BIG DATA & SOCIETY

Commentary

Big Data & Society
April-June: 1–5
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517241249447
journals.sagepub.com/home/bds
Sage

Generating reality and silencing debate: Synthetic data as discursive device

Paula Helm[1], Benjamin Lipp[2] and Roser Pujadas[3]

Abstract

In addition to tapping data from users' behavioral surplus, by drawing on generative adversarial networks, data *for* artificial intelligence is now increasingly being generated *through* artificial intelligence. With this new method of producing data synthetically, the data economy is not only shifting from "data collection" to "data generation." Synthetic data is also being employed to address some of the most pressing ethical concerns around artificial intelligence. It thereby comes with the sociotechnical imaginary that social problems can be cut out of artificial intelligence, separating training data from real persons. In response to this technical solutionism, this commentary aims to initiate a critical debate about synthetic data that goes beyond misuse scenarios such as the use of generative adversarial networks to create deep fakes or dark patterns. Instead, on a more general level, we seek to complicate the idea of "solving," i.e., "closing" and thus "silencing" the ethico-political debates for which synthetic data is supposed to be a solution by showing how synthetic data itself is political. Drawing on the complex connections between recent uses of synthetic data and public debates about artificial intelligence, we therefore propose to consider and analyze synthetic data not only as a technical device but as a discursive one as well. To this end, we shed light on their relationship to three pillars that we see associated with them (a) algorithmic bias, (b) privacy, (c) platform economy.

Keywords

Synthetic data, artificial intelligence, generative adversarial networks, algorithmic bias, privacy, platform economy

Introduction

In the venture capital sector, the industrial complex, and in academia there is now unprecedented excitement about how to best meet requirements to drive artificial intelligence (AI). Apart from security and ethico-political concerns, this includes generating and benchmarking the necessary data to train models on mass-scale, across domains, and in various languages (Zha et al., 2023). However, since not all needed data can be easily obtained and also not for every case and requirement, alternative solutions are being sought. One of these is synthetic data (Heaven, 2021). In addition to tapping data from users' behavioral surplus, by drawing on generative adversarial networks (GANs), data *for* AI is now also being generated *through* AI (Marriott et al., 2020).

This new method shifts the data economy from "data collection" to "data generation." With its emphasis on generating data, it seems to reproduce some of the basic tenets of critical data studies, while at the same time fundamentally contradicting them. On the one hand, the discussion of synthetically generated data complicates the notion of raw data, debated in critical data studies for more than a decade (Bowker, 2008). With the proliferation of synthetic data, misleading notions of data as a direct, immediate reflection of reality are giving way to a terminology and understanding that not only the results of AI, but also the data with which it is trained, are produced, and thus undergo processes of manufacture and design. Claims to transparency are thereby extended from the conditions of model production to those of data production. On the other hand, though, the emphasis on generated data as opposed to collected data can underline a dualistic cut between the real and the artificial. Here, only "the real" is

[1]Media Studies Department/Data Science Center, Critical Data & AI Studies Group, University of Amsterdam, Amsterdam, The Netherlands
[2]Department of Technology, Management and Economics, Section of Human-Centered Innovation, Technical University of Denmark, Copenhagen, Denmark
[3]Department of Science, Technology, Engineering, and Public Policy, University College London, London, UK

Corresponding author:
Paula Helm, Media Studies Department/Data Science Center, University of Amsterdam, Turfdraagsterpad 9, 1012 XT Amsterdam, Netherlands.
Email: p.m.helm@uva.nl

exposed to ethical scrutiny while synthetic data is shielded from it.

Originally, one of the main reasons for using synthetic data was to add anomalies and variations to systems that were mainly trained on the standard state, e.g., to supplement information about the healthy body, the well-functioning car or normal weather conditions with data about tumors, accidents, and extreme weather, without having to document these rare and dangerous events through a plethora of real cases (Nikolenko 2021; Raghunathan 2021). Synthetic data is hence seen as somewhat detached from reality and therefore less risky (Jacobsen, 2023). In addition to solving such problems that are primarily constructed on a technical level, this new technology is also increasingly being touted as a promising technical solution to difficult ethico-political problems associated with AI. The underlying idea is that social problems can be excluded from AI by separating data from people. However, by supposedly securing data, its synthetic production can serve as a shortcut that silences, rather than advances, critical debates about the socio-ethical implications of AI.

In this commentary, we problematize this silencing that the transition to synthetic data could engender. Rather than delving deeper into more fundamental issues underlying debates about digital privacy (Solove, 2007), training data bias (Boulamwini and Gebru, 2018), and platform capitalism (Srnicek, 2017), synthetic data allows companies to simply sidestep concerns by offering yet another technical fix for socio-technical problems. Against this, we insist that any, including synthetic solutions, are ethico-political least because they shape our imagination of strategies and alternatives to come to terms with issues of surveillance, discrimination, and capital accumulation in data-intensive economies.

The aim of this commentary is therefore to initiate a critical debate on synthetic data that goes beyond misuse scenarios such as the deployment of GANs to create deep fakes or dark patterns (de Vries, 2020). Instead, on a more general level, we intend to complicate the idea of “solving,” i.e., “closing” and thus “silencing” the debates for which this technology is supposed to provide a solution by showing how synthetic data itself is political. Based on the complex connections between recent uses of synthetic data and public debates on AI, we, therefore, propose to consider and analyze synthetic data not only in its technical functionality, but in its functionalities as a discursive-political device as well. By this, we mean that synthetic data not only have material effects but also shape ethico-political debate, while negating the need for critical examination of the farther-reaching effects of generative forms of data processing. As the problem of data collection becomes translated into issues of data production, we might, paradoxically, lose the intuition that this constructedness requires critical scrutiny. Against this, we highlight three pillars that we see associated with synthetic data discourse: (a) algorithmic bias, (b) privacy, (c) platform economy.

Algorithmic bias

One of the ethical problems that synthetic data is supposed to address is algorithmic discrimination caused by distortions in training data sets. Companies praise synthetic data as a way to provide “unbiased”[1] data or to “reduce bias”[2] significantly. The idea behind this is that by artificially generating data sets, the statistical distribution of features such as different skin colors can be ensured, and, furthermore, anomalies be included. However, this idea of including endless possibilities in the data has limitations.

Giuffré and Shung (2023) argue, for the case of healthcare, that there is also a real danger that those who design and use AI systems trained with synthetic data overgeneralize or overestimate their results, thus potentially worsening the issue of bias they are meant to address. This can lead to the “creation of non-existent or incorrect correlations.” (ibid, 3). Furthermore, synthetic data produced by GANs add new layers of difficulty in correctly interpreting and checking AI-driven decision-making in clinical practice, as GANs, like other deep neural networks, are black boxes (Chen et al. 2021: 494). Hence, while synthetic data is used to account for diversity in datasets blind spots in AI development as well as the lack of contextual fit are not really *solved*. Rather, synthetic data shift issues of bias to other dimensions of data production that require critical inquiry themselves.

More fundamentally, using synthetic data as a technical shortcut to deal with problems of discriminatory bias may promote a view that discourages critical ethical scrutiny into the systemic and social conditions of bias. For instance, the institution of medicine is widely recognized for exhibiting considerable structural bias toward marginalized groups (Hammond et al., 2021). Digital technology adds another layer of complication and additional sources of bias to these existing problems, for example, via the combined underrepresentation and marginalization of people of color and women in both: Medical research *and* tech.

As research shows, bias in the tech industry stems to a considerable degree from the structural misrepresentation of those groups in its workforce and management (Neely et al., 2023). So, even when we acknowledge that social problems can partly be mitigated by technology, the promissory discourse of synthetic data reduces the scope of the problem to data rather than enquiring into the wider systems that run on and through them. While data-intensive systems building on synthetic solutions do not have to be concerned with infrastructures of collection as well as the resistance and politics of data subjects, the social locus of data production becomes even more entrenched in the very same communities and institutions that have given rise to these biases in the first place. The language of synthetic data as a guarantor of diversity might thereby result in a problematic combination of a simplification of what is at stake when we talk about diversity, while at the same time reducing the pressure of installing guardrails for the ongoing experimentation with AI in society (Helm et al., 2022).

Privacy

Synthetic data is also used to respond to problems related to data scarcity caused by requirements for privacy protection. However, the use of synthetic data for, i.e., anonymization, has been drawn into question. Stadler et al. (2022: 15), for instance, argue that "synthetic data shares similar tradeoffs with previous techniques, highlighting the unpredictable nature of privacy gains in synthetic data publication." Even setting aside these doubts on a functional level, on a more fundamental level, it is worth considering the different constructions of data, how they function and what effects they have, i.e., what distinguishes synthetic data from other types of data production (*or not*). In information capitalism, users' activities are constantly tracked, not as individual traces but as aggregates, that is, statistics that can be turned into probabilities and prediction. Despite this, data protection is still mostly associated with a specific "user" (Taylor et al., 2018). In line with this limited view, the attractiveness of synthetic data for privacy protection is presented as detaching data from the user. Yet, to be of value, synthetic data still needs to correspond closely to a "real world" context. It must thus strike a balance between detachment from real-world persons while closely resembling real-world problems (Jacobson, 2023).

A striking example of this is how IDEMIA employs synthetic data for criminal investigation solutions, where data scarcity is not only and/or primarily caused by the rarity of events but by privacy issues pertaining to the protection of involved third parties (Helm and Hagendorff 2021). To fix this, IDEMIA turns to synthetic data: "In compliance with relevant privacy regulations (…) we create synthetic images (…) that are completely fictional"[3]. But are these images really "completely fictional"? In practice, producing synthetic data for a concrete application might look as such: (a) Based on statistics and end-user input a stereotypical crime scene is scripted. (b) This scene is then reproduced and recorded. (c) The so obtained footage is turned into data. (d) The data is multiplied via GANs, and (e) used as training data.

Such use of synthetic data to circumvent privacy concerns underlies a narrow, individualistic conceptualization of what the protection of privacy is about. Synthetic data may be separated from me as an individual, but not from me as a member of a group: A point in a statistic, an inhabitant of a district, a person fitting into a certain norm, while deviating from another. Synthetic data might, as a result, still encode sensitive information about real people (Renieris, 2023). Given the relevance of data protection not just on the individual, but also on the group level (Helm 2017; Taylor et al. 2018), individualist notions of privacy have long been proven to be inadequate when it comes to data processing and digital networking (Solove 2007). Instead, if we understand privacy in its relation to contexts (Nissenbaum, 2010), democracy (Seubert and Helm, 2017), and society (Rössler and Mokrosinska, 2015), synthetic data complicates, rather than solves problems related to privacy-preserving handling of metadata, statistical data, mobility data, etc.

Platform economy

Data has become a central asset to present-day economies. Notions such as platform capitalism (Srnicek, 2017) capture "a new mode of capitalist production in which digital data, harvested via surveillance, is of central importance to valorization." (Steinhoff 2022: 4). Indeed, in today's economy, the trading of data has become an integral part of the business models of the world's largest companies. How should synthetic data be positioned in this context?

Capital accumulation results not only from data harvesting but increasingly involves prediction and the production of real-time insights through the analytical capabilities of AI (Pujadas et al., 2024). Thus, synthetic data can be seen as both the input and output of an economy on the move towards the hyperreal simulacrum (Baudrillard, 1994), in which synthetic data is "pitched as represent[ing] the real world in a very even way, better than the real world does" (Staff 2021, in Steinhoff 2022: 11). Furthermore, synthetic data is more accessible and cheaper than other data and presented by companies like Datagen as "free from the headaches of manual data acquisition, annotation, and cleaning" (Renieris, 2003: 87). If quantity of data is not a problem anymore, current extractivist AI models seem to be removed from almost all barriers. Thus, the discourse of synthetic data reinforces imaginaries of the unavoidability of current AI-driven innovation trajectories.

Like other forms of technological innovation, synthetic data can be seen as a means to make sure that distributions of symbolic and material reward remain within entrenched circuits of privilege and capital (Bishop and Suchman, 2000: 332). This could amplify the existing reinforcing mechanisms that place Big Tech into monopolistic positions. This raises ethical questions around data justice (Dencik et al., 2022), and the unequal material conditions of possibility that AI may perpetuate (Verdegem, 2023). Since synthetic data are part of these conditions, calls for data justice should not halt here.

Apart from these considerations, it is worth mentioning that synthetic data is not the only viable option to solve problems of training data availability. For example, visual analytics techniques, which can be used in combination with active learning, offer an alternative. These are systems that present data in various, often sophisticated ways and then learn through active use and adoption by experts in real-life situations. They are therefore not dependent on huge amounts of training data. Another advantage is the ability to pause the learning process manually if appropriate (Fischer et al., 2022). This also enables professionals to tailor systems to their needs, thus increasing diversity and user empowerment (Ametowobla and Prechelt, 2020). Hence, the hype around synthetic data might offer big companies' new opportunities to preserve their market power, but it also renders invisible alternatives that favor responsibility and decentralization.

Conclusion

Synthetic data does not actually resolve ethico-political questions but shifts them from the mode of data collection to data production, from problems of representation to problems of design. Instead of learning a more fundamental lesson about the limits of capitalism and the persistence of historical domination, ethico-political problems are addressed through the very same logic that created them. However, the argument of "relocating" a problem rather than "displacing" it is (a) a familiar one when it comes to automation technology (and the supposed replacement of human labor), and (b) raises the more general question: Where does the shift move the problem to? In this case, the answer seems clear: It shifts ethical questions that arise in the context of AI technology to the data science departments and laboratories of powerful corporations and other influential stakeholders. In doing so, these companies are working from the almost hyperbolic idea that to solve complex, historically rooted, onto-epistemic problems related to algorithmic discrimination, surveillance, and exploitation, we can simply invent the reality we want—but by drawing from the very patterns that originally created the problems that make that reality impossible. While the idea of synthetic data is charming due to its performative dimension, it is also flawed in the ways in which it is currently employed. This flaw stems above all from the persistence of the distinction between raw, collected data and synthetic, produced data. As long as this distinction is maintained, synthetic data functions as a discursive device that excludes the latter from ethical scrutiny by neglecting the performative power of data, including synthetic data.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Data Science Center University of Amsterdam,

ORCID iDs

Paula Helm https://orcid.org/0000-0002-2719-9721
Benjamin Lipp https://orcid.org/0000-0003-4220-1532
Roser Pujadas https://orcid.org/0000-0002-8837-6744

Notes

1. https://syntheticus.ai/synthetic-data-for-healthcare-and-pharma, Accessed February 29, 2024.
2. https://www.ibm.com/topics/synthetic-data, accessed February 29, 2024.
3. https://www.idemia.com/insights/why-artificial-intelligence-so-crucial-modern-identity-and-security-technologies, accessed, February 29, 2024.

References

Ametowobla D and Prechelt L (2020) How layered reuse can support harmful micropolitics: SAP ERP in surgery planning. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Software Engineering in Society, ICSE-SEIS '20*. New York, NY, USA: Association for Computing Machinery. pp. 39–48.

Baudrillard J (1994) *Simulacra and Simulation*. Ann Arbor, Michigan: University of Michigan Press.

Bowker G (2008) *Memory Practices in the Sciences*. Illustrated edition. Cambridge, Mass.: MIT Press.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: *Conference on fairness, accountability and transparency*, pp.77–91.

Chen R, Lu M, Chen T, et al. (2021) Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engin* 5(6): 493–497.

Dencik L, Hintz A, Redden J, et al. (2022) *Data Justice*. Thousand Oaks: SAGE.

De Vries K (2020) You never fake alone: Creative AI in action. *Information, Communication & Society* 23(14): 2110–2127.

Fischer M, Hirsbrunner S, Jentner W, et al. (2022) Promoting ethical awareness in communication analysis: Investigating potentials and limits of visual analytics for intelligence applications. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul Republic of Korea, pp.877–889.

Giuffrè M and Shung DL (2023) Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *NPJ Digit Med* 6: 186.

Hammond M, Stehli J, Stavros G, et al. (2021) Bias in medicine. *JACC: Basic to Translational Science* 6(1): 78–85.

Heaven W (2021) Synthetic Data for AI. *MIT Technology Review*. Available at: https://www.technologyreview.com/2022/02/23/1044965/ai-synthetic-data-2/.

Helm P (2017) Group privacy in times of big data. *Digital Culture & Society* 16(2): 137–152.

Helm P and Hagendorff T (2021) Beyond the prediction paradigm: Challenges for AI in the struggle against organized crime. *Law and Contemporary Problems* 84(3): 1–17.

Helm P, Michael L and Schelenz L (2022) Diversity by Design? Balancing the Protection and Inclusion of Users in Online Social Networks. In *2022 AAAI/ACM Conference on AI, Ethics, and Society*, Oxford, United Kingdom,

Jacobsen B (2023) Machine learning and the politics of synthetic data. *Big Data & Society* 10(1): 1-12.

Marriott R, Madiouni S, Romdhani S et al. (2020) An Assessment of GANs for Identity-Related Applications. *Arxiv*. 978-1-7281-9186-7/20/

Neely M, Sheehan T and Williams C (2023) Social inequality in high tech: How gender, race, and ethnicity structure the world's most powerful industry. *Annual Review of Sociology* 49(1): 319–338.

Nikolenko S (2021) *Synthetic Data for Deep Learning* (Vol. 174). Cham: Springer International Publishing.

Nissenbaum H (2010) *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, Calif: Stanford Law Books.

Pujadas R, Valderrama E and Venters W (2024) The value and structuring role of web APIs in digital innovation ecosystems: The case of the online travel ecosystem. *Research Policy*, 53(2): 104931.

Raghunathan T (2021) Synthetic data. *Annual Review of Statistics and Its Applications* 8: 129–140.

Renieris E (2023) *Beyond data: Reclaiming human rights at the dawn of the metaverse.* Cambridge, Mass.: The MIT Press.

Roessler B and Mokrosinska D (2015) *Social Dimensions of Privacy: Interdisciplinary Perspectives*. Cambridge, UK: Cambridge University Press.

Seubert S and Helm P (2017) Privatheit und demokratie. *in: Forschungsjournal Soziale Bewegungen* 2017(2): 120–124.

Solove D (2007) I've got nothing to hide" and other misunderstandings of privacy. *San Diego Law Review* 44(4): 745–772.

Srnicek N (2017) *Platform capitalism*. Cambridge, UK: Polity.

Stadler T, Oprisanu B and Troncoso C (2022) *Synthetic Data—Anonymisation Groundhog Day* (arXiv:2011.07018). arXiv.

Steinhoff J (2022) Toward a political economy of synthetic data: A data-intensive capitalism that is not a surveillance capitalism? *New Media & Society* 0(0): https://doi.org/10.1177/14614448221099217

Suchman L and Bishop L (2000) Problematizing 'innovation' as a critical project. *Technology Analysis & Strategic Management* 12(3): 327–333.

Taylor L, Floridi L and Sloot B (2018) *Group Privacy: New Challenges of Data Technologies*. Cham: Springer International Publishing.

Verdegem P (2023) Critical AI studies meets critical political economy. In: *Handbook of Critical Studies of Artificial Intelligence*. Cheltenham: Edward Elgar Publishing, 302–311.

Zha D, et al. (2023) "Data-Centric Artificial Intelligence: A Survey." http://arxiv.org/abs/2303.10158 (December 6, 2023).