Surveillance & Society

Critical Provocations for Synthetic Data

Daniel Susser and Jeremy Seeman

Volume 22, Number 4, 2024

Open Issue

URI: https://id.erudit.org/iderudit/1115676ar
DOI: https://doi.org/10.24908/ss.v22i4.18335

See table of contents

Publisher(s)

Surveillance Studies Network

ISSN

1477-7487 (digital)

Explore this journal

Cite this document

Susser, D. & Seeman, J. (2024). Critical Provocations for Synthetic Data. *Surveillance & Society*, *22*(4), 453–459. https://doi.org/10.24908/ss.v22i4.18335

Article abstract

Training artificial intelligence (AI) systems requires vast quantities of data, and AI developers face a variety of barriers to accessing the information they need. Synthetic data has captured researchers' and industry's imagination as a potential solution to this problem. While some of the enthusiasm for synthetic data may be warranted, in this short paper we offer critical counterweight to simplistic narratives that position synthetic data as a cost-free solution to every data-access challenge—provocations highlighting ethical, political, and governance issues the use of synthetic data can create. We question the idea that synthetic data, by its nature, is exempt from privacy and related ethical concerns. We caution that framing synthetic data in binary opposition to "real" measurement data could subtly shift the normative standards to which data collectors and processors are held. And we argue that by promising to divorce data from its constituents—the people it represents and impacts—synthetic data could create new obstacles to democratic data governance.

Surveillance & Society

Dialogue | Critical Provocations for Synthetic Data

Daniel Susser

Cornell University, USA
susser@cornell.edu

Jeremy Seeman

University of Michigan, USA
jhseeman@umich.edu

Abstract

Training artificial intelligence (AI) systems requires vast quantities of data, and AI developers face a variety of barriers to accessing the information they need. Synthetic data has captured researchers' and industry's imagination as a potential solution to this problem. While some of the enthusiasm for synthetic data may be warranted, in this short paper we offer critical counterweight to simplistic narratives that position synthetic data as a cost-free solution to every data-access challenge—provocations highlighting ethical, political, and governance issues the use of synthetic data can create. We question the idea that synthetic data, by its nature, is exempt from privacy and related ethical concerns. We caution that framing synthetic data in binary opposition to "real" measurement data could subtly shift the normative standards to which data collectors and processors are held. And we argue that by promising to divorce data from its constituents—the people it represents and impacts—synthetic data could create new obstacles to democratic data governance.

Introduction

More than a decade ago, in a similar moment during a previous period of enthusiasm for new data-driven technologies, danah boyd and Kate Crawford (2012) put forward an incisive set of "Critical Provocations for Big Data." In the same spirit, in this short paper we offer critical provocations for synthetic data, highlighting emerging ethical, political, and governance questions synthetic data raises.

Unlike big data, synthetic data has not (yet) become "mythological" (boyd and Crawford 2012: 663). The term is not as pervasive; it doesn't organize public discourse or structure academic inquiry. But synthetic data has begun to capture the imagination of researchers and industry practitioners, and it has become deeply intertwined with this moment's most powerful technological mythology of all—big data's successor, artificial intelligence (AI). For that reason, it is worth thinking carefully, now, about synthetic data, about its sociotechnical affordances and disaffordances, about the normative assumptions its proponents make, about the rhetorical work the language of "real" vs "synthetic" accomplishes, and about how reliance on synthetic data may impact data governance.

One note of clarification before we begin. Many readers will have encountered discussions about synthetic *content*—i.e., text, images, video, or related media produced by generative AI systems like OpenAI's ChatGPT and DALL-E and intended for human consumption. Synthetic *data* is related—it is a specific type of synthetic content that is often (though not always) produced by AI systems. Rather than being created as system outputs for people to view, however, synthetic data (which can include text, images, numerical values, or other forms of data) is used as inputs to other data processing systems, as a supplement to or replacement for "real" or "original" data sources containing empirical measurements. As a report from the

Susser, Daniel, and Jeremy Seeman. 2024. Critical Provocations for Synthetic Data. *Surveillance & Society* 22 (4): 453-459.
https://ojs.library.queensu.ca/index.php/surveillance-and-society/index | ISSN: 1477-7487

Alan Turing Institute defines it, synthetic data is "generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)"—including, importantly, the task of training AI systems (Jordon et al. 2022).

Our discussion therefore centers on technical actors—AI developers, academic scientists, and others engaged in the production of data-driven technologies—rather than everyday technology users. As we hope to show, however, the use of synthetic data by these actors has significant implications for everyone.

Synthetic Data Does Not Emerge in a Political and Economic Vacuum

Like all technologies, synthetic data introduces specific affordances and disaffordances—it makes some activities easier and some harder—for particular actors in particular contexts (Davis 2020). First, then, it bears asking basic questions about the political economy of synthetic data: Who is using synthetic data? What particular affordances are driving its adoption? How are synthetic data's benefits distributed? Is anyone disadvantaged by it?

Historically, methods for generating synthetic data were developed to enable privacy-preserving analysis of sensitive data (Raghunathan 2021). Today, synthetic data is being put to use for a wider variety of purposes. While it is still a valuable tool for privacy-minded social science and public policy researchers, recent excitement about synthetic data seems to be driven more by its potential for solving problems in the development of artificial intelligence (e.g., Brodsky 2024; Grossman 2021; Wiehn, in this issue). Training AI systems requires vast amounts of data, and AI companies are encountering a number of obstacles in their pursuit of it. Despite ubiquitous digital surveillance, many data are tightly controlled—either for privacy reasons (e.g., tax data and sensitive medical data held in electronic health records, and data collected and controlled by governments and public institutions) or as a means of protecting competitive advantage (the data controlled by private firms, and the insights contained within it, are often one of a firm's main sources of value). Data collection is highly uneven—in any context there is likely to be more data available about some groups than others—resulting in AI systems with uneven performance. And data often exist in formats that are incompatible with the specific machine learning models powering AI, making data difficult to utilize even when they are abundant and easily available.

Proponents of synthetic data believe that it affords solutions to many, if not all, of these problems, enabling more shareable, representative, and interoperable training data for AI, thus making the resulting systems more useful, equitable, and accessible (Savage 2023). Synthetic data could afford more data sharing, proponents argue, by promising (rightly or wrongly, as we discuss next) to ease worries about privacy. It could, theoretically, afford more representative data by offering tools for creating data related to groups about whom little measurement data exists—e.g., demographic minorities or patients with rare medical conditions. And it could afford more interoperable data by enabling the automatic generation of datasets in the specific formats most suitable to training particular models, providing greater access to the resources needed to develop new AI systems.

It is equally important, however, to interrogate synthetic data's disadvantages and disaffordances. While it promises to make life easier for AI developers, synthetic data threatens to make things difficult for others. As we discuss in what follows, naive or simplistic notions that, by its nature, synthetic data is exempt from privacy and related ethical concerns could legitimize the construction of new systems of surveillance and social sorting. More broadly, framing synthetic data in binary opposition to "real" measurement data could function to subtly shift the normative standards to which data collectors and processors are held, exacerbating existing wrongs and harms experienced by data subjects. Finally, by promising to divorce data from its constituents—the people it represents and impacts—synthetic data may create new obstacles to democratic data governance.

Using Synthetic Data Is Not Necessarily More Ethical or Private

The "synthetic" label might suggest that synthetic data is artificial, unconnected to real people, places, or things, and much of the enthusiasm for using synthetic data stems from the belief that sharing or analyzing it thus poses no ethical or privacy risks (Jacobsen 2023). In fact, like synthetic oil or synthetic meat, synthetic data is a highly processed or carefully cultivated version of "real" measurement data, and its use can affect people's rights and interests, depending on "how the sausage is made." Moreover, ethical concerns about data-driven systems are not limited to questions about where the data that fuels them come from or how those data are constructed. Concerns extend (or ought to extend) to questions about the broader implications of such systems for the individuals, groups, institutions, and social orders they interact with (Susser 2022a).

The idea that using synthetic data automatically circumvents all privacy issues reflects both mistaken assumptions about synthetic data and mistaken assumptions about privacy. First, "synthetic" datasets necessarily retain some information about the real-world measurement datasets they mimic—if they didn't, it wouldn't be useful to analyze them. Thus, sharing and processing synthetic data carries some risk of disclosing information about real data subjects, depending on how much and what kind of information about the original measurement data is preserved (Bellovin, Dutta, and Reitinger 2019). "Partially synthetic" datasets contain a mix of synthetically generated records and actual measurement records, the latter of which can leak during data processing. And even some "fully synthetic" datasets generated by ML systems can overfit their training data, exposing information about real-world data subjects. The UK Office of National Statistics describes the range of disclosure risks posed by synthetic data on a spectrum from "structural synthetic datasets," which preserve only high-level organizational features of measurement datasets (e.g., the types of variables they contain), to "replica datasets," which mirror many of the complex statistical relationships between variables (Bates et al. 2019).

Second, synthetic data tools—like many "privacy-enhancing technologies" (PETs)—aim to facilitate inferences (i.e., learning) about populations, while preventing inferences about the individuals that comprise them. But privacy protects against more than just identification or the disclosure of individual personal information (especially given that the line between inferences about individuals and inferences about very small subpopulations can quickly blur in practice) (Seeman and Susser 2024). Privacy is the defense against inappropriate information flows generally, and against the harms—such as censorship, manipulation, and social sorting—unconstrained information flows can cause (Nissenbaum 2009). Whether synthetic data is "privacy-preserving" is therefore a function not only of how well it obscures the identities of individual data subjects but also whether its affordances and disaffordances align with shared values. This distinction is critical because obscuring data subject identities can continue to enable surveillance (Yew, Qin, and Venkatasubramanian 2024).

As discussed above, much of the excitement around synthetic data stems from its potential to help AI developers overcome obstacles to acquiring the data they need to train AI systems. And while many of those systems will undoubtedly bring widely shared social goods—for example, AI systems that optimize energy use or accelerate drug discovery—others are just as likely to increase and deepen surveillance. As Louis Ravn argues in this issue, synthetic data is being incorporated into larger "surveillant assemblages." For example, face recognition and related AI-powered computer vision technologies promise to identify people in public spaces, analyze their behavior, and anticipate outbreaks of protest and social unrest (Delussu, Putzu, and Fumera 2024). Similarly, workplace tracking technologies analyze employee behavior to help managers predict and quell worker organizing. While processes like these may be ethically suspect with "original" data, the same processes using "synthetic data" may (wrongly) attract less ethical scrutiny or exemption from regulation. To the extent that synthetic data is used to develop and refine such systems it will fuel surveillance rather than protect against it (see also Ridgway and Malevé, in this issue).

The "Real" vs "Synthetic" Distinction Is More Rhetorical than Ontological

Describing some data as "synthetic" implies the existence of "real" or "natural" counterparts. But critical data studies scholars have long emphasized that data—all data—are "made not found" (Wiggins and Jones 2023). Which is to say, the structure and meaning of data are always, necessarily artifacts of the human interests, perspectives, decisions, technical capabilities, and evaluative standards that shape the process of imagining, capturing, sharing, and interpreting them. No data are "entirely raw," as Lisa Gitelman and Virginia Jackson (2013: 2) put it, "the data are always already 'cooked.'"

Yet, as Geoffrey Bowker (2013: 168) argues, designating some data as "raw" and other data as (e.g.) "processed"—like distinguishing between "natural" and "social" phenomena—is "politically and philosophically powerful." As a rhetorical move, it carries implicit assumptions about each of the two terms and about how they are, or ought to be, related. Similarly, while the distinction between "real" and "synthetic" data seems to suggest differences in their fundamental nature, it is perhaps more helpful to focus on the distinction's rhetorical effects and their normative implications. If all data are, in some sense, the outcomes of "synthesis," what does designating some data as "real" and some data as "synthetic" accomplish?

The language of "synthetic" data suggests such data are artificial, constructed, unconnected to real people with rights and interests, while "real" data are natural, found, intrinsically bound to their source. As we saw above, this implies that collecting, analyzing, and using synthetic data carries different risks than doing the same with "real" data—perhaps no risks at all. In this way, one effect of the real/synthetic distinction is to subject each side to different normative standards. "Real" data comes laden with privacy expectations, consent procedures, data security practices, and other mainstays of responsible data collection and use. Synthetic data, one is led to assume, is free from these expectations. While using synthetic data in place of measurement data in certain application contexts may indeed help to mitigate some ethics and privacy risks, however, that has to be evaluated in each case rather than assumed in advance.

At a deeper level, emphasizing the synthetic dimension of some data functions to draw attention away from the end-to-end data-making processes through which all data are produced, reducing these nuances of data-making to a real/synthetic binary. As a result, the "synthetic" data label frames questions about data governance—how to use synthetic data ethically and responsibly—in terms of this relationship. The question becomes, to what extent does synthetic data successfully stand in for real, "ground truth" datasets? Questions about how such "ground truth" was established in the first place, or the functions it serves, are put to the side.

If we recognize synthesis as simply one step in the data-making process, however, then we can see that these two questions are inextricably intertwined. For example, assessing the clinical feasibility of synthetic electronic health records for medical research requires evaluating the interplay between the quality of the original patient data (e.g., whether the original records were representative of the clinical population, how precisely the medical instruments captured their measurements, etc.) and the synthetic data generation process (e.g., how well features of the original data were preserved by the synthetic data generating algorithm). Unless we attend carefully to these complexities, reliance on synthetic data risks diverting us away from the challenges of producing and using data—all data—ethically and in service of social goods. We can't simply hand off the difficult work of data-making (collection, curation, processing, dissemination) to synthetic data algorithms.

Synthetic Data Solutionism Creates New Challenges for Democratic Data Governance

Synthetic data can serve different statistical purposes: it can expand access to sensitive or confidential data, augment previously observed data with additional statistical properties, simulate measurements of

unobserved or unobservable qualities, and more. Regardless of why one might choose to generate synthetic data, though, any synthetic data product necessarily contains no more empirical information than is available from accessing the original data. Synthetic data is, by definition, less information-rich than the data it mimics.

To technical readers, this should be obvious—such ideas date back to the origins of information theory. Nevertheless, it has not dampened the appeal of what we might call synthetic data solutionism.[1] Tech executives continue to tout synthetic data as a panacea for the scaling woes of large generative AI models (Castellanos 2021). Healthcare systems look to synthetic data as a key tool for studying novel and emergent diseases (Chen et al. 2021). The finance industry expects that synthetic data will "transform" fraud detection and credit risk modeling (Ribeiro 2024).

Synthetic data solutionism mistakes a data quality problem for a data quantity problem—the key to overcoming any AI performance challenge, it's believed, is simply more data. Focusing on the amount of available data rather than questions about its quality can have catastrophic consequences; a now-canonical example is "model collapse," wherein large-scale models trained recursively on synthetic data and other synthetic content see rapidly degrading performance (Shumailov et al. 2024). While excitement about synthetic data ranges across a variety of data-driven endeavors, from developing computer vision systems to quantitative social science research, data users ought to think carefully about how suitable synthetic data is for specific purposes—i.e., whether using less informative data is epistemically and ethically appropriate—and how its utility can be determined in specific contexts.

Further, synthetic data solutionism could exacerbate a more subtle problem, familiar from recent discussions about data governance. Data ethics and governance are about more than merely preventing bad behavior; they aim to encourage good data practices—collecting, analyzing, utilizing, and sharing data in ways that advance public values (Susser 2022b; Viljoen 2022). As many have argued, central to that project must be efforts to make data and data governance more democratic (Viljoen 2021; Cuéllar and Huq 2022). That is, to engage broad publics in the creation and governance of data-driven systems (especially the communities most likely to be impacted by them), and to make such systems responsive to those publics. By distancing data as much as possible from real data subjects, synthetic data solutionism pushes in the opposite direction, encouraging a technocratic attitude indifferent to data's constituents—to the people it represents and impacts. Rather than create new ways to include more people in governing our data-driven social order, synthetic data aims to minimize the role real people play in it (Susser et al. 2024).

Conclusion

Synthetic data can help solve genuine problems. The provocations, above, are not meant to detract from the good it could bring, but rather to provide ballast—critical counterweight to the uncritical exuberance for synthetic data that seems to be growing in academia, industry, and elsewhere. By highlighting how the use of synthetic data could create ethical, political, and governance problems, we hope to encourage skepticism toward simplistic narratives that position synthetic data as a cost-free solution to every data-access challenge, and to help data practitioners think constructively about how to use it to advance public values. To close, we offer some guiding prompts for navigating these challenges.

First, one should zoom out and think carefully at the macro level about the social, political, and economic contexts in which synthetic data-driven technologies are being developed and implemented. Who are the actors involved? What are their goals? Who stands to benefit or be empowered by these tools? To whose detriment? Synthetic data is not absolution—harmful technologies, such as AI-driven surveillance systems, will harm regardless of whether they are trained on synthetic data or "real" measurement data. By contrast,

[1] Here, we are building on Evgeny Morozov's (2014) idea of "technological solutionism."

in cases where the only obstacle to achieving a genuine social good is the risk of exposing individual personal information—e.g., in certain efforts to make academic science more open and reproducible—if deployed thoughtfully, synthetic data could be a valuable tool.

Second, one should zoom in and interrogate the specific role synthetic data is meant to play—the problem it is intended to solve—in the project at hand. Is this really a data quantity problem or is it a data quality problem? Why is access to data an issue? For whom does synthetic data expand access, and to what ends? If existing datasets misrepresent certain subpopulations, what caused those disparities? What aspects of the data-making process are obscured by relying on synthetic data? How will using less information-rich datasets in this application context affect outcomes? Sometimes barriers to accessing information exist for good reason; in other contexts, they result from discrimination or neglect. Avoiding synthetic data solutionism requires recognizing when technical data and data access problems are actually symptoms of deeper, underlying social and political injustice and grappling carefully with the potential for data-driven technologies to improve or worsen them.[2]

Finally, one should consider the role of data's constituents in data governance—does using synthetic data make data production more responsive to the needs and interests of the people likely to be impacted by it, or does it insulate the data-making process from public scrutiny and oversight? Democratic governance can be difficult and time-consuming, and in some cases, using synthetic data could help. For example, rather than asking minority or marginalized communities to spend time and energy fixing data representation problems, carefully developed synthetic data could minimize that burden. In other cases, however, relying on synthetic data simply serves as a means and justification for excluding data's constituents from data-making practice. Democratic data governance demands approaching this delicate balancing act head-on, delegating difficult data work to experts when it serves public interests and recognizing when broad participation in data-making is indispensable to realizing shared values.

References

Bates, Andrew G., Iva Špakulová, Iain Dove, and Andrew Mealor. 2019. ONS Methodology Working Paper Series Number 16—Synthetic Data Pilot. *UK Office for National Statistics.* https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot [accessed November 20, 2024].

Bellovin, Steven, Preetam K. Dutta, and Nathan Reitinger. 2019. Privacy and Synthetic Datasets. *Stanford Technology Law Review* 22: 1–52.

boyd, danah, and Kate Crawford. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15 (5): 662–679.

Brodsky, Sascha. 2024. Examining Synthetic Data: The Promise, Risks and Realities. *IBM,* August 20. https://www.ibm.com/blog/ai-synthetic-data/ [accessed November 20, 2024].

Bowker, Geoffrey. 2013. Data Flakes: An Afterward to "Raw Data" is an Oxymoron. In *"Raw Data" is an Oxymoron*, edited by Lisa Gitelman, 167–171. Cambridge, MA: The MIT Press.

Castellanos, Sara. 2021. Fake It to Make It: Companies Beef up AI Models With Synthetic Data. *The Wall Street Journal,* July 23. https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601 [accessed November 20, 2024].

Chen, Richard J., Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. 2021. Synthetic Data in Machine Learning for Medicine and Healthcare. *Nature Biomedical Engineering* 5 (6): 493–497.

Cuéllar, Mariano-Florentino, and Aziz Z. Huq. 2022. The Democratic Regulation of Artificial Intelligence. *Knight First Amendment Institute.* https://knightcolumbia.org/content/the-democratic-regulation-of-artificial-intelligence [accessed November 20, 2024].

Davis, Jenny. 2020. *How Artifacts Afford: The Power and Politics of Everyday Things*. Cambridge, MA: MIT Press.

Delussu, Rita, Lorenzo Putzu, and Giorgio Fumera. 2024. Synthetic Data for Video Surveillance Applications of Computer Vision: A Review. *International Journal of Computer Vision* 132: 4473–4509.

[2] For a helpful discussion about navigating a version of this challenge in the context of algorithmic fairness, see Green (2022).

Gitelman, Lisa, and Virginia Jackson. 2013. Introduction. In "*Raw Data" is an Oxymoron,* edited by Lisa Gitelman, 1–14. Cambridge, MA: The MIT Press.

Green, Ben. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology* 35 (90): 1–32.

Grossman, Gary. 2021. How Synthetic Data Could Save AI. *VentureBeat*, March 20. https://venturebeat.com/ai/how-synthetic-data-could-save-ai/ [accessed November 20, 2024].

Jacobsen, Benjamin. 2023. Machine Learning and the Politics of Synthetic Data. *Big Data & Society* 10 (1): 1–12.

Jordon, James, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, Adrian Weller. 2022. Synthetic Data—What, Why and How? *The Alan Turing Institute.* https://arxiv.org/abs/2205.03257 [accessed November 20, 2024].

Morozov, Evgeny. 2014. *To Save Everything, Click Here: The Folly of Technological Solutionism.* New York: PublicAffairs.

Nissenbaum, Helen. 2009. *Privacy as Contextual Integrity: Technology, Policy, and the Integrity of Social Life.* Stanford, CA: Stanford University Press.

Raghunathan, Trivellore E. 2021. Synthetic Data. *Annual Review of Statistics and Its Application* 8: 129–140.

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

Ribeiro. Gonçalo. 2024. Synthetic Data Applications in Finance. *Forbes,* April 3. https://www.forbes.com/councils/forbestechcouncil/2024/04/03/synthetic-data-applications-in-finance/ [accessed November 20, 2024].

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

Savage, Neil. 2023. Synthetic Data Could Be Better than Real Data. *Nature*, April 27. https://www.nature.com/articles/d41586-023-01445-8 [accessed November 20, 2024].

Seeman, Jeremy, and Daniel Susser. 2024. Between Privacy and Utility: On Differential Privacy in Theory and Practice. *ACM Journal on Responsible Computing* 1 (1): 1–18.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI Models Collapse When Trained on Recursively Generated Data. *Nature* 631: 755–759.

Susser, Daniel. 2022a. Decision Time: Normative Dimensions of Algorithmic Speed. In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Korea, June 21–24,* 1410–1420. New York: Association for Computing Machining.

———. 2022b. Data and the Good? *Surveillance & Society* 20 (3): 297–301.

Susser, Daniel, Daniel S. Schiff, Sara Gerke, Laura Y. Cabrera, I. Glenn Cohen, Megan Doerr, Jordan Harrod, Kristin Kostick-Quenet, Jasmine McNealy, Michelle N. Meyer, W. Nicholson Price II, and Jennifer K. Wagner. 2024. Synthetic Health Data: Real Ethical Promise and Peril. *Hastings Center Report* 54 (4): 1–6.

Viljoen, Salomé. 2021. A Relational Theory of Data Governance. *Yale Law Journal* 131 (2): 573–654.

———. 2022. An Argument for Positive Political Theories of Data Governance. *Georgetown Law Technology Review* 6: 464–472.

Wiehn, Tanja. 2024. Synthetic Data: From Data Scarcity to Data Pollution. *Surveillance & Society* 22 (4): 472–476.

Wiggins, Chris, and Matthew Jones. 2021. *How Data Happened: A History from the Age of Reason to the Age of Algorithms.* New York: W. W. Norton & Company.

Yew, Rui-Jie, Lucy Qin, and Suresh Venkatasubramanian. 2024. You Still See Me: How Data Protection Supports the Architecture of ML Surveillance. Arxiv, October 6. https://arxiv.org/abs/2402.06609v3 [accessed November 20, 2024].