Surveillance & Society

Synthetic Data: From Data Scarcity to Data Pollution

Tanja Wiehn

Volume 22, Number 4, 2024

Open Issue

URI: https://id.erudit.org/iderudit/1115679ar
DOI: https://doi.org/10.24908/ss.v22i4.18327

See table of contents

Publisher(s)

Surveillance Studies Network

ISSN

1477-7487 (digital)

Explore this journal

Cite this document

Wiehn, T. (2024). Synthetic Data: From Data Scarcity to Data Pollution. *Surveillance & Society*, *22*(4), 472–476. https://doi.org/10.24908/ss.v22i4.18327

Article abstract

The increasing development and adaptation of synthetic data raises critical concerns about the perpetuation of datafication logics. In examining some of synthetic data's core promises, this dialogue paper aims to uncover the potential harm of further de-politicizing synthetic data. With synthetic data, technological opportunities are introduced that promise to resolve a growing demand for data needed to train AI models. Furthermore, models trained on synthetic data are praised as more precise and effective while bring cheaper than collected data (Zewe 2022). With this dialogue paper, I aim to nuance the ways in which synthetic data complicate a critique directed at AI-driven technologies. I build my argument on two elements fundamental to the debate on the promises and perils of synthetic data. The first is the notion of data scarcity—often leveraged to argue for the implementation and further development of synthetic data to train bespoke models. Second, I discuss the concerns of data pollution and contamination with synthetic data. Through these entry points, I argue that synthetic data re-ignites issues previously raised by scholars in the field of critical data and surveillance studies. Therefore, the aim of this dialogue paper is to call for a critical understanding of synthetic data as living information, much like collected data, and to account for synthetic data and the conditions of its generation in the context of simulated environments.

Surveillance & Society

Dialogue

Synthetic Data: From Data Scarcity to Data Pollution

Tanja Wiehn

Roskilde University, Denmark
tanjaaw@ruc.dk

Abstract

The increasing development and adaptation of synthetic data raises critical concerns about the perpetuation of datafication logics. In examining some of synthetic data's core promises, this dialogue paper aims to uncover the potential harm of further de-politicizing synthetic data. With synthetic data, technological opportunities are introduced that promise to resolve a growing demand for data needed to train AI models. Furthermore, models trained on synthetic data are praised as more precise and effective while bring cheaper than collected data (Zewe 2022). With this dialogue paper, I aim to nuance the ways in which synthetic data complicate a critique directed at AI-driven technologies. I build my argument on two elements fundamental to the debate on the promises and perils of synthetic data. The first is the notion of data scarcity—often leveraged to argue for the implementation and further development of synthetic data to train bespoke models. Second, I discuss the concerns of data pollution and contamination with synthetic data. Through these entry points, I argue that synthetic data re-ignites issues previously raised by scholars in the field of critical data and surveillance studies. Therefore, the aim of this dialogue paper is to call for a critical understanding of synthetic data as living information, much like collected data, and to account for synthetic data and the conditions of its generation in the context of simulated environments.

Introduction

In 2014, José van Dijck discussed the problematics of datafication and dataveillance in an article for *Surveillance & Society.* In analysing the persistent logic of collecting information about users and citizens, van Dijck (2024) describes here dataveillance as a process that extends surveillance for specific purposes and datafication as "a means to *access*...and *monitor* people's behavior" (Van Dijck 2014: 198; italics in the original). Synthetic data promises to remove the issue of surveillance and dataveillance and the frameless collection of data (Andrejevic 2020; Steinhoff 2022). However, in line with Susser and Seeman (in this issue), I argue that synthetic data presents a similar moment of techno deterministic hype, much like big data a decade ago. The capability to create bespoke synthetic data for AI development holds the potential to perpetuate logics of datafication with processes and conditions under which value is generated from data, with harmful consequences (Mejlias and Couldry 2019; van Dijck 2014; Zuboff 2019).

Synthetic data is roughly defined as data that has not been collected nor mined from subjects or events in the real world but generated by algorithmic systems (Jordon et al. 2022). Governments and industries are highly invested in the generation and governance of synthetic data, for reasons as diverse as AI development, bias mitigation, or economic advantages on a global scale (de Wilde 2024; Helm, Lipp, and Pujadas 2024).

Wiehn, Tanja. 2024. Synthetic Data: From Data Scarcity to Data Pollution. *Surveillance & Society* 22 (4): 472-476.
https://ojs.library.queensu.ca/index.php/surveillance-and-society/index | ISSN: 1477-7487

Furthermore, a growing privacy industry adapts synthetic data technologies, even though there are reasonable doubts about the feasibility of data privacy without the loss of utility for data sets (Munkholm and Wiehn 2025). Data policy and regulation are already affected by the ways in which synthetic data challenges definitions of personal data and data privacy (Beduschi 2024; De Wilde et al. 2024). My critical provocations towards synthetic data open a way to think through synthetic data as living information, inherently political and always on a threshold of becoming (Amoore 2020; Kaufmann 2023; Thylstrup 2022). In outlining concerns around the development of AI models, that is data scarcity, data pollution, and contamination, my aim is to trace how synthetic data does not place AI models out of the realm of risk of being harmful (Jacobsen 2023).

From Data Scarcity to Data Pollution

In early 2024, the United Nations University published a policy guideline with recommendations on the global use of synthetic data for the training of AI models (De Wilde et al. 2024). The guideline, directed at the IT sector, companies, and governments, is formulated as a reaction to the increasing use of synthetic data on a global scale (De Wilde et al. 2024). It is an indicator for the importance of synthetic data and an awareness of regulatory needs. Here, the nascent synthetic data industry is explained through a higher rate of adaptation of AI systems that require more and better training data (Wilde et al. 2024). Synthetic data is thus a reaction to data scarcity, a term invoked by computer scientists (Nikolenko 2021) that relates to the lack of precise and high-quality data for the training of AI models. Data can be lacking because they would be difficult, risky, labour intensive, and expensive to collect and label. Scarcity is a key argument for the further development and distribution of synthetic data for data industries, especially as models trained on synthetic data are praised as more precise and effective while being cheaper than collected data (Zewe 2022).

While the frictions of real-world datasets often resist computational analysis, synthetic data can be made to perform beautifully in statistical models by correcting biases, seamlessly filling gaps, sanitizing and regularizing outliers, scaling up resolution by in-painting plausible details, increasing variability, and decreasing ambiguity (Offenhuber 2024: 13).

Synthetic data industry players, like NVIDIA's (n.d) Omniverse 3D generator or Synthesis.ai's (n.d.) Synthesis Scenarios, collect synthetic data in fully simulated environments such as in simulated digital factory facilities and public spaces for applications including activity classification, threat detection, and pedestrian traffic analytics. Synthesis.ai (n.d.) describes the purpose and use of their synthetic scenarios as follows: "Synthesis Scenarios lets CV [computer vision] teams create labelled 3D data scenes populated with synthetic humans that have realistic facial, body and hand motion across a wide variety of body types, camera angles, backgrounds, and environments." What is often left out in marketing claims like these is the fact that people make decisions about the design and constellations of these simulated pedestrians, objects, and environments (Korenhof, Blok, and Kloppenburg 2021). Synthetic data generated from simulations reinvigorates the question: to what extent does synthethic data hold the potential to perpetuate harmful AI models and surveillance practices, similar to machine learning models trained on non-synthetic data (Kaufmann 2023)?

The data scarcity resolved with synthetic data generation is thus not only justification to generate data on demand and to create new supply chains for the access, portability, and potential re-usability of data (Thylstrup et al. 2022). It further underpins the logics of algorithmic regimes of verification "as new forms of identifying a wrong or of truth telling in the world." (Amoore 2020: 5f). Synthetic data fundamentally changes the ways in which the data industry satisfies a need for data for AI applications (Jacobsen 2023; Steinhoff 2022). Synthetic data sets out to overcome strategic disadvantages in AI development for regions like the Global South, which could lead to less dependence on globally operating tech companies (Wilde et al. 2024). As Ravn rightly implies in this issue, there also lies an emancipatory potential for synthetic data in overcoming lack of data. However, I echo Susser and Seeman (in this issue) when arguing for further analyses of for whom new synthetic data economies become most (dis)advantageous. This is also relevant when looking at the concerns of data pollution and contamination invoked by synthetic data sets and data sets blended with real and synthetic data: "If the synthetic data are not balanced, misrepresent a population group, or are otherwise biased, their biases could propagate throughout trained models and even to other synthetic datasets" (De Wilde et al. 2024: 5). Pollution and contamination imply how data sets can turn unusable, misrepresentative, and unfair. Synthetic data industries chime into a narrative of the fixable, correctable datasets to achieve unbiased and effective AI systems. Data pollution reiterates the core issues of data sets collected in the "real" world. What is an adequate representation of groups of population, gender, race, and other attributes in a data set? Where is agency located in the curation of a data set made up of a combination of "real" data and synthetic data?

The incorporation of enough high quality and otherwise missing data aims, at best, to achieve a form of algorithmic fairness, a notion that describes the achievement of statistical equity. Jacobsen (2024: 8) emphasizes the underlying issues of supplementing data as part of data sets stating, "this promise to resolve political imbalances is problematic, because it obfuscates how machine learning can reconfigure the very notions of race and ethnicity." Critical scholarship has debunked the value of this fairness in AI systems and further surfaced structures of dominance created by efforts to fix bias with technical means (Hoffmann 2020; West 2020). In the context of generative AI, recent studies have leveraged feminist and intersectional analysis to note the significance of asking *how* patterns of bias are generated and how they figure in specific socio-cultural contexts (Deviancy, Björklund, and Björklund 2024). Synthetic data's potential to perpetuate bias is further enhanced when it is generated without the engagement and participation of marginalized groups (Wilde et al. 2024). Another concern lies in data contamination, when a blend of synthetic data and real data becomes very difficult to separate. In other words, new efforts and research techniques are needed to detect, analyse, and challenge these issues brought about by synthetic data. This marks the fragile dependence and dynamic of data sets and their synthetic supplements (Jacobsen 2024). It underlines how synthetic data prolong techno deterministic understandings of datafication with data (sets) as apolitical entities (Jacobsen 2023; Thylstrup 2022). This is an imposition appropriately described by Fitzgerald (in this issue) as the "intensification of a pre-existing lack of accountability inherent within automated systems more generally." Critical provocations of synthetic data need to address data and data sets as inherently political, where every change, adjustment, and adaptation can cause real life consequences (D'Ignazio and Klein 2020; Thylstrup 2022).

Through the discussion of pollution, contamination, and scarcity, my aim is to signal to a vitality of data that stands in stark contrast to the projected inertness of *synthetic* data. This allows us to understand data as living information, always on the threshold of becoming and existing together in data life cycles, which proves to be relevant for the case of surveillance (Amoore 2020; Kaufmann 2023). Kaufmann (2023: 68) underlines how a look into data life cycles allows us to understand how surveillance practices, such as predictive policing efforts, shape the creation of information through the involvement of people, infrastructures, tools, and imaginaries: "Information is imbued with its particular history of being imagined, generated, and stored. Such histories travel to more abstract, harder-to-comprehend contexts as information rematerializes in new association processes." In the case of synthetic data, technical and statistical imaginaries shape preconceptions of the demand and needed quantity of data while hindering ethical debates (Jacobsen 2023). Synthetic data might become a new roadblock for the ethico-political discourse in AI development. Nevertheless, synthetic data can demonstrate how an information life cycle is shaped by a multitude of human and non-human agencies (Kaufmann 2023). My examples of synthetic data generation in fully simulated environments are one way to further discuss preconceptions of data and datafication imaginaries. They beg the questions: In what ways do human and non-human agencies, imaginaries and speculations impact the generation of synthetic data? And when synthetic data create an "idealized representation heavily mediated by beauty filters and image manipulations" (Offenhuber 2024:13), what structures of dominance are perpetuated?

Conclusion

Synthetic data technologies are being introduced to react to a growing demand for data required to train AI models. It sounds particularly promising that data can be provided without the usual constraints and costs of collected data (Gitelman 2013; Kitchin 2021; Thylstrup 2022). In the development and further implementation of synthetic data, persisting *technical* and *statistical* issues can be resolved. However, synthetic data promotes a further de-politization of data whilst constructing industries of further capitalization of data (Steinhoff 2022). Considering the underlying promise for the abundance and malleability of data, and the prospect of mixed and polluted data sets, I conclude that synthetic data requires a new sensibility on its impact—especially in the case of technologies linked to the purpose of surveillance, such as algorithmic prediction and computer vision, as these have been shown to be particularly harmful for minorities, such as racialized and gendered subjects (Browne 2015; D'Ignazio and Klein 2020; Phan and Wark 2021). Synthetic data add new layers to the life cycles of data (Kaufman 2020, 2023). With industries on the rise that introduce synthetic data generation in fully simulated 3D environments or synthetic data augmented from existing data sets, new notions of datafication practices are being implemented worthy of scholarly attention. In a similar vein as the other authors in this dialogue section, I conclude that further scrutiny and analysis is needed to understand the (industrial) development and practices around synthetic data. Ridgway and Malevé (in this issue) emphasize in their case study on reverse image search how new potentials of surveillance practices unfold with synthetic data. Future scholarship on synthetic data needs to foster insights into concrete moments like these, into the interaction between AI systems trained with synthetic data and people, policies and governance.

References

Amoore, Louise. 2020. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham, NC: Duke University Press.

Andrejevic, Mark. 2020. "Framelessness." In *Automated Media*, 113–132. London: Routledge

Beduschi, Ana. 2024. Synthetic Data Protection: Towards a Paradigm Change in Data Regulation? *Big Data & Society* 11 (1): https://doi.org/10.1177/20539517241231277.

Browne, Simone. 2015. *Dark Matters: On the Surveillance of Blackness*. Durham, NC: Duke University Press.

D'Ignazio, Catherine, and Lauren F. Klein. 2020. *Data Feminism*. Cambridge, MA: The MIT Press.

De Wilde, Phillippe, Payal Arora, Fernando Buarque, Tik Chan Chin, Mamello Thinyane, Serge Stinckwich, Eleonore Fournier-Tombs, Eleonore, and Tshilidzi Marwala. 2024. *Recommendations on the Use of Synthetic Data to Train AI Models*. Tokyo, JP: United Nations University.

Devinney, Hannah, Jenny Björklund, and Henrik Björklund. 2024. "We Don't Talk About That": Case Studies on Intersectional Analysis of Social Bias in Large Language Models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Bankgok, Thailand, DATES, August 16,* 33–44. Kerrville, TX: Association for Computational Linguistics.

Fitzgerald, Andrew. 2024. Why Synthetic Data Can Never Be Ethical: A Lesson from Media Ethics. *Surveillance & Society* 22 (4): 477–482.

Gitelman, Lisa, ed. 2013. "*Raw Data" Is an Oxymoron*. Cambridge, MA: MIT Press.

Helm, Paula, Benjamin Lipp, and Roser Pujadas. 2024 Generating Reality and Silencing Debate: Synthetic Data as Discursive Device. *Big Data & Society* 11 (2): https://doi.org/10.1177/20539517241249447.

Hoffmann, Anna Lauren. 2020. Terms of Inclusion: Data, Discourse, Violence. *New Media & Society* 23 (12): 3539–3556.

Jacobsen, Benjamin N. 2023. Machine Learning and the Politics of Synthetic Data. *Big Data & Society* 10 (1): https://doi.org/10.1177/20539517221145372.

———. 2024. The Logic of the Synthetic Supplement in Algorithmic Societies. *Theory, Culture & Society* 41 (4): https://doi.org/10.1177/02632764231225768.

Jordon, James, Lukasz Szpruch, Florimon Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. 2022. Synthetic Data—What, Why and How? ArXiv, May 6. http://arxiv.org/abs/2205.03257 [accessed October 23, 2024].

Kaufmann, Mareile. 2020. Vocations, Visions and Vitalities of Data Analysis. An Introduction. *Information, Communication & Society* 23 (14): 1981–1995.

———. 2023. *Making Information Matter: Understanding Surveillance and Making a Difference*. Bristol, UK: Bristol University Press.

Kitchin, Rob. 2021. Data Lives: How Data Are Made and Shape Our World. Bristol, UK: Bristol University Press.

Korenhof, Paulan, Vincent Blok, and Sanneke Kloppenburg. 2021. Steering Representations: Towards a Critical Understanding of Digital Twins. *Philosophy & Technology* 34 (4): 1751–1773.

Mejias, Ulises A., and Nick Couldry. 2019. Datafication. *Internet Policy Review* 8 (4): https://policyreview.info/concepts/datafication.

Munkholm, Johan Lau, and Tanja Wiehn. 2025. Synthetic Data: Servicing Privacy. In B*eyond Privacy: People, Practices, Politics*, edited by Sille Obelitz Søe, Tanja Wiehn, Rikke Frank Jørgensen, and Bjarki Valtýsson. Bristol, UK: Bristol University Press.

Nikolenko, Sergey I. 2021. *Synthetic Data for Deep Learning*. New York: Springer.

Nvidia. N.d. Nvidia Omniverse: The Platform for Developing Openusd Applications for Industrial Digitalization and Generative Physical AI. https://www.nvidia.com/en-us/omniverse/ [accessed October 23, 2024].

Offenhuber, Dietmar. 2024 Shapes and Frictions of Synthetic Data. *Big Data & Society* 11 (2): https://doi.org/10.1177/20539517241249390.

Phan, Thao, and Scott Wark. 2021. Racial Formations as Data Formations. *Big Data & Society* 8 (2): https://doi.org/10.1177/20539517211046377.

Ravn, Louis. 2024. Synthetic Training Data and the Reconfiguration of Surveillant Assemblages. *Surveillance & Society* 22 (4): 460–465.

Ridgway, Reneé, and Nicolas Malevé. 2024. Synthetic Data and Reverse Image Search: Constructing New Surveillant Indexicalities. *Surveillance & Society* 22 (4): 466–471.

Steinhoff, James. 2022. Toward a Political Economy of Synthetic Data: A Data-Intensive Capitalism That Is Not a Surveillance Capitalism? *New Media & Society* 26 (6): https://doi.org/10.1177/14614448221099217.

Susser, Daniel, and Jeremy Seeman. 2024. Critical Provocations for Synthetic Data. *Surveillance & Society* 22 (4): 453–459.

Synthesis.ai. N.d. “Synthesis Scenarios.” https://synthesis.ai/synthesis-scenarios/ [accessed October 23, 2024].

Thylstrup, Nanna Bonde. 2022. The Ethics and Politics of Data Sets in the Age of Machine Learning: Deleting Traces and Encountering Remains. *Media, Culture & Society* 44 (4): 655–671.

Thylstrup, Nanna Bonde, Kristian Bondo Hansen, Mikkel Flyverbom, and Louise Amoore. 2022. Politics of Data Reuse in Machine Learning Systems: Theorizing Reuse Entanglements. *Big Data & Society* 9 (2): https://doi.org/10.1177/20539517221139785.

van Dijck, Jose. 2014. Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology. *Surveillance & Society* 12 (2): 197–208.

West, Sarah Myers. 2020. Redistribution and Rekognition: A Feminist Critique of Algorithmic Fairness. *Catalyst* 6 (2): https://doi.org/10.28968/cftt.v6i2.33043

Zewe, Adam. 2022. In Machine Learning, Synthetic Data Can Offer Real Performance Improvements. MIT News, November 3, 2022. https://news.mit.edu/2022/synthetic-data-ai-improvements-1103 [accessed October 23, 2024].

Zuboff, Shoshana. 2019. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: Public Affairs.