developers simply need to increase their proportions in the training data. These methods are not novel to synthetic data; for example, naive rebalancing is commonly used to duplicate the minority class to increase its proportion in the dataset (Aysha, 2023). Synthetic approaches, though, propose to generate *new* data, not just repeat existing data.

Researchers at the synthetic data platform MOSTLY AI (Tiwald, Ebert, & Soukup, 2021) describe how to generate tabular data to correct "gender" bias in an adult census dataset. This is done by selecting gender as a protected attribute and training the generative model with a parity fairness constraint. The constraint aims to eliminate any correlation between gender and income in the new dataset. Put differently, the synthesized dataset controls the correlation between gender and income, while the statistical relationships between the other attributes (e.g., age, education level, marital status) replicate that of the original dataset.[1]

This approach attempts to fix bias through equalized representation in training data, revealing an assumption that bias is defined by the level representation of all social groups and the presence of unfavorable correlations. Google and data ethics researchers suggest that concerns about bias and fairness are often reductively framed as a purely technical issue of not enough data existing to be solved with additional, diversified data (Denton, Hanna, Amironesei, Smart, & Nicole, 2021, p. 9). Diversity—a capacious and murky term in itself—is codified and enacted as synthetic variability, which is achieved by creating high numbers of possibilities and probability distributions.

[1] This method also supposedly accounted for hidden proxy attributes: Developers added an artificial feature labeled "proxy," strongly correlated with "gender," and demonstrated that the synthetic correlation between income, gender *and* the proxy was lower than that of the original dataset. They stated the limitation that they could train it with "race" as a protected attribute but were less successful with multiple protected attributes.

The narrative of rebalancing training datasets presumes neutrality as a possibility, understood primarily through the logics of statistical distribution. If developers eradicate unwanted correlations, they can consider the dataset unbiased. For MOSTLY AI, the assumption is that the source of bias was a correlation between gender and income, rather than education or marital status, for example. Hence, this idea of an unbiased state is actually based on particular theories of which relationships should be neutral.

Moreover, proponents of using synthetic data to rebalance datasets tend not to focus on the real, systemic conditions driving dataset imbalances, such as the very real gender pay gap causing fewer women to be in the "high income" bracket. Questions of how these models are used and the potential impact of their deployments on marginalized groups are overshadowed by a focus on the fairness of the model's outputs. Even if datasets can be rebalanced synthetically, the models are often being used in contexts of the same structural inequalities that produced the unbalanced, biased original datasets in the first place. It is unclear what types of futures synthetic data hopes to produce by artificially fixing these AI models. Rebalancing is then a fantasy.

Deconstructing and Reconstructing

The generation of synthetic data relies on the replication of identified attributes.[2] To explain how synthetic data imagines real life, we turn to facial data, which can be generated from original images or from scratch.

Notably, researchers at Microsoft Research Asia (Bao, Chen, Wen, Li, & Hua, 2018) proposed a GAN-based method of face synthesis that could disentangle the identity and attributes (e.g., pose, expression, illumination) of a face image and recombine them to generate a synthetic face image. In other words, a real face is disassembled into isolated features that are then randomly reassembled to produce new faces that do not belong to any one person.

More recently, Microsoft developed the Face Synthetics dataset for face-related computer vision, containing 100,000 synthetic faces with "unprecedented realism and diversity" (Wood et al., 2021, para. 1). Each image was rendered by mixing and matching a random identity, expression, texture, hair, and clothing—assets authored by 3D artists and visual designers at Microsoft—atop a template facial model. Rather than beginning with a sample, real-life face, facial assets are created through Photoshop and computer-generated imagery (CGI).

In both these examples, the face is understood as a collection of fragmented, separable assets. Persons, and their faces, are used as elements to improve a model's performance. The logics of deconstructing the face are underpinned by a desire to extract this instrumentality, disassembling faces into a set of features legible to the ML model. In this disassemblage, the identity and context behind the subject is made irrelevant. Kate Crawford (2022) describes a shift from image to infrastructure, in which the meaning and context of an image is "presumed to be erased at the moment it becomes part of an aggregate mass" (p. 93).

[2] In the case of tabular data, this is usually a reproduction of statistical distributions, and with image data, it is often the isolation of selected features.

Synthetic data models a politics of fragmentation. It recognizes and replicates real life by breaking down human subjects, echoing Simone Browne's (2009) concept of digital epidermalization: the disembodied gaze of surveillance biometric technologies fragments individuals into body components, "alienating the subject by producing a 'truth' about the body and one's identity" (p. 135). If the validity of synthetic data is its ability to mimic reality, and this mimicry is done primarily through these disassemblages, then the ontological basis of data-centric AI is a decontextualized, fragmented subject. In other words, any attempt to ameliorate the error of bias depends on the ability to produce representations of a fragmented, marginalized subject. State-of-the-art face-recognition methods like SynFace and Microsoft's DigiFace-1M (Bae et al., 2023), which boast rivaling error rates, highly depend upon fragmentation to render subjects more visible, readable, and legible to the ML model. Synthetic data may promise to reduce bias, but such promise depends on deconstructing an identity into arbitrarily selected components. While this fragmentation is not new, the creation of synthetic facial images is newly concerning because the images can be an essentializing projection of the developers' beliefs of what certain races, ethnicities, genders could and should look like.

Classifying

To discover and diagnose racial bias, one needs a coherent set of operationalized racial categories. It is probably unsurprising that synthetic data generative models rely on problematic classification systems, documented and criticized by scholars such as Wendy Chun (2021) and Kate Crawford (2022). The continuity of this violence of classification in synthetic data exposes the limitations of a conception of bias that relies on stable and unproblematic categorizations of marginalized groups.

Revisiting the notion of rebalancing a dataset, its fundamental logic requires consensus about which groups are underrepresented, what attributes are protected, and how to define and measure features. In a tabular dataset, protected attributes are often just the labels "race" or "gender." For image data, engineers or system designers decide which features are under or identifiable with these labels (e.g., relationships between skin color and race). To claim that the dataset is biased against a particular social group, there must be *a priori* assumptions of existing stable and identifiable groups. Synthetic data actually *produces* images according to these often arbitrary and reductive classifications.

In 2023, a team at New York University's Tandon School of Engineering developed a bias mitigation loss function to construct an unbiased GAN model for facial data. Jain, Memon, and Togelius (2023) used an auxiliary classifier that identifies the class label of an image, ensures equal amounts of data for each class, and generates new facial data accordingly. They claim that this method leads to fairer models, defined as a model that is "equally likely to generate images belonging to any of the classes in the training distribution" (Jain et al., 2023, para. 2). The auxiliary classifier was based on a pretrained ethnicity classifier of 6 racial groups—Blacks, Indians,[3] Asians, Whites, Hispanic Latinos, and Middle Easterners. Racial groups were treated as discrete and fixed, with no consideration of mixed-race individuals or the wide variety of

[3] The pretrained classifier does not explain this classification system, but from looking at their dataset, it appears like they mean people from the subcontinent of India and not Native American, although this is not made explicit.

people who could fall into these broad categories. While these problematic labels have long been used to divide and categorize faces, synthetic data takes it one step further by *producing* faces based on these assumptions, thereby "systematically engineering" (Ng, 2022, para. 1) representations of race.

These practices not only demand an oversimplification and stagnation of dynamic, complex categories but also ignore how categories are often decisions based on the needs of the ML model. Analyzing the history of mechanical emotion recognition, Crawford (2022) describes how one theory of emotion was chosen because it "seemed ideal for the emerging field of computer vision because they could be automated at scale. . . . The theory fits what the tools could do" (p. 175).

Thus, the novelty of synthetic data conceals how it relies on conventional data processing techniques: classification and annotation, or labeling. Ian Hacking's (2006) looping effect argues that classifications not only stabilize or reinforce norms and standards but also *create* the community and behavior it aims to identify. Rather than as reflections of seemingly objective hierarchies, classifications are actually mechanisms of inequality in of themselves. In the context of synthetic data, taxonomies gain a life of their own and persist as developers use them to generate images and information. This not only cements these classifications' problematic assumptions but reinforces the scale at which Generative AI can create and circulate seemingly neutral, objective, or natural representations of people.

Conclusion

In viewing bias as an error to ameliorate, proponents of synthetic data make several assumptions about the definition and impact of bias, as well as about the marginalized subject that needs protection from computational inequalities. Jacobsen (2023) argues that synthetic variance is capable of and culpable for enabling new divisions of the world (p. 6). Following Jacobsen (2023), we contend that synthetic data further pigeonholes bias as a technical problem—rather than a social one—that can be corrected easily by adjusting, or equalizing, the sample size.

At the different stages of synthetic data generation, its promise of reducing bias is validated by results that show ostensibly equal outcomes for different social groups. The history and foundations of inequality are seen as things that artificial training data can avoid, with the complexities of identity representations neglected in favor of higher accuracy and seemingly fairer outcomes. In operationalizing this narrow, outcome-oriented conception of bias, not only are foundations of inequality and injustice never addressed, but they are also exacerbated by and replicated in data synthesizing processes.

Ironically, concerns with the social, political, and technical senses of bias drive the promises and production of synthetic data. In pitches for synthetic data, companies co-opt "algorithmic bias" to market the value of both these datasets and the models that are trained on them. Yet, in looking at the actual practices to generate this data, it is apparent that this emergent technology replicates many of the same issues in models that use real-life data. In undermining these promises to eliminate bias, this article hopes to highlight the fundamental pitfalls of recapitulating bias into a merely technical error.

References

Aysha, A. (2023, September 25). *Rebalancing your data for ML classification problems—MOSTLY AI*. Retrieved from https://mostly.ai/blog/rebalancing-your-dataset-for-ml-classification-problems

Bae, G., de La Gorce, M., Baltrušaitis, T., Hewitt, C., Chen, D., Valentin, J., . . . Shen, J. (October, 2023). DigiFace-1M: 1 Million digital face images for face recognition. *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2023)*, 3526–3535. doi:10.48550/arXiv.2210.02579

Bao, J., Chen, D., Wen, F., Li, H., & Hua, G. (August, 2018). Towards open-set identity preserving face synthesis. *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)*, 6713–6722. doi:10.48550/arXiv.1803.11182

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Medford, MA: Polity.

Broussard, M. (2023). *More than a glitch: Confronting race, gender, and ability bias in tech*. Cambridge, MA: MIT Press.

Browne, S. (2009). Digital epidermalization: Race, identity and biometrics. *Critical Sociology*, *36*(1), 131–150. doi:10.1177/089692050934714

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*(1), 1–15. Retrieved from https://proceedings.mlr.press/v81/buolamwini18a.html

Chun, W. H. K. (2021). *Discriminating data: Correlation, neighborhoods, and the new politics of recognition*. Cambridge, MA: MIT Press.

Crawford, K. (2022). *Atlas of AI: Power, politics, and the planetary costs of Artificial Intelligence*. New Haven, CT: Yale University Press.

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, *8*(2), 1–14. doi:10.1177/20539517211035955

Dobson, J. E. (2023). Objective vision: Confusing the subject of computer vision. *Social Text*, *41*(3), 35–55. doi:10.1215/01642472-10613653

Hacking, I. (2006). Making up people. *London Review of Books*, *28*(16). Retrieved from https://www.lrb.co.uk/the-paper/v28/n16/ian-hacking/making-up-people

Jacobsen, B. N. (2023). Machine learning and the politics of synthetic data. *Big Data & Society*, *10*(1), 1–12. doi:10.1177/20539517221145372

Jain, A., Memon, N., & Togelius, J. (2023). *Fair GANs through model rebalancing for extremely imbalanced class distributions*. doi:10.48550/arXiv.2308.08638

Martineau, K. (2023, February 7). *What is synthetic data?* Retrieved from https://research.ibm.com/blog/what-is-synthetic-data

Ng, A. (2022). *How would you define data-centric AI development?* Retrieved from https://www.linkedin.com/posts/andrewyng_how-would-you-define-data-centric-ai-development-activity-6849022282543849472-jwtO/

Tiwald, P., Ebert, A., & Soukup, D. T. (April, 2021). Representative & fair synthetic data. *Synthetic Data Workshop—ICLR 2021*. doi:10.48550/arXiv.2104.03007

Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T. J., & Shotton, J. (September, 2021). Fake it till you make it: Face analysis in the wild using synthetic data alone. *2021 IEEE/CVF International Conference on Computer Vision (ICCV 2021)*, 3681–3691. doi:10.48550/arXiv.2109.15102