Surveillance & Society

Synthetic Data, Synthetic Media, and Surveillance

Aaron Martin et Bryce Clayton Newell

Volume 22, numéro 4, 2024

Open Issue

URI : https://id.erudit.org/iderudit/1115675ar
DOI : https://doi.org/10.24908/ss.v22i4.18334

Aller au sommaire du numéro

Éditeur(s)

Surveillance Studies Network

ISSN

1477-7487 (numérique)

Découvrir la revue

Citer ce document

Martin, A. & Newell, B. (2024). Synthetic Data, Synthetic Media, and Surveillance. *Surveillance & Society*, *22*(4), 448–452. https://doi.org/10.24908/ss.v22i4.18334

Surveillance & Society

Dialogue | Synthetic Data, Synthetic Media, and Surveillance

Aaron Martin

University of Virginia, USA
xqt3xa@virginia.edu

Bryce Clayton Newell

University of Oregon, USA
bcnewell@uoregon.edu

Public and scholarly interest in the related concepts of *synthetic data* and *synthetic media* has exploded in recent years. From issues raised by the generation of synthetic datasets to train machine learning models to the public-facing, consumer availability of artificial intelligence (AI) powered image manipulation and creation apps and the associated increase in synthetic (or "deepfake") media, these technologies have shifted from being niche curiosities of the computer science community to become topics of significant public, corporate, and regulatory import. They are emblematic of a "data-generation revolution" (Gal and Lynskey 2024: 1091) that is already raising pressing questions for the academic surveillance studies community.

Defining and distinguishing between synthetic data and synthetic media is important. It helps to understand the key differences in terms of *inputs* and *outputs*. 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. In other words, 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, increasingly, the task of training AI systems (Jordon et al. 2022). Often, these synthetic datasets are designed to simulate and replace the need for actual empirical observation by mimicking the distribution of data "corresponding to real-world phenomena" (Offenhuber 2024: 1; Ridgway and Malevé, in this issue). Alternatively, synthetic outputs include text, images, video, or other content that are often (though not always) produced by generative AI systems and intended for people to consume. Herein, we refer to these forms of synthetic content as "synthetic media," though these are pejoratively referred to as AI "slop" when they are deemed to be of low quality (Hoffman 2024).

The recent mass popularization of generative AI has made synthetic data the subject of ongoing mainstream media reporting (see, for example, Bhatia 2024). Public-facing scholars have critically described it as the "diet soda" of data—branded in a way to convince companies of its (debatable) value (Gallon 2024). Still, industries including banking (JPMorgan 2024) and healthcare (Giuffrè and Shung 2023) are embracing the possibilities afforded by these innovations, in part due to "the sensitive nature of the data [in these sectors], or where the costs to acquire real data would be prohibitive," as Fitzgerald (in this issue) points out in his contribution to this Dialogue.

Following increased business interest in and use of synthetic data, we are also witnessing growing regulatory, legislative, and policy activity in this area: in March 2024, the data protection authorities of the G7 jointly published a "use case" to demonstrate the privacy value of synthetic data (ICO 2024). Legislators

Martin, Aaron, and Bryce Clayton Newell. 2024. Synthetic Data, Synthetic Media, and Surveillance. *Surveillance & Society* 22 (4): 448-452.
https://ojs.library.queensu.ca/index.php/surveillance-and-society/index | ISSN: 1477-7487

in the key US states of New York and California have pursued rules that would mandate additional transparency requirements for synthetic media in different contexts including advertising (DataGuidance 2024a; DataGuidance 2024b). And authorities in geopolitically important countries like China are taking the lead in developing regulations for "deep synthesis" (i.e., synthetically generated content) as part of their broader AI regulatory strategies (Sheehan 2023). "Synthetic data protection" has also entered our lexicon (Beduschi 2024). Insofar as regulation might be seen as having a legitimating effect for new technologies, this flurry of policy interest might lead one to conclude that synthetic data and synthetic media are innovations whose time has finally come even if, as Susser and Seeman (in this issue) note in this Dialogue, they are not yet "mythology" like their big data predecessor.

Within surveillance studies scholarship, Fussey (2022: 348) has argued that synthetic media is one of several "issues of urgent societal and planetary concern" and that it has "arguably never been more important" for surveillance studies "researchers to understand these dynamics and complex processes, evidence their implications, and translate esoteric knowledge to produce meaningful analysis." Yet, while fields adjacent to surveillance studies have begun to explore the ethical risks of synthetic data (see, for example, Whitney and Norman 2024), we currently perceive a lack of attention to the surveillance implications of synthetic data and synthetic media in published literature within our field. In response, this Dialogue is designed to help promote thinking and discussion about the links and disconnections between synthetic data, synthetic media, and surveillance.

Scholars working within the field of surveillance studies have grappled with the implications of AI. As the implications of generative AI have become more front-and-center in recent years, *Surveillance & Society* hosted a special issue on AI and surveillance in 2023 (see vol. 21 [3]: 236–287). Yet, surveillance scholarship has also been called out for lacking historical analyses of the links between AI and surveillance, even as "the stakes of an ahistorical analysis" in this domain "are especially high" (Gluck-Thaler 2023: 260). Surveillance scholarship has also not yet engaged as deeply as it should with questions surrounding the creation and use of synthetic data and synthetic datasets, and the implications these technologies and practices have for surveillance in contemporary society.

Beyond simply a paucity of scholarship, we also see researchers arguing that synthetic data operates outside the logics and concerns of surveillance and privacy. For example, Steinhoff (2024: 3291) writes that "synthetic data is data which is not collected via surveillance; rather it is 'produced artificially'" (quoting Nikolenko 2021: v). Regulators make similarly curious claims: the Personal Data Protection Commission of Singapore (the country's privacy regulator) describes synthetic data as "fictitious data" in its proposed guide on synthetic data generation (PDPC 2024), suggesting that these data are somehow imaginary. These sorts of claims raise intriguing ontological questions about the status of synthetic data(sets) as data or information and about the (lack of) syntactic connections between source (or "real") data and synthetically generated data. After all, much synthetic data draws from, or is at least designed to replicate relationships and distributions that exist in, real world datasets (Lee 2024: 16; Lucini 2021). Can we say that forms of synthetic data retain links to human beings (or data subjects) in ways that generate surveillance and privacy concerns, or is synthetic data really a solution to the surveillance- and privacy-related problems of the past? In other words, and to extend the diet soda metaphor, can surveillance based on synthetic data be understood as "surveillance lite"? These are crucial questions that need more robust answers within the field. Thus, the Dialogue papers in this issue aim to begin to chart a path for theory and research at the intersection of synthetic technologies and surveillance studies.

Problematizing Syntheticity for Surveillance Studies

Susser and Seeman open the Dialogue with a series of provocations meant to challenge the ethical, political, and governance narratives surrounding synthetic data. They encourage us to take synthetic data's political economy seriously by critically analyzing its users, drivers, and consequences—which actors stand to

benefit from the mainstreaming of synthetic datasets? To what end? And who loses out as a result of the synthetic turn? The authors also challenge the accepted knowledge that synthetic data is necessarily more private than "real" data, or that its use in surveillance systems automatically renders these systems more ethical—themes that other contributors further expound on elsewhere in the Dialogue. Susser and Seeman push back on the idea that synthetic data is ontologically distinct from "real" data and implore us to explore the rhetorical effects of these distinctions as well as their normative implications. Lastly, they question whether "synthetic data solutionism"—which prioritizes data quantity over data quality, and which further distances data from the people and world it purports to represent—might require new governance solutions.

Next, Ravn's contribution helps us begin (re)theorizing synthetic data for surveillance scholarship. Ravn conceptualizes synthetic data in terms of Haggerty and Ericson's (2000) surveillant assemblage: "From this perspective, synthetic training data are neither good, nor bad, nor neutral in relation to surveillance; instead, they may be seen to intensify or dissipate the potentialities of the surveillant assemblages in which they figure." Through the assemblage lens, his piece engages empirically to explore both the problematic and anticipatory dimensions of using synthetic data to train video surveillance applications.

Third, drawing on insights from the history of photography, Ridgway and Malevé explore the use of synthetic data in the context of "fake" images and reverse image search. They are particularly interested in images that have a strong claim to indexicality—i.e., those "presenting themselves as an emanation of their referent." Focusing on PimEyes, a popular reverse search engine for facial images, they ponder what it means when a search using fake images generates results that may or may not refer to "real" people: "Synthetic data both expands and complicates technologies of identification…. To be able to establish a correlation between an image (the image submitted by the user) and another (the search result), the reverse image search engine is forced to negotiate the possibility of a synthetic image and therefore translate from one ontology of the photograph (indexicality) to another (simulation). The same goes for the human user who needs to critically appraise the results."

Next, Wiehn highlights concern about data's scarcity and the possible contamination and pollution of synthetic datasets as entry points to re-energize contemporary debates within surveillance studies and critical data studies. Specifically, she calls on us to treat synthetic data not as "inert" but rather as "living information, inherently political and always on a threshold of becoming." Wiehn argues that concerns about data scarcity, contamination, and pollution become especially pressing in the context of surveillance applications that have been shown to have disproportionate and harmful racialized and gendered impacts, and that "synthetic data do not place AI models out of the realm of risk."

Finally, Fitzgerald closes the Dialogue with a perspective from media ethics—one that is arguably even more provocative than Susser and Seeman's opening salvo. Drawing on a normative stance whereby being ethical means being accountable, Fitzgerald argues that "that the use of synthetic data can *never* be ethical" (emphasis added). His critique is, in part, a reaction to certain brands of AI ethics, which entail "vague procedures for stakeholder input or voluntary ethics guidelines that disproportionately come from corporations themselves." He asserts that the growing use of synthetic data, which purports to be ethically informed, in fact "intensifies a pre-existing lack of accountability inherent within automated systems more generally, and through this, entrenches and compounds surveillant practices." He also warns that "the biggest danger of synthetic data is the broader normalization of its use in automated or semi-automated systems, and the migration of models—and mindsets—from commercial to state deployment in law enforcement and military contexts" (cf. Sharon 2021).

Future Research Directions

This Dialogue raises a number of important issues related to synthetic data and synthetic media that merit further empirical research and theoretical exploration. Here we outline just a few avenues.

First, surveillance studies scholars ought to grapple more, and more deeply, with the relationship(s) between synthetic data and surveillance. To what extent (if any) does the synthetic "data revolution" (Gal and Lynskey 2024: 1094) avoid or sidestep surveillance practices or represent something like "surveillance lite"? Answering these questions may require surveillance scholars to engage more directly with critical data studies and information science scholarship, insofar as scholars in these fields are tackling definitional and ontological questions about the nature and forms of data and information.

Second, more research is needed to assess the drivers for, and the broad effects of, the use of synthetic data in surveillance-intensive contexts and sectors. Where do we see synthetic data being explored or deployed to enrich surveillance systems? Which actors are promoting the adoption of synthetic data-driven surveillance, and what are their motives? Insofar as synthetic data is being used operationally for surveillance purposes, in what ways has it shaped the subjectivities and lived experiences of surveillance? And what does its popularization mean for resistance? In contrast, for which surveillance applications has the use of synthetic data failed to take off, and why? Is a new synthetic-surveillance paradigm almost upon us or not? The field would also benefit from analyses and empirical research on these questions from researchers based in Majority World countries (or what is often referred to as the Global South).

Third, building on the contribution by Ridgway and Malevé (in this issue), future research could also explore the various ways in which AI technologies, including generative AI, are shaping the development and (mis)use of synthetic identities. Which actors are using AI to create and propagate fake identities and to what end (Biddle 2024)? How are these technologies challenging the effectiveness of biometric identification and surveillance tools? What are policy makers and industry groups doing in response to synthetic identity fraud (cf. FS-ISAC 2024). And what are the broad societal consequences of synthetic identities for trust and personhood in the digital age?

Lastly, what are the opportunities for surveillance regulation and data governance in our emerging synthetic society (van der Sloot 2024)? If, in fact, synthetic data does not constitute personal data and is thus largely outside the scope of data protection law, how do we ensure that surveillance activities and other critical decisions made on the basis of synthetic data are effectively governed?

The ethical, social, and political implications of this data revolution are significant, and these developments raise myriad questions of "urgent societal and planetary concern" (Fussey 2022: 348) that sit squarely within the scope of surveillance research. We hope that this Dialogue can help spur and inform additional scholarship, discussion, and debate within the field in the coming years.

Acknowledgments

Aaron Martin's work is supported by a grant from the Robert Bosch Stiftung GmbH as well as by the National Security Data & Policy Institute at UVA.

References

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

Bhatia, Aatish. 2024. When A.I.'s Output Is a Threat to A.I. Itself. *The New York Times*, August 26. https://www.nytimes.com/interactive/2024/08/26/upshot/ai-synthetic-data.html [accessed November 4, 2024].

Biddle, Sam. 2024. The Pentagon Wants to Use AI to Create Deepfake Internet Users. *The Intercept*, October 17. https://theintercept.com/2024/10/17/pentagon-ai-deepfake-internet-users/ [accessed November 4, 2024].

DataGuidance. 2024a. New York: Bill Requiring Advertisements to Disclose Synthetic Data Referred to Assembly Committee. https://www.dataguidance.com/news/new-york-bill-requiring-advertisements-disclose [accessed November 4, 2024].

———. 2024b. California: Bill on Watermarks for Synthetic Content Passes Committee on Appropriations. https://www.dataguidance.com/news/california-bill-watermarks-synthetic-content-passes [accessed November 4, 2024].

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

FS-ISAC. 2024. Deepfakes in the Financial Sector: Understanding the Threats, Managing the Risks. Financial Services Information Sharing and Analysis Center. October https://www.fsisac.com/hubfs/Knowledge/AI/DeepfakesInTheFinancialSector-UnderstandingTheThreatsManagingTheRisks.pdf.

Fussey, Pete. 2022. Seeing Surveillance: Twenty Years of Surveillance & Society. *Surveillance & Society* 20 (4): 346–352.

Gal, Michal S., and Orla Lynskey. 2024. Synthetic Data: Legal Implications of the Data-Generation Revolution. *Iowa Law Review* 109 (3): 1087–1156.

Gallon, Kim. 2024. The "Diet Soda" of Data. Public Books, September 18. https://www.publicbooks.org/the-diet-soda-of-data/ [accessed November 4, 2024].

Giuffrè, Mauro, and Dennis L. Shung. 2023. Harnessing the Power of Synthetic Data in Healthcare: Innovation, Application, and Privacy. npj Digital Medicine 6 (1): 1–8.

Gluck-Thaler, Aaron. 2023. Surveillance Studies and the History of Artificial Intelligence: A Missed Opportunity? *Surveillance & Society* 21 (3): 259–268.

Haggerty, Kevin D., and Richard V. Ericson. 2000. The Surveillant Assemblage. *British Journal of Sociology* 51 (4): 605–622.

Hoffman, Benjamin. 2024. First Came "Spam." Now, With A.I., We've Got "Slop." *The New York Times*, June 11. https://www.nytimes.com/2024/06/11/style/ai-search-slop.html [accessed November 4, 2024].

ICO. 2024. G7 DPAs' Emerging Technologies Working Group Use Case Study on Privacy Enhancing Technologies. https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/case-studies/g7-dpas-emerging-technologies-working-group-use-case-study-on-privacy-enhancing-technologies/ [accessed October 23, 2024].

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, May 6. https://arxiv.org/abs/2205.03257 [accessed November 18, 2024].

JPMorgan. 2024. Synthetic Data. https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/synthetic-data [accessed October 23, 2024].

Lee, Peter. 2024. Synthetic Data and the Future of AI. SSRN, March 26. https://ssrn.com/abstract=4722162.

Lucini, Fernando. 2021. The Real Deal About Synthetic Data. MIT Sloan Management Review, October 20. https://sloanreview.mit.edu/article/the-real-deal-about-synthetic-data/ [accessed November 18, 2024].

Nikolenko, Sergey I. 2021. *Synthetic Data for Deep Learning*. Cham, CH: Springer.

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

PDPC. 2024. Proposed Guide on Synthetic Data Generation. Personal Data Protection Commission of Singapore. https://www.pdpc.gov.sg/help-and-resources/2024/07/proposed-guide-on-synthetic-data-generation [accessed November 4, 2024].

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.

Sharon, Tamar. 2021. Blind-Sided by Privacy? Digital Contact Tracing, the Apple/Google API and Big Tech's Newfound Role as Global Health Policy Makers. *Ethics and Information Technology* 23 (1): 45–57.

Sheehan, Matt. 2023. China's AI Regulations and How They Get Made. Carnegie Endowment for International Peace, July 10. https://carnegieendowment.org/research/2023/07/chinas-ai-regulations-and-how-they-get-made [accessed November 4, 2024].

Steinhoff, James. 2024. Toward a Political Economy of Synthetic Data: A Data-Intensive Capitalism That Is Not a Surveillance Capitalism? *New Media & Society* 26 (6): 3290–3306.

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

van der Sloot, Bart. 2024. *Regulating the Synthetic Society: Generative AI, Legal Questions and Societal Challenges*. Oxford, UK: Hart Publishing.

Whitney, Cedric Deslandes, and Justin Norman. 2024. Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, June 3–6,* 1733–1744. New York: Association for Computing Machinery.

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