Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention

Anonymous Authors[1]

Facial recognition technology (FRT) has become commonplace, used from flight check-in at airports to police crowd-monitoring. Bias in FRT models has resulted in misidentification and expanded surveillance, causing unjust incarceration and other discriminatory outcomes. Attempts to solve these issues by increasing the accuracy of FRT run headfirst into problems; for a machine learning-based computer vision system to be considered robust enough for a given real-world task, it must "generalize" to images that vary widely in quality and domain (image granularity, race, age, gender, background, head pose, hats, glasses, etc.). Datasets with this level of granular design and annotation, that are also large enough for use in deep learning, are nearly impossible to find due to logistical and ethical concerns. As a result, researchers have turned to synthetic data generation, where data is generated to resemble something without being a representation of an instance of it — a drawing of a generic face as compared to a photograph of a real person. Synthetic data has been used to augment existing datasets and create new datasets for better training and evaluation of FRT models. Logistical and ethical challenges to data collection exist outside of FRT, and synthetic data usage has become commonplace across machine learning, from computer vision to large language models. This paper examines two key risks of using synthetic data.

Synthetic data is fundamentally useful where real data is not fit to task, necessitating that synthetic data must be both similar enough to be meaningful, but different enough to mitigate the reason the real data is not usable (Jordon et al., 2022). Jordan et al. propose three attributes of synthetic data that must be met for it to function in lieu of real data: utility, fidelity, and privacy. This paper focuses on facial recognition because it clearly articulates the risks of synthetic data, inherently forcing trade-offs between these attributes. There is high difficulty in making a picture of a face private but still usable as training data (privacy vs. utility) — a face which has been obscured to the point where an identity could not be gleaned is less useful (Chamikara et al., 2020). Achieving fidelity in facial datasets, a measure of how well synthetic data matches the real world, is also saliently difficult in facial recognition use cases, as we examine below.

This paper proceeds as follows: We begin by summarizing related prior work. We first examine work on datasets and representation, before discussing participation and consent and power over data and models. Finally, we discuss synthetic data and its use. We then proceed to the two titular risks of synthetic data that this paper focuses on — diversity-washing and the circumvention of consent. We draw upon two real-world examples: a facial recognition evaluation task conducted using synthetic data, and the FTC's enforcement actions against models trained on deceptively collected data to illustrate these risks. Finally, we expand upon how these two risks are examples of irresponsible use of synthetic data: consolidating power in the hands of model creators, and decoupling data from those it represents and those who are harmed by its improper use. It is our intention for this research to contribute to the field by presenting tangible examples and background for the challenges inherent in responsible use of synthetic data, thus laying foundations for further work and debate. We call for future work to examine the breadth and usage of synthetic data and to work towards both mitigating synthetic data's risks and enabling its potential for participatory empowerment.

Expanding upon the content of the paper, the first risk we focus on is the *high risk of false confidence* in the ability of synthetic datasets to mitigate bias in data distribution and representation. We demonstrate this through the real-world example of using synthetic data for a facial recognition model evaluation. This paper was motivated by the realization of the under-explored risks of synthetic data while conducting the evaluation, and we present it both to provide an example of how synthetic data is used and to detail the concerns that conducting it made apparent. In brief, synthetic data offers a way of diversifying datasets, but diversity in real-world faces often follows from cultural practices that are qualitative and meaning-laden rather than quantitative. Creating a synthetic dataset or adding synthetic data to existing datasets in an attempt to diversify that dataset runs the risk of *diversity-washing* — appearing to

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

resolve valid criticism regarding a dataset's distribution and representation but in a way that is superficial. As a result, using synthetic data risks legitimizing technologies such as FRT despite potentially continuing to propagate bias by achieving false fidelity, as well as risks creating datasets — and subsequently training and evaluating models from that data — that fail to mitigate bias in data distribution and representation. Furthermore, there is a risk of propagating harm through a patina of legitimacy, and through diversity-washing potentially harmful models.

The second risk we examine is how using synthetic data risks *circumventing consent for data usage*, a key element of the U.S. Federal Trade Commission's regulation of data collection and affected models. Synthetic data provides an avenue for model developers to side-step thorny issues around collecting large-scale representative facial datasets. Proper consent to data usage is foundational to the privacy enforcement tools that the FTC has used to require companies delete ML models trained on improperly collected data. Consent also plays a role in U.S. sectoral statutes such as HIPPA, U.S. state laws such as California's California Privacy Rights Act and Illinois' Biometric Information Privacy Act, as well as laws in many other countries, with the EU's GDPR a notable example. These statutes share a common goal: to provide people with control over their personal data, via notification, access, and consent regarding the collection, use, and disclosure of personal data. This type of privacy regulation is referred to as "privacy self-management" by Solove (Solove, 2012), and focuses solely on whether or not data subjects have consented, rather than on value judgements of privacy practices. This section focuses on illustrating the risk that synthetic data poses to consent-based frameworks by expanding upon how the Federal Trade Commission (FTC) has functionally used consent as a key aspect of conducting enforcement against companies using ML systems, by enforcing their deception authority. The analysis is guided by one author's experience at the FTC, but draws upon solely public knowledge.

Finally, we discuss how synthetic data risks consolidation of power, as showcased when considering the inherent relational aspects of data privacy. Solon Barocas and Helen Nissenbaum identify the risk of a "tyranny of the minority" in big-data analytics when "the volunteered information of the few can unlock the same information about the many" (Barocas & Nissenbaum, 2014). More recently, Salome Viljoen emphasizes the importance of a relational theory of data governance (Viljoen, 2021). As Viljoen explains, dataflows entail not only 'vertical' relations between a particular individual and a data collector, but also 'horizontal' relations between the individual and others sharing relevant population features. Viljoen focuses on the manner in which informational infrastructures rely on group classification to make sense of individuals by taking a 'relevant shared feature,' generating a prediction based upon that shared feature, and then applying this prediction. When those shared features are derived from synthetic data, decoupled from any real context and perhaps even specifically created to rectify gaps in representation, we hand power to those creating that synthetic data. We risk imposing designers' decision-making in lieu of and upon those least likely to have been represented and most likely to be harmed by both the diversity-washing and the side-stepping of consent. After all, if they were represented or able to consent in the first place, there would be no need for additional synthetic data. Data minimization and lineage principles (Hutson & Winters, 2022) are a first step towards mitigating this issue by requiring documentation and its requisite transparency into where data has come from. The need for this is also readily apparent when considering contexts such as the EU's Right to be Forgotten (Razmetaeva, 2020), where synthetic data further complicates the ability to be removed from a dataset. In making it harder to decouple data from its context through the use of synthetic data, there is an avenue for mitigating consolidation of power and ensuring consent.

To conclude, in this paper, we build on prior responsible dataset development work by focusing on the underexplored impacts of synthetic data on dataset development. Synthetic data will continue to play an ever-increasing role in the training of machine learning systems as real-world data becomes harder to capture, and we must attend to language that paints it as a panacea. We show two examples of the risks of synthetic data, diversity-washing and consent circumvention, and discuss how it is a complicated tool, gravitationally prone to consolidation of power, but with potential for being used to enable participatory governance instead of squashing it.

References

Barocas, S. and Nissenbaum, H. Big data's end run around procedural privacy protections. *Communications of the ACM*, 57(11):31–33, October 2014. ISSN 0001-0782. doi: 10.1145/2668897.

Chamikara, M. A. P., Bertok, P., Khalil, I., Liu, D., and Camtepe, S. Privacy Preserving Face Recognition Utilizing Differential Privacy. *Computers & Security*, 97: 101951, October 2020. ISSN 0167-4048. doi: 10.1016/j.cose.2020.101951.

Hutson, J. and Winters, B. America's Next 'Stop Model!': Model Disgorgement, September 2022.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. Synthetic Data – what, why and how?, May 2022.

Razmetaeva, Y. The Right to Be Forgotten in the European

Perspective. *TalTech Journal of European Studies*, 10(1): 58–76, May 2020. doi: 10.1515/bjes-2020-0004.

Solove, D. J. Privacy Self-Management and the Consent Dilemma. November 2012.

Viljoen, S. A Relational Theory of Data Governance. *the yale law journal*, 2021.