Original Research Article

BIG DATA & SOCIETY

Big Data & Society
April–June: 1–13

DOI: 10.1177/20539517251318289
journals.sagepub.com/home/bds
Sage

The ontological politics of synthetic data: Normalities, outliers, and intersectional hallucinations

Francis Lee[1,2], **Saghi Hajisharif**[3] and **Ericka Johnson**[4]

Abstract

Synthetic data is increasingly used as a substitute for real data due to ethical, legal, and logistical reasons. However, the rise of synthetic data also raises critical questions about its entanglement with the politics of classification and the reproduction of social norms and categories. This paper aims to problematize the use of synthetic data by examining how its production is intertwined with the maintenance of certain worldviews and classifications. We argue that synthetic data, like real data, is embedded with societal biases and power structures, leading to the reproduction of existing social inequalities. Through empirical examples, we demonstrate how synthetic data tends to highlight majority elements as the "normal" and minimize minority elements, and that the slight changes to the data structures that create synthetic data will also inevitably result in what we term "intersectional hallucinations." These hallucinations are inherent to synthetic data and cannot be entirely eliminated without compromising the purpose of creating synthetic datasets. We contend that decisions about synthetic data involve determining which intersections are essential and which can be disregarded, a practice which will imbue these decisions with norms and values. Our study underscores the need for critical engagement with the mathematical and statistical choices in synthetic data production and advocates for careful consideration of the ontological and political implications of these choices during curatorial style production of synthetic structured data.

Keywords

Synthetic structured data, ontological politics, intersectionality, data bias, classification, data ethics

The ontological politics of synthetic data

Today, synthetic data is increasingly used as a substitute for "raw" or "real" data—for instance, when there are ethical or legal reasons to address privacy and anonymity, when data sharing is restricted for regulatory or secrecy reasons, or when the amount of existing data is too small or has significant gaps.[1] The use of synthetic data brings to the fore a number of questions about how it is intertwined with the politics of classification and valuation of people, objects, and phenomena and what can happen to datasets during the production of synthetic data.

We have two aims with this paper:

First, we aim to problematize the increasing use of synthetic and simulated data as a means to produce knowledge about the world. Through the concept of "intersectional hallucinations" (Johnson and Hajisharif, 2024), we ask how the production of synthetic data is entangled with the reproduction of particular versions of the world, and how these versions of the world can reproduce specific classifications and categorizations. Our argument is that synthetic data can have crucial ontological consequences and contribute to the reproduction of social facts and categories, such as class, race, gender, or age (Fourcade and Healy, 2013). The making of data—perhaps especially synthetic data—is always already entangled with historically contingent normalities, classifications, and social categorizations (Bowker and Star, 1999; D'Ignazio and Klein, 2019; Foucault, 2007; Fourcade and Healy, 2013; Högberg,

[1]Division of Science, Technology, and Society, Chalmers Technical University, Goteborg, Sweden
[2]Department of Science and Technology, Linköping University, Linkoping, Sweden
[3]Department of Science and Technology, Linköping University, Linkoping, Sweden
[4]Department of Thematic Studies – Gender Studies, Linköping University, Linkoping, Sweden

Corresponding author:
Francis Lee, Division of Science, Technology, and Society, Chalmers Technical University, Goteborg, Sweden.
Email: francis@francislee.org

2025). The promise of synthetic data does not automatically replace or fix knowledge making practices developed in historically unjust systems, as demonstrated by well-known examples like the sexist hiring algorithms of Amazon and the racist algorithms for assessing the risk of recidivism (Cf. Angwin et al., 2016; Reuters, 2018; Sandvig et al., 2016). We posit that synthetic data is just as deeply entangled with societal injustices as "real" data, despite claims it can be used to do the opposite—reduce bias and balance data (Jacobsen, 2023).

Our second aim is to attend to the ontological politics of synthetic structured data as they are instantiated in mathematical and statistical assumptions, as well as the concrete intersectional effects of synthesizing data. In this we attend to the politics of what is produced as "normal" when it comes to synthetic data, including the effects of statistical and mathematical choices in simulating data. The selection of particular data distributions to simulate the data, the handling of outliers, and the existence of relational glitches are essential to reproducing the "normal" with synthetic structured data.

To show this, we first explore how the statistics and mathematics of synthetic data are always entwined with performing particular versions of the world as normal—sometimes producing ontological overflows that exclude certain categories, objects, or people, and sometimes magnifying other aspects of the world (Lee, 2023). We aim to bring to the fore the ontological politics (cf. Mol, 1999, 2002) of synthetic data and highlight some of the ways in which we can bring this critique into the various situations and places where synthetic data is constructed and used.

Then we show the concrete effects of the ontological politics of synthetic data as we report on our experiments with producing synthetic data. Here we start at the assertion that the AI techniques currently being used (often GANs and diffusion models) can learn the "essential" aspects of a dataset and reproduce those while modifying other aspects in order to create a new dataset that is relevant but synthetic. Rather than trusting this assertion, we explore it empirically with two examples.

We find that the production of synthetic data often tends to bring to the fore majority elements as the "normal," and minimize minority elements (cf. Bhanot et al., 2019; Cheng et al., 2021: 150; Ganev et al., 2022). But more disturbingly, we also find that even when one can control for this (through the use of training methods) the synthetic dataset will still contain intersectional hallucinations (Johnson and Hajisharif, 2024).

We take intersectional hallucinations to be relational glitches in synthetic structured datasets. They are an artefact of the synthesis; a requirement of the need to be slightly different from the original dataset. The implications of how we can identify and curate for these intersectional hallucinations (and how we should be nudged to do this by policy guardrails) are significant for any use of synthetic structured data. They are an issue for any synthetic dataset because of the complex intersections between different "features" of the data (for early work on relations within data, see Bowker and Star, 1999).

We use the term intersectional hallucination to remind the reader that synthetic data (all tabular data, for that matter) are an ontological flattening of the world which tries to capture the relations between elements of a dataset. Therefore, we reference discussions about Intersectionality—a theoretical tool from black feminism to describe and analyze complex, intersecting power dynamics (Cho et al., 2013; Crenshaw, 1989; Monk, 2022). Additionally, we appropriate the concept of AI generated **hallucinations**—when results from an AI algorithm appear unrealistic or represent non-existent phenomena. Hallucination is often used to describe nonsensical text answers from Large Language Models like ChatGPT or the pictures of hands with too many fingers. While "hallucination" carries baggage (not least, its tendency to anthropomorphize, or humanize, AI by implying the AI itself is hallucinating (Placani, 2024)), it also connects the relatively hard-to-see phenomenon in structured data to more easily understood examples more people may be familiar with.

Our assertion is that intersectional hallucinations are an essential part of synthetic data. They are what makes it synthetic. They cannot be "fixed" to identically mirror the original data without losing the whole point of making synthetic data. Therefore, deciding which versions of synthetic data to use involves deciding which intersections are essential and which can be discarded.

In short, we start with a description of what synthetic data is in the AI domain, which is followed by a discussion of the theories we have found useful to interrogate its nuances. Then we explore the reliance on particular ways of doing statistics when producing synthetic data, after which we report on a concrete example from our experiments with making and examining synthetic data. Here we detail the phenomenon of intersectional hallucinations. Finally, we conclude with some thoughts on the implications of synthetic data use and suggestions for its regulation. Throughout, we attend to the practices and politics of synthetic data and argue that micro decisions about data distributions and outliers can lead to far reaching effects in society.

Synthetic data: Notion and critique

Synthetic structured data are datasets that claim to reproduce the essential aspects of an original dataset (see Jacobsen, 2023; Mackenzie, 2017; Offenhuber, 2024). Synthetic data is not a completely new concept. It has been used in census data to ensure privacy for a long time (Bouk and boyd, 2021; Gigerenzer et al., 1989), but the AI-generated kind is currently gaining traction (Guépin et al., 2024; Li et al., 2023).

Our use of the term "AI" references machine learning techniques applied to existing datasets for analysis and "discovery", but also used to generate new data—synthetic data. Here we address the generative production of synthetic data using machine learning techniques. This particular use of AI produces a particular type of data. Following Suchman, we see it as our task to "challenge discourses that position AI as ahistorical, mystify "its" agency and/or deploy the term as a floating signifier" (Suchman, 2023:1). We attempt to narrow our focus, look at a specific example of AI, and address synthetic data through a close examination of what is happening in the datasets described below. Doing so, we show how methods and theories from STS can be applied to the questions this approach triggers.

Synthetic data are often presented as solving "external" data use problems—like privacy and regulation—specifically because it they are not real data (Jacobsen, 2023; Offenhuber, 2024; Savage, 2023). By introducing noise, for example, synthetic data can allegedly preserve privacy and "de-risk" data (Jacobsen, 2023). They can also solve the problem of small datasets, by amplifying real data (Axelrod et al., 2020; Hao and Orlitsky, 2020; Rajabi and Garibay, 2022)andand.

This use of synthetic data can also allegedly produce the "rare, the unusual and the infinitely variable" (Jacobsen, 2023:9; Wang et al., 2022), supposedly helping to address known biases, and possibly some unknown ones. If a real dataset is biased, say is missing women or over-represents people racialized as white, synthetic data can be generated to balance it out. This is, at least in the discourse, imagined to be particularly useful for medical data (cf. Yoon et al., 2023). We claim one can view these assertions with trepidation. We draw from theoretical work within Science and Technology Studies (STS) and our argument is supported by the results of our experiments with producing and analyzing structured synthetic data.

Our work and this article are focusing on synthetic structured data—synthesized versions of numerical datasets. This is in contrast to much of the work on synthetic data which focuses on synthetic images for computer vision (Buolamwini and Gebru, 2018; Li et al., 2023) or synthetic texts produced through large language models (Wenger, 2024).

Theoretical concepts for troubling synthetic data

Data is always shaped by the settings, people, and machines that produce it. There is never such a thing as raw data (cf. Gitelman, 2013). The politics of classifying the world—making it into data—has caused concern in the social sciences since the birth of sociology in the 1700s and 1800s, and in the fields of social physics and statistics. Struggles about classifying the world were tied to debates about the reality of *l'homme moyen*—the average man—in society, and how social statistics should be interpreted in relation to the reality of social groups (cf. Desrosières, 1998; Gigerenzer et al., 1989). The classification of the world in categories—and the production of statistics about society—was also implicated in the development of questionable social reforms, with ties to the eugenic movement striving for creating a society of geniuses while culling the less desirable elements of society (MacKenzie, 1976).

In STS, there has existed a longstanding critical engagement with data and the practices and infrastructures of classification. This work has brought to the fore how the making of classifications—such as disease classifications, race classifications, or gender classifications—are intertwined with politics and power struggles. Seemingly mundane technologies like databases have been shown to be implicated in political and epistemic choices that have impacted what becomes included and excluded in knowledge making practices (Bowker, 2000; Bowker and Star, 1999; Hine, 2006; Star, 1990). We expect that the generation and use of synthetic datasets will become normalized and part of machine learning pipelines. They will become part of the background infrastructures of society's information systems. They are important to understand because these aspects of data—synthetic or otherwise—produce categories of identity and structural opportunities and discrimination, which impact an individual's life and health.

Algorithmic performativity

There has also been a longstanding discussion about the *performativity* of models and equations. Work emerging from the study of markets in STS has shown that models, equations, and algorithms shape the very phenomena that they are set to describe. The act of introducing a particular way of describing a phenomenon with equations algorithms or mathematics contains the possibility that people, objects, or phenomena are performed differently based on these new descriptions (cf. Callon, 1998; Callon and Muniesa, 2005; D'Ignazio and Klein, 2019; MacKenzie, 2008).

The same arguments also hold true for synthetic data, perhaps even more so. Synthetic data—just as other data—are produced by particular actors, at particular times, at particular places, with particular algorithms and equations, as well as in relation to particular political and social projects. While these relations are implicated in the production of "raw" or "real" data, they are also maintained and reproduced in the production of synthetic data, reflecting particular choices made by particular actors in those particular situations. But in contrast to other data, synthetic data adds another layer of "production:" the changes to data made by the statistical patterns engaged in the synthesizing

processes (see below) (Mackenzie and Spears, 2014a, 2014b).

Much of the work on fairness and bias in AI data has begun to address representation of minority groups or elements in data. Many researchers from within the machine learning community are aware of the diversity problems in synthetic data and are working on metrics to address this (Bhanot et al., 2019; Cheng et al., 2021; Mehrabi et al., 2022; Verma and Ruben, 2018). However, their work, while important and necessary, is not sufficient. It tends to adopt a simplified understanding of identity as static, and social categories as ontologically predetermined (a critique also rightfully directed towards intersectional theory (Monk, 2022)). By not recognizing the role of context and power dynamics in the production of data subjects, this analysis of synthetic data will not help produce data applicable to real-world research questions (cf. Miceli et al., 2021).

Addressing this problem requires a shift in perspective to include more than complex mathematical definitions of "fair" representation. Lessons from theoretical work on intersectionality assert that static categories like class, race, gender, age, etc. are artifacts of sociological knowledge making practices, just like any other categorizing practice. The important aspect of these identity holders is not their terminology or the elements contained in a category, but the ways that temporal, contextual structural power dynamics interact on/with subjects to position them in moments (and lives) of discrimination or privilege (Cho et al., 2013; Ciston, 2019; Washington, 2017).

Social categories like race, age, class or economic status, geographical location, or resource access have become important considerations when planning and executing research and the collection and analysis of human data (Mena et al., 2019). Yet these categories are fluid. For example, our understanding of sex as a binary category has been challenged by the trans community, forcing a reconsideration of some of the categories used for analysis and data collection in social, medical and biological research (Restar et al., 2021). Race is another example of a category that is constantly undergoing reconfigurations. Critiques of these categories have led to significant changes in the data gathering practices (Epstein, 2007; Mapes et al., 2020).

Categories and structures

The recognition of categories in flux still often presumes a static mapping of categories and identities onto individuals and does not engage with intersectional insights on how dynamical social structures shape what lives we can live (Epstein, 2007). One tool to address this has been the development of intersectionality theory (see Cho et al., 2013; Collins, 2000 [1990]; Cooper, 2016; Crenshaw, 1989), which argues that we need to understand individuals and their identities as produced through contexts, including social locations and power structures (Bowleg, 2008: 314). People do not "have" identities, they position themselves (and are positioned by others) within fluid spectrums of power (see Washington, 2017). As Monk (2022) notes, especially when engaged in quantitative data practices which reproduce existing categories, this is still a simplification. But it also reflects the data currently being used in many research and policy contexts.

For our analysis of synthetic data we engage an intersectional sensitivity which understands that people are positioned by the categories that are created to hold them, something one of the foundational texts in STS showed through the analysis of apartheid racial categories (cf. Bowker and Star, 1999). This is important because demographic labels (and their resulting columns in a tabular dataset) like class, race, sex, age are not explanatory (as stress or prejudice would be), they are merely labels. Intersectionality theory, with its emphasis on power dynamics, would suggest columns based on dimensions of experience, like earnings, access to healthcare, etc. (Bowleg, 2008: 316), stigma, marginalization (Benjamin, 2019; Rouhani, 2014), and, just as importantly, the access to privilege (Varley and Kaminski, 2021), rethinking the categories and what is placed in them in the production of tabular data columns (Monk, 2022). In data collected from a recognition of the generative effects of power dynamics, columns would be categories that explore contexts rather than collecting head counts.

This is not often done, however. Thus, the analyses of intersectional subjectivities tend to follow categorization practices that have developed over time (the usual suspects of class, gender, race, age, ability, coloniality…). These categories are used as shorthand for diversity. However, we emphasize that in its richest form, intersectionality (in general and in AI data generation) is not just about gathering and analyzing more and nuanced data with different categories. A true intersectional analysis would engage discussions of how categories are made by context, see the power structures shaping opportunities and discrimination, and explore the fluidity of the subject positions within them (cf. Cho et al., 2013: 787; Lykke, 2011). Still, in the below, we will start with what we have (census data from 1990) and see what happens to the categories and their intersections as if they were static and stable labels. Even though this is a simplified way of using intersectionality, it is still going to show some troubling elements of synthetic data and its conceits.

Thus, by basing our understanding of synthetic data on the critical insights from the social history of statistics, STS, and feminist theory, we want to highlight some of the classification politics in the age of synthetic data. In particular, we are pushed to ask questions such as: Which categories—and intersections of categories—are produced with synthetic data? What people, natures, phenomena, or

objects are excluded? How are these choices made in simulating data? In which situations and practices? And for whose benefit—*cui bono*?

The synthetic data vault and the challenge of data distribution

Our first close reading of synthetic data involves a software package called the Synthetic Data Vault—the SDV, developed at MIT which was later spun off in a company named DataCebo. The SDV aims to create a generally applicable data synthesizer that can be applied to any type of data. The aim is to make "an open source software ecosystem for generating synthetic data" (DataCebo Blog, 2021). The company argues that the need for synthetic data is one of the key solutions to being able to share data freely, in light of sharing personal information and company secrets. Their data synthesis techniques use machine learning to create synthetic data that mimic the mathematical, relational, and formatting properties of the original database:

> The SDV uses machine learning to analyze data. Then, it creates fully synthetic datasets that mimic the original. Although the synthetic data is entirely machine generated, it maintains the original format and mathematical properties. (DataCebo Blog, 2021)

Their original approach to synthetic data, which was published in a paper in 2016, hinged on statistical analyses of the original database based on a number of assumptions and techniques (Patki et al., 2016). This publication outlines three main statistical techniques to produce synthetic data: The Gaussian Copula, Probability Distributions, and Covariance, which are the backbone of the tool, engaging a number of statistical concepts and assumptions that tie into the politics of synthetic data. These algorithmic and machine learning tools produce particular conditions of possibility for the generated data, which begs the question: What kind of synthetic world is produced through the mathematical and machine learning techniques that are used? This question is becoming ever more relevant in a world that is teeming with generative AI (Jacobsen, 2023; Offenhuber, 2024). So here we explore the production of specific conditions of possibility through the algorithms and statistical techniques of the SDV.

Normals and outliers: What is the right way to enact a simulated world?

In order to simulate the mathematical and relational properties of any dataset, the team behind the SDV have employed a number of statistical techniques to analyze the distribution of the data in statistical space. They used "a generative model [that] relies on knowing the distribution shapes of each of its columns" (Patki et al. 2016, 3). That is, the SDV attempts to mimic the statistical distribution of the original data. The SDV has a few different models for synthesizing datasets, and the one we experiment with below is called the CopulaGANSynthesizer, which is based on particular statistical techniques. For instance, if the original data distribution has a Gaussian shape—the shape of a normal distribution—the SDV will attempt to generate data that also has a bell-curve shape.

However, since calculating the exact distribution of a dataset is complex and time-consuming, the team behind the original SDV package (Patki et al., 2016) included four ready-made probability distributions that could be used. These were the truncated Gaussian, uniform, beta, and exponential distributions (Patki et al., 2016: 3) (see Figure 1). Thus, the SDV package was built to avoid the need for analyzing the original data directly because that would take too much computational resources. But, by default the package uses a statistical test to choose between only two of the predefined distributions—the Gaussian and uniform distributions.

Importantly, the assumption inscribed into the software is that the original dataset—which is the basis for making synthetic data—has the characteristics of one of the two probability distributions. And more to the point, these four pre-packaged distributions also include an assumption about the properties and nature of the world that is described by these datasets, entangling the machine learning algorithm and the production of synthetic data in the history of statistics as a research field and resource for scientific knowledge production (cf. MacKenzie, 1976).

Consequently, generative data modeling is inherently political, reflecting underlying assumptions inscribed in these distributions. Our point is that the shape of a statistical distribution contains assumptions about how the world is structured and functions. Assuming that the world can be described by a Gaussian curve assumes that outliers are rare—on the ends of the bell curve. Assuming an exponential function assumes that there is a long tail of diminishing values. Assuming a uniform distribution assumes that there are equal chances for each outcome. And the beta distribution is flexible and complicated. By default—and due to computational efficiency—the mathematics of generating data in the original SDV package depended on these four probability distributions. *Thus, by default, the complexity of a world is reduced to four probability distributions for computational efficiency—probably mathematically defensible in many situations, but probably problematic in many other situations.*

Importantly, each statistical distribution included in the SDV contains an assumption about what a "normal" world looks like if visualized statistically. What does this mean? Take, for instance, the scene in Michael Crichton's book *Jurassic Park*, where the mathematician Malcolm,

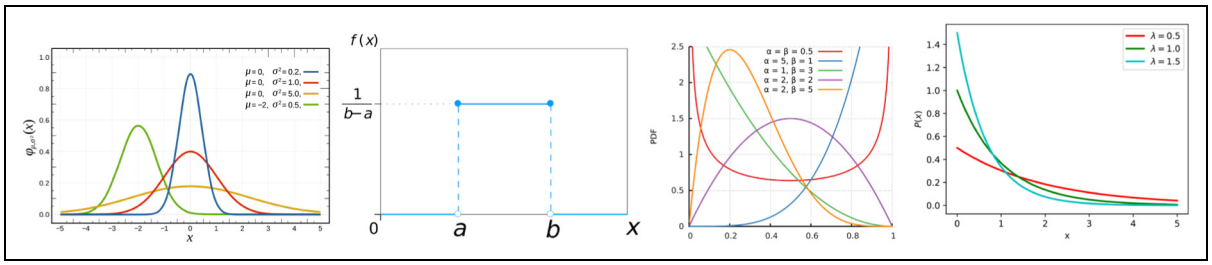

Figure 1. The pre-packaged data distributions in the SDV. Gaussian distribution, uniform distribution, beta distribution, and exponential distribution.
Image source: Wikipedia.

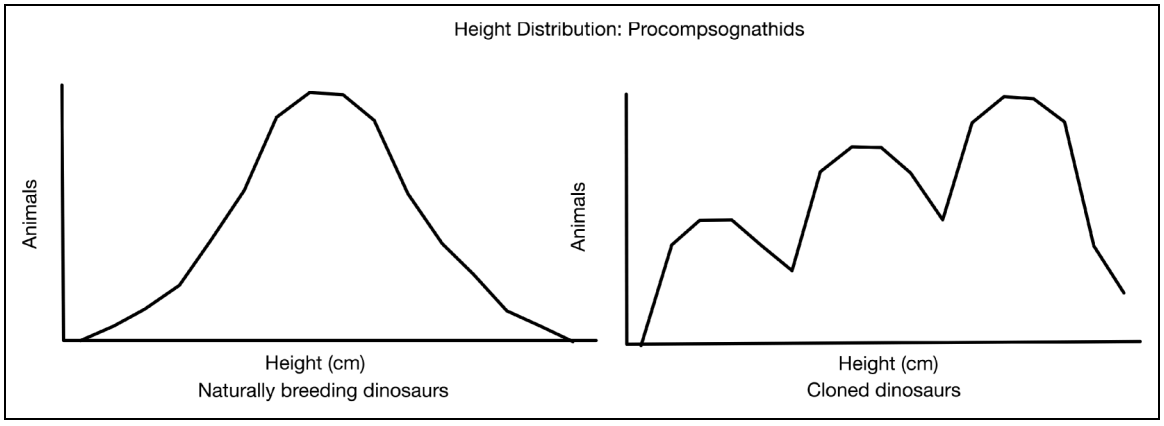

Figure 2. The dinosaur height distributions redrawn from Michael Crichton's book, Jurassic Park.

played by Jeff Goldblum in the 1993 movie of the same name, brings up a Gaussian curve to show that the dinosaurs are indeed breeding—not only being cloned using biotechnology. The assumption that the mathematician brings to the fore is that the dinosaurs' heights in the park are normally distributed along a Gaussian curve when they breed in the wild—while artificially cloned dinosaurs would follow a different distribution of data, visualized with three peaks in the book by Crichton (see Figure 2).

In fiction and in real life, distributions of data are built on assumptions about the world—for instance, that breeding dinosaurs would lead to a Gaussian curve. These assumptions are not innocent, but contain ideas that can be very consequential in the world—not least in a world of generative data.

For instance, in economics the choice of how to model the economy and how often it crashes is a good example of the politics of data distributions in action. In financial calculations of risk there have been two dominant models that describe the world using different statistical distributions. One, the Black-Scholes-Merton model (BSM) used widely in options trading (discussed at length in MacKenzie, 2008), understands market crashes as outliers on a Gaussian curve—as very unusual events—thus relying on a Gaussian curve to model market movement (cf. Lee et al., 2019). This choice of a Gaussian data distribution assumes that small changes in markets are very common (they are in the middle of the Gaussian curve) and that larger changes (like market crashes) are less likely and would be at the end of the Gaussian curve. Thus, an idea about how the world of markets works is inscribed into the model; an ontological politics of data and markets.

However, another competing economic model, created by Benoit Mandelbrot, relies on a fractal representation of data, which assumes that very large changes (market crashes) occur more often than the Gaussian curve assumes. One model assumes that crashes are rare, and another assumes that crashes are normal (cf. Lee et al., 2019).

In the words of Mandelbrot and Taleb "The traditional Gaussian way of looking at the world begins by focusing on the ordinary, and only later deals with exceptions or so-called outliers as ancillaries" (Mandelbrot and Taleb, 2010: 48). Mandelbrot and Taleb describe the politics of these different ways of modeling economics in the following manner:

> [M]any fundamental quantities follow distributions that have "fat tails"—namely, a higher probability of extreme

> values that can have a significant impact on the total. One can safely disregard the odds of running into someone several miles tall, or someone who weighs several million kilograms, but similar excessive observations can never be ruled out in other areas of life. […] So, while weight, height, and calorie consumption are Gaussian, wealth is not. Nor are income, market returns, size of hedge funds, returns in the financial markets, number of deaths in wars, or casualties in terrorist attacks. (Mandelbrot and Taleb, 2010: 49–50)

Assumptions about data distributions—like the assumptions in the economic models or the SDV or the breeding of dinosaurs—are not innocent. They are powerful statistical assumptions that shape how the world is understood and described using mathematics and statistics. This is an example of the ontological politics of describing the world through statistics. Furthermore, assuming that a dataset is distributed in a particular way is performative to the highest degree when that distribution is used to synthesize data. The mathematical assumptions about the world fundamentally reshape the synthetic world.

Returning to the SDV, in the first iteration of the software published by Patki et al., 2016, the SDV makes a statistical test to determine if the data is more likely to be a uniform distribution or a truncated Gaussian distribution. For the sake of discussion, let us focus on the truncated Gaussian distribution. The truncated Gaussian distribution is a mathematical technique that cuts off the outliers from a normal distribution. It is sometimes used for simulating data while preserving privacy (Parsa, 2009). It has been described as

> a mathematically defensible way to preserve the main features of the normal distribution while avoiding extreme values [that] involves the truncated normal distribution, in which the range of definition is made finite at one or both ends of the interval. (Burkardt, 2023)

In plain language, the truncated Gaussian distribution uses a description of the original data that is cut off at the top and the bottom of the bell curve. Thus, the unusual data points—outliers—are removed from the distribution that the generative model, the SDV, is trying to replicate. Outliers are often seen as a problem for machine learning models as they tend to make the model "overfit"—which often results in the predictive capacity of the model going down.

Our questions of the ontological politics of synthetic data relate to what types of normal worlds are enacted with these machine learning models (Lee et al., 2019). For instance, by truncating the normal distribution in the SDV, rare occurrences, outliers, become excluded from the dataset that is used to train the generative model. The people, things, or phenomena that are unusual—outliers, that are outside *l'homme moyen,* the medium man of nineteenth century statistics—become removed from the distributions that are used to create synthetic data.

The consequence is that unusual data points will become even more unusual for the model that is supposed to produce the synthetic data, unless specific techniques to include them are used (cf. Chen et al., 2024) people, objects, or phenomena that are unusual outliers are removed from the data-world that the model is trained on. And importantly, the parameters that are chosen for the model also create the conditions of possibility for the simulated data. Thus, by relying on a truncated normal distribution the SDV creates particular conditions of possibility for the generative model—and for how it creates synthetic data.

Knowing how the outliers would affect the model is a matter of knowing what the outlier represents, and how it would affect the model in specific cases (Grace-Martin, 2008). What might often be "mathematically defensible," might not be defensible in the politics of simulating the world. In essence, truncated normal distributions might make mathematical sense, but we need to be careful that we do not perform a narrower world—where the performative effects of synthesizing data are exclusionary to unusual people or things. We can pose the question: when is it mathematically defensible to truncate outliers? To make the world of data narrower? Given these questions, we now explore what happened when we used a machine learning algorithm to produce synthetic data.

Synthetic data and intersectional hallucinations

As detailed above, the production of synthetic data from the SDV relies on finding statistical distributions of the data within the columns of the original data. However, the important aspects of a dataset—the reason one collects data in tabular form in the first place, is to analyze the *relations between the columns*. In addition to being local (Loukissas, 2022) and contextual (D'Ignazio and Klein, 2019), all data is relational. Here we attend to what happens when synthetic data is made with a sensitivity to the distributions—first within the columns—then between the columns at multiple intersections.

To explore the effects of the truncated Gaussian distribution discussed above (when outliers disappear), we first decided to test for some of the things we had read about the tendency of ML algorithms (especially generative adversarial networks, GANs) to over-represent majority elements in a dataset and under-represent minority elements (Bhanot et al., 2019; Chen et al., 2024). We were concerned that if one started producing synthetic data with a technology that minimized existing edge-cases, there would be the risk that the (proportionally few) data about minoritized groups would slowly (or quickly) disappear.

Using a couple of different GANs, we made synthetic versions of the 1990 US Adult Population Census Data. One of these GANs was specifically meant to address fairness (Rajabi and Garibay, 2022), but we found that the synthetic data generated by that had a tendency to miss edge cases due to mode-collapse of GAN models (see the truncating discussion above) and thereby create a new dataset that was focused on the normal. In other words, some of the edge-cases in these initial runs did, indeed, disappear.

For example, in the original data, there were people from 40 different countries of origin. In one of the first synthetic datasets we generated, there were only 31 countries of origin. Nine countries which had very few emigrants to the USA in that dataset simply disappeared. The phenomenon of disappearing edge cases occurred in all of the different categories in the data (country of origin, educational level, professional occupation, etc). Given the use of truncated Guassian distributions, it is expected that this would happen and other researchers are also noting this phenomenon (see Shumailov et al., 2024; Wenger, 2024; Xu et al., 2019).

Because we were concerned about edge-cases and looking for them, we were able to catch it when they went missing and develop ways to tweak the GANs to produce synthetic datasets that still contained the edge-cases (Johnson and Hajisharif, 2024). One can, for example use a diffusion model (another type of AI) to produce synthetic data in an even more carefully curated way. Doing so, however, required that we knew what we were looking for ahead of time and could curate the synthetic data to ensure the edge-cases were represented. In some ways, this demonstrates a point that Jacobsen (2023) and others make about synthetic data: with careful curation, it can be used to change representation dynamics in a dataset, potentially adjusting for known biases (see also Dehdarirad et al., 2024). However, it also shows that without careful curation, the use of machine learning to produce synthetic structured data can actually do the opposite—increase bias and decrease representation of minority elements.

Disappearing edge cases were not all we found, however. Going through the data, we noticed something else happening: intersectional hallucinations (Johnson and Hajisharif, 2024).

We are more than just our country of origin. Or our educational level. Or our age. Or just our gender. Intersectionality theory has shown us that our identities and subject positions are created through the complex intersections of power structures in society (Cho et al., 2013; Crenshaw, 1989). As detailed in our theory section, translating these into a series of columns of population data is fraught (Bouk and boyd, 2021; Monk, 2022). But while census data is far from perfect, it does at least give a hint of the complexity of different subject positions—of different data points—a complexity that is very relevant for any analysis or use of population data. Given this, intersectional complexity in a synthetic version of population data is also important.

The scatter plot matrix (SPLOM) in Figure 3 is a visualization of what happened to some of the intersectional relations in a synthetic dataset generated in the experiments we were running on the 1990 Adult census data.

The lighter squares in the bottom left half of the diagram are three-part intersections of datapoints in the real 1990 US Adult Census Dataset. The greyer squares in the upper right half of the diagram are the same intersections but in the synthesized data (made with CTGAN). The bar graphs on the diagonal show how these two datasets compare at the single column level—for example, the data in the purple circle compares the age distribution in the original data with the synthetic data. This shows that distribution is close but not identical in the two datasets, which is good. Synthetic data must be slightly different than the original data or else it is not synthetic (and would not assure privacy, avoid regulation, allow portability, facilitate amplification, etc.).

But look at what happens in the data at an intersectional level: For example, in the green circles, one sees the intersection of age-income-gender in the original dataset and the synthetic one. The synthetic dataset represented this intersection fairly well. Again, it was not identical (which is good, or else it would not be synthetic data), but pretty close. This is an example of what we are calling intersectional fidelity.

Now look at the intersection of marital status, occupation and gender (in the orange circles). At this intersection in the data, there are a lot more females in the synthetic data than in the original data, and at particular occupations (the synthetic data square is pinker in the middle). When we tried to figure out why, we saw that in the original dataset there was one husband who was also classified as female. In the synthetic dataset there were 259 husbands who were also female. Today, and in some countries, that is perfectly fine. But remember, this is a synthesis of the 1990 US census data. Those extra female-husbands are an intersectional hallucination.

There were also many other intersectional hallucinations. For example, the synthetic data also had 333 datapoints labelled husband/wife *and* single. The AI had not learned (or been told) that this is an unrealistic representation of the original data. Of these, over 100 datapoints were never married-husbands earning under 50,000 USD a year, an intersection that did not exist in the original data. On the other hand, there were widowed-females-working in tech support in the original dataset, but they were missing in the synthetic version. These are all intersectional hallucinations.

Implications

The narrow implications of these particular intersectional fidelities and hallucinations in our dataset is that *maybe*

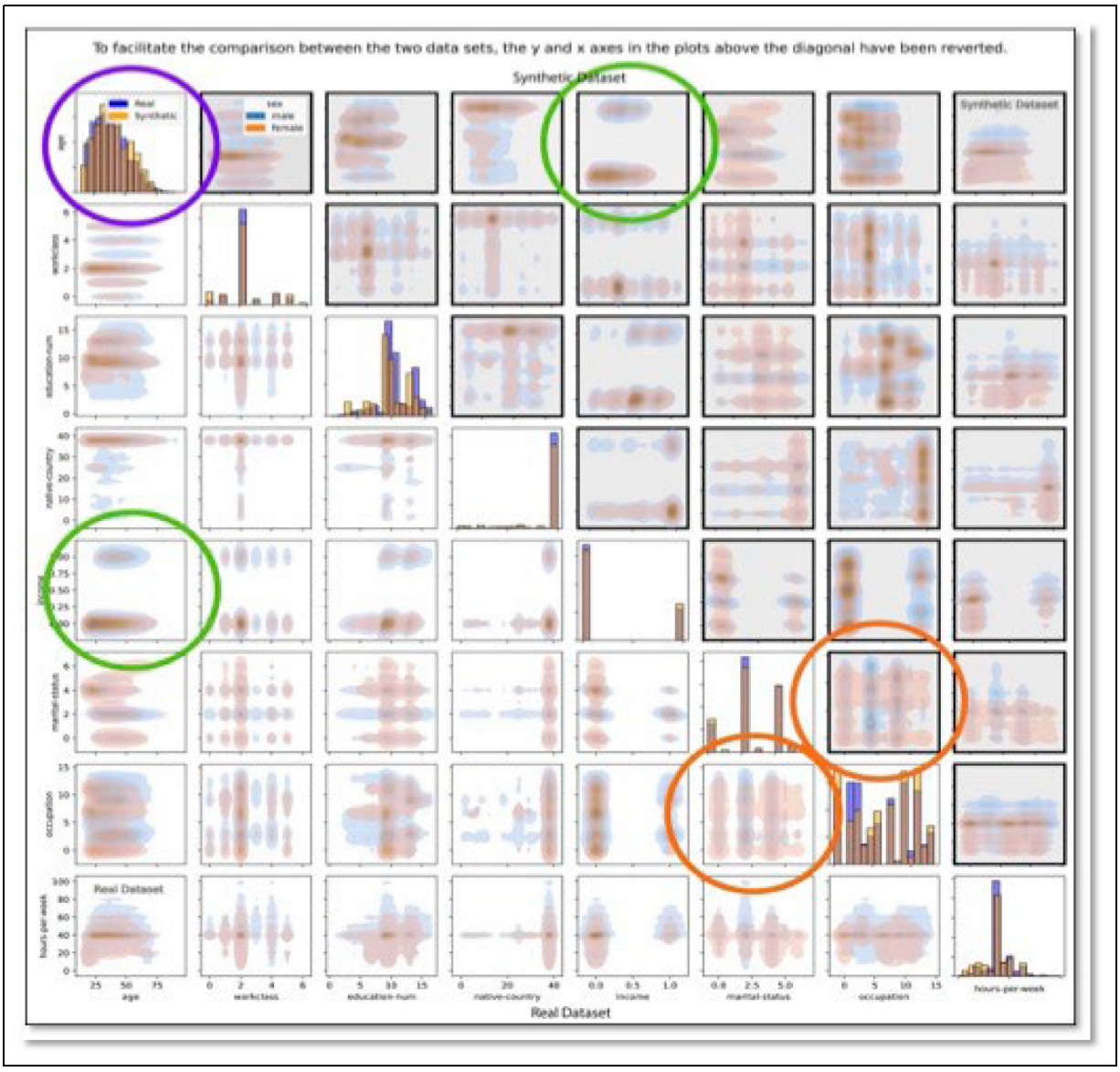

Figure 3. A SPLOM of three-point intersections in the original data (white) and synthetic data (grey). Circles explained in text. Data: 1990 US adult census.

the synthetic dataset could be used for research on age-income-gender questions (where there was an intersectional fidelity) but not if one were interested in using it for research related to widowed females working in tech support. And one should watch out for never-married husbands appearing in the results. In other words, this particular synthetic dataset might be useful for some purposes, but it would not be appropriate to claim it (or any other synthetic dataset) could be a general-purpose dataset.

But there are wider implications than that.

For one thing, the presence of intersectional hallucinations triggers the question: Where should one stop? The fidelities and hallucinations described above are two-part and three-part intersections. What about four-part intersections? Or five-part? One could go on and on and on, checking more and more complex fidelities and hallucinations in the synthetic data… At what point would a complex intersectional hallucination make the synthetic data irrelevant or misleading? And for which purposes. Here we note that Offenhuber (2024) suggests that synthetic data may need to be evaluated "relationally," based on their contexts of use, rather than on their relevance within a representational paradigm. We agree, though note that concern about the contexts of data (collection and use) is also relevant to "real" data (cf. Asdal and Moser, 2012; Huvila et al., 2024).

Additionally, we suggest that intersectional hallucinations need to be approached with "caution" and

"guardrails." At the same time, in some discourses, the existence of intersectional hallucinations is what makes synthetic data useful, both for ensuring privacy and for amplifying existing datasets prior to machine learning analysis. They are the "added noise" that protects privacy. They are the "additional information" that a machine learning algorithm is trying to learn. But they may also be one of the reasons for model collapse (cf. Shumailov et al., 2024) and we wonder about the implications of using synthetic data and its inherent hallucinations as the basis for scientific research or policy decisions. Will this not lead to incorrect outcomes, with potentially serious consequences for science and for society?

Finally, we claim that intersectional hallucinations are especially problematic as synthetic data are added to existing datasets or deposited in data repositories without sufficiently detailed meta data, readme texts or labels. Synthetic structured data is often being generated because of a desire to use machine learning in domains that do not necessarily have access to sufficiently large datasets. By generating synthetic data and adding it to an existing dataset, one can produce enough data to "run an AI." But these new, heterogeneous datasets are often currently impossible to trace and untangle. And they are being shared. Existing datasets are being contaminated with the intersectional hallucinations of synthetic data. Over time, we may have a problem of data validity and reliability. We suspect this will lead to new practices of data pedigrees that are going to disrupt many domains.

Conclusion: Lessons for a reality with synthetic data

All data are made. Whether the data are described as raw, cooked, objective, subjective, real, or synthetic it takes work to make them (cf. Gitelman, 2013). Data do not just exist in a state of nature for us to harvest. Data in a database are painstakingly made, chosen, curated, and systematized. In the creation of each database, choices have been made about what to include and exclude (Bowker, 2000). Data are performative. In this article, we have attended to yet another such performative version of data: "synthetic" data.

The production of any version of data—raw/cooked, objective/subjective, real/synthetic—is connected to enacting a particular version of the world (Mol, 1999; 2002). Be it data about predictive policing (Benbouzid, 2019; Hälterlein, 2021), disease surveillance (Lee, 2021a; 2021b), or valuing credit derivatives (MacKenzie and Spears, 2014a; 2014b)—data holds particular conditions of possibility.

Above, we have demonstrated the importance of attending to mathematical assumptions. Assumptions about data distribution, outliers, and the world have effects on how we synthesize data. We have used these sensitivities to tinker with synthetic data to highlight some of the potential pitfalls with synthetic data. We identified several performative effects in using the Synthetic Data Vault. In one of our synthetic datasets, whole countries were removed from the world. In another synthetic data set, husbands who were also single became very common in the dataset. These results show how important it is to think outside of single categories and attend to performative relations between categories.

This means that the instantiations of synthetic data which we encounter as analysts need to be understood both in terms of their mathematical assumptions, as well as for the particular intersectional hallucinations they contain. Thus, in our call for considering the politics of data, we want to advocate for careful consideration of the politics of mathematical modeling—in how the world is imagined in terms of data distributions and outliers, and categories —but also careful consideration of the politics enacted by caring for some intersections over others.

Any version of data contains particular conditions of possibility for putting together a particular world. By truncating a normal distribution, unusual phenomena can become enacted as more unusual, often prompting performative effects on the phenomena that they are set to describe. By using these statistical definitions of what is essential and important in data—the relations between and within columns, we are also making other things invisible and potentially changing them, including the relations between the columns.

Intersectional hallucinations in synthetic data occur in the relations between categories and *structured data is about relations*. Developing a vocabulary to talk about them in synthetic data needs to engage the statistical terms currently employed in the synthetic data literature and industry, but it also needs to use theoretical tools from within STS research, feminist technoscience and critical data studies on the power of categorization, relationality, and ontological politics (cf. Barad, 2007; Bowker and Star, 1999; Crawford, 2021; Klein and D'Ignazio, 2024; Mol, 2002; Suchman, 2007; Suchman, 2023). This will help researchers consider the practices that produce intersectional hallucinations as ontological multiplicities made through relations.

In order to understand the ontological politics of synthetic data, we need to sensitize ourselves to the various choices, algorithms, mathematical operations, and software that produce synthetic data. We need to attend to the choices of what to include and exclude in simulating new datasets—whether the operation makes mathematical sense or not. And we need to attend to the performative effects that data, statistics, and mathematics can have. Especially as their enactment in particular situations, cultures, or organizations can generate potentially catastrophic consequences. But we also need to attend to the intersectional fidelities and hallucinations that are occurring in the

synthetic data—make them visible and ensure information about them is also ported with the synthetic data as it is released out into the world. This is essential for heterogeneous domains—medicine and financial markets, but also the natural sciences and tech. And it is important for individuals whose lives might be shaped by the use of such 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 WASP-HS (NetX).

ORCID iDs

Francis Lee https://orcid.org/0000-0002-7206-2046
Saghi Hajisharif https://orcid.org/0000-0002-0176-5852
Ericka Johnson https://orcid.org/0000-0001-5041-5018

Note

1. In citation marks to indicate the ontological complexity hidden in the term "raw data", see Gitelman (2013).

References

Angwin J, Larson J, Mattu S, et al. (2016) Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Asdal K and Moser I (2012) Experiments in context and contexting. *Science, Technology, & Human Values* 37(4): 291–306. http://www.jstor.org/stable/41511177.

Axelrod B, Garg S, Sharan V, et al. (2020et al. (2020) Sample amplification: Increasing dataset size even when learning is impossible. In: *International Conference on Machine Learning*, pp.442–451: PMLR.

Barad K (2007) *Meeting the Universe Half Way*. Durham: Duke University Press.

Benbouzid B (2019) Values and consequences in predictive machine evaluation: A sociology of predictive policing. *Science & Technology Studies* 32(4): 119–136.

Benjamin R (2019) *Race After Technology*. London: Polity.

Bhanot K, Qi M, Erickson JS, et al. (2019) The problem of fairness in synthetic healthcare data. *Entropy* 23(9): 1165.

Bouk D and Boyd D (2021) *Democracy's Data Infrastructure: The Technopolitics of the US Census*. New York: Knight First Amendment Institute.

Bowker GC (2000) Biodiversity datadiversity. *Social Studies of Science* 30(5): 643–683.

Bowker GC and Star SL (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge: MIT Press.

Bowleg L (2008) When black + lesbian + woman $\not=$ black lesbian woman: The methodological challenges of qualitative and quantitative intersectionality research. *Sex Roles* 59(5–6): 312–325.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research* 81: 1–15.

Burkardt J (2023) The truncated normal distribution. Available at: https://people.sc.fsu.edu/~jburkardt/presentations/truncated_normal.pdf.

Callon M (1998) Introduction: The embeddedness of economic markets in economics. *The Sociological Review* 46(S1): 1–57.

Callon M and Muniesa F (2005) Peripheral vision: Economic markets as calculative collective devices. *Organization Studies* 26(8): 1229–1250.

Chen W, Yang K, Yu Z, et al. (2024) A survey on imbalanced learning: Latest research, applications and future directions. *Artificial Intelligence Review* 57(6): 137.

Cheng V, Suriyakumar VM, Dullerud N, et al. (2021) Can you fake it until you make it?: Impacts of differentially private synthetic data on downstream classification fairness. In: *FAccT '21*, Virtual Event, Canada, March 3–10, 2021. ACM.

Cho S, Crenshaw K and McCall K (2013) Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs* 38(4): 785–810.

Ciston S (2019) Imagining intersectional AI. *xCoAx*. 2019.

Collins PH (2000 [1990]) *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*, 2nd edn. New York: Routledge.

Cooper B (2016) Intersectionality. In: Disch L and Hawkesworth M (eds) *The Oxford Handbook of Feminist Theory*, Vol. 1. Oxford: Oxford University Press, 385–406.

Crawford K (2021) *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Crenshaw K (1989) Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum* 1989: 139–167.

DataCebo Blog (2021) *Meet the synthetic data vault*. Available at: https://datacebo.com/blog/intro-to-sdv/ (accessed 14 March 2023).

Dehdarirad T, et al. (2024) Enhancing tabular GAN fairness: The Impact of Intersectional Feature Selection. *ICMLA*. 2024.

Desrosières A (1998) *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge: Harvard University Press.

D'Ignazio C and Klein L (2019) *Data Feminism*. Cambridge: MIT Press.

Epstein S (2007) *Inclusion*. Chicago: University of Chicago Press.

Foucault M (2007) *The Order of Things: An Archaeology of the Human Sciences*. London: Routledge.

Fourcade M and Healy K (2013) Classification situations: Life-chances in the neoliberal era. *Accounting, Organizations and Society* 38(8): 559–572.

Ganev G, Oprisanu B and De Cristofaro E (2022) Robin Hood and Matthew effects: Differential privacy has disparate impact on synthetic data. In: *Proceedings of the 39th International Conference on Machine Learning*, vol. 162, 2022. Baltimore, MD: PMLR.

Gigerenzer G, Swijtink Z, Porter T, et al. (1989) *The Empire of Chance: How Probability Changed Science and Everyday Life. Ideas in Context*. Cambridge [England]; New York: Cambridge University Press.

Gitelman L (2013) *Raw Data' Is an Oxymoron*. Cambridge: MIT Press.

Grace-Martin K (2008) Outliers: To drop or not to drop. In: *The Analysis Factor*. Available at: https://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/ (accessed 17 March 2023).

Guépin F, Meeus M, Creţu A-M, et al. (2024) Synthetic is all you need: Removing the auxiliary data assumption for membership inference attacks against synthetic data. In: Katsikas S (ed) *Computer Security. ESORICS 2023 International Workshops. ESORICS 2023. Lecture Notes in Computer Science*, Vol. 14398. Cham: Springer, 182–198.

Hälterlein J (2021) Epistemologies of predictive policing: Mathematical social science, social physics and machine learning. *Big Data & Society* 8(1): 20539517211003118.

Hao Y and Orlitsky A (2020, November) Data amplification: Instance-optimal property estimation. In: *International Conference on Machine Learning*, pp. 4049–4059: PMLR.

Hine C (2006) Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science* 36(2): 269.

Högberg C (2025) This ground truth is muddy anyway. Ground truth assemblages for medical AI development. *Sociologisk Forskning* 2025(1). Forthcoming.

Huvila I, Andersson L and Sköld O (eds) 2024) *Perspectives on Paradata: Research and Practice of Documenting Data Processes*. Cham: Springer. http://doi.org/10.1007/978-3-031-53946-6.

Jacobsen BN (2023) Machine learning and the politics of synthetic data. *Big Data & Society* 10(1): 20539517221145372.

Johnson E and Hajisharif S (2024) The intersectional hallucinations of synthetic data. *AI & Society*. 1–3. Epub ahead of print. DOI: 10.1007/s00146-024-02017-8.

Klein L and D'Ignazio C (2024) Data feminism for AI. In: *FAccT '24*. June 03-03, 2024, Brazil: Rio De Janeiro.

Lee F (2021a) Enacting the pandemic: Analyzing agency, opacity, and power in algorithmic assemblages. *Science & Technology Studies* 34(1): 65–90.

Lee F (2021b) Sensing Salmonella: Modes of sensing and the politics of sensing infrastructures. In: Witjes N, Pöchhacker N and Bowker GC (eds) *Sensing In/Security: Sensors as Transnational Security Infrastructures*. London: Mattering Press, 97–131.

Lee F (2023) Ontological overflows and the politics of absence: Zika, disease surveillance, and mosquitos. *Science as Culture* 33(1): 417–442.

Lee F, Bier J, Christensen J, et al. (2019) Algorithms as folding: Reframing the analytical focus. *Big Data & Society* 6(2): 1–12.

Li X, Wang K, Gu X, et al. (2023) Parallel eye pipeline: An effective method to synthesize images for improving the visual intelligence of intelligent vehicles. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 53(9): 5545–5556.

Loukissas Y (2022) *All Data Are Local*. Cambridge: MIT Press.

Lykke N (2011) Intersectional analysis: Black box or useful critical feminist thinking technology. In: Lutz S (ed) *Framing Intersectionality: Debates on a Multi-Faceted Concept in Gender Studies*. Farnham: Ashgate, 207–219.

MacKenzie A (2017) *Machine Learners*. Cambridge: MIT Press.

MacKenzie D (1976) Eugenics in Britain. *Social Studies of Science* 6(3–4): 499–532.

MacKenzie D (2008) *An Engine, Not a Camera: How Financial Models Shape Markets*, 1st edn. Cambridge: MIT Press.

MacKenzie D and Spears T (2014a) A device for being able to book P&L': The organizational embedding of the Gaussian copula. *Social Studies of Science* 44(3): 418–440.

MacKenzie D and Spears T (2014b) The formula that killed Wall Street': The Gaussian copula and modelling practices in investment banking. *Social Studies of Science* 44(3): 393–417.

Mandelbrot B and Taleb N (2010) Mild vs. wild randomness: Focusing on those risks that matter. In: Diebold F, Doherty N and Herring R (eds) *The Known, the Unknown, and the Unknowable in Financial Risk Management: Measurement and Theory Advancing Practice*. Princeton, NJ: Princeton University Press, 47–58.

Mapes BM, Foster CS, Kusnoor SV, et al. (2020) Diversity and inclusion for the All of Us research program: A scoping review. *PLOS ONE* 15(7): e0234962.

Mehrabi N, Morstatter F, Saxena N, et al. (2022) A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54(6): 1–35.

Miceli M, Posada J and Yang T (2021) Studying up machine learning data. *ArXiv:*2109.08131.

Mol A (1999) Ontological politics: A word and some questions. In: Law J and Hassard J (eds) *Actor-Network Theory and After*. Oxford: Blackwell, 74–89.

Mol A (2002) *The Body Multiple: Ontology in Medical Practice*. Durham: Duke University Press.

Monk EP (2022) Inequality without groups: Contemporary theories of categories, intersectional typicality, and the disaggregation of difference. *Sociological Theory* 40(1): 3–27.

Offenhuber D (2024) Shapes and frictions of synthetic data. *Big Data & Society.* 11 (2): 20539517241249390.

Mena E and Bolte G on behalf of the ADVANCE GENDER Study Group (2019) Intersectionality-based quantitative health research and sex/gender sensitivity: A scoping review. *International Journal of Equity Health* 18(1): 199.

Parsa RA, Kim JJ and Katzoff M (2009) Application of the truncated distributions and copulas in masking data. In: *Joint Statistical Meetings*, pp. 2770–2780.

Patki N, Wedge R and Veeramachaneni K (2016) The synthetic data vault. In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Montreal, QC, Canada, October 2016, pp. 399–410. IEEE. Available at: http://ieeexplore.ieee.org/document/7796926/ (accessed 14 March 2023).

Placani A (2024) Anthropomorphism in AI: Hype and fallacy. *AI Ethics* 4 (1): 691–698.

Rajabi A and Garibay OO (2022) TabFairGAN: Fair tabular data generation with generative adversarial networks. *Machine Learning and Knowledge Extraction* 4(2): 488–501.

Restar A, Jin H and Operario D (2021) Gender-Inclusive and gender-specific approaches in trans health research. *Transgender Health* 6(5): 235–239.

Reuters (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (accessed 2 April 2020).

Rouhani S (2014) Intersectionality-informed quantitative research: A primer. *American Journal of Public Health* 103 (6): 1082.

Sandvig C, Hamilton K, Karahalios K, et al. (2016) When the algorithm itself is a racist. *International Journal of Communication* 10 (2016): 4972–4990.

Savage N (2023) Synthetic data could be better than real data. *Nature Machine Intelligence*. Epub ahead of print. DOI: 10.1038/d41586-023-01445-8.

Shumailov I, Shumaylov Z, Zhao Y, et al. (2024) AI Models collapse when trained on recursively generated data. *Nature* 631 (8022): 755–759.

Star SL (1990) Power, technology and the phenomenology of conventions: On being allergic to onions. *The Sociological Review* 38(1 Suppl.): 26–56.

Suchman L (2007) *Human-Machine Reconfigurations*. Cambridge: Cambridge University Press.

Suchman L (2023) The uncontroversial 'thingness' of AI. *Big Data & Society* 10(2): 20539517231206794.

Varley T and Kaminski P (2021) Intersectional synergies: Untangling irreducible effects of intersecting identities via information decomposition. *Arxiv* 1–10. DOI: 10.48550/arXiv.2106.10338.

Verma S and Ruben J (2018) Fairness definitions explained. In: 2018 ACM/IEEE International Workshop on Software Fairness. FairWare'18, Gothenburg, Sweden, May 29, 2018. https://doi.org/10.1145/3194770.3194776.

Wang A, et al. (2022) Towards intersectionality in machine learning. In: FAccT '22, June 21–24, 2022.

Washington M (2017) *Race, Rhetoric and Media Studies*. Jackson, MS: University Press of Mississippi.

Wenger E (2024) AI Produces gibberish when trained on too much AI-generated data. *Nature* 631(8022): 742–743.

Xu L, Skoularidou M, Cuesta-Infante A, et al. (2019) Modeling tabular data using conditional GAN. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp.7335–7345. Red Hook, NY: Curran Associates Inc.

Yoon J, Mizrahi M, Ghalaty NF, et al. (2023) EHR-Safe: Generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digital Medicine* 6(1): 141.