Synthetic Data: Legal Implications of the Data-Generation Revolution

Michal S. Gal[*] *& Orla Lynskey*[**]

ABSTRACT: A data-generation revolution is underway. Until recently, most of the data used for algorithmic decision-making was collected from events that took place in the physical world ("collected" data). Yet it is forecast that by 2024, sixty percent of data used to train artificial intelligence systems around the world will be synthetic (!). Synthetic data is artificially generated data that has analytical value. For some purposes, synthetic datasets can replace collected data by preserving or mimicking its properties. For others, synthetic data can complement collected data in ways which increase its accuracy or enhance privacy or security protections. The importance of this data revolution for our economies and societies cannot be overstated. It affects data access and data flows, potentially changing the competitive dynamics in markets where data cannot be easily collected and affecting decision-making in many spheres of our life. In many ways, synthetic data does to data what synthetic threads did to cotton.

This data-generation revolution requires us to reevaluate and potentially restructure our legal data governance regime, which was designed with collected data in mind. As we show, synthetic data challenges the equilibrium erected by existing laws to ensure the protection of competing values, including data utility, privacy, security, and human rights. For instance, by revolutionizing data access, synthetic data challenges assumptions regarding the height of access barriers to data. As such, it may affect the need for and

[*] Professor and Director of the Center of Law and Technology, University of Haifa Faculty of Law, and the former President of the International Association of Competition Law Scholars (ASCOLA).

[**] Associate Professor, London School of Economics. We thank Avigdor Gal, Florencia Marotta-Wugler, Paul Schwartz, Kathy Strandburg, participants in the Intellectual Property Scholars Annual Conference, Cambridge IP and Information Law Center workshop, NYU Privacy Regulation seminar, NYU Engelberg Center for Innovation Law and Policy faculty workshop, Berkeley Topics in Privacy Law seminar, and Cornell Tech Digital Initiative workshop, for excellent discussions or comments on previous drafts. Daniel Greenberg, Adiel Eithan Mustaki and Stav Zeitouni for excellent research assistance. This work was supported by NYU School of Law research assistance funds, as well as the Israel Science Foundation (grant no. 2737/20). Any mistakes or omissions remain the authors'.

the application of antitrust and direct regulation to some firms whose comparative advantage is data-based.

Even more importantly, by potentially making data about individuals more granular, and by increasing the accuracy and completeness of data used for decision-making about individuals, synthetic data also challenges the governance structures and basic principles underpinning current privacy laws. Indeed, many argue that synthetic data does not constitute personal data, and thus avoids the application of privacy laws. We challenge this claim. We also show that synthetic data exposes deep conceptual flaws in the data governance framework. It raises fundamental questions, such as whether data which is not linked to a person in the original dataset should still be treated as personal data, and how inferences based on collected data should be treated.

We then reevaluate the justifications for legal requirements regarding data quality, such as data completeness and accuracy, as well as those relating to fair and informed decision-making, such as data transparency and explainability. The claim is often made that such obligations enhance social welfare. Yet, as we show, synthetic data changes the balance between the protected values, potentially leading to different optimal legal requirements in different contexts. For example, where synthetic data significantly increases consumer welfare, yet the underlying processes are not easily explained, requirements to look under the hood of datasets and provide a detailed explanation of what led to the decision might not always be welfare-maximizing.

This Article seeks to bring state-of-the-art data generation methods into the legal debate and to propose legal reforms which capture the unique characteristics of synthetic data. While some of the challenges discussed here also arise with the use of collected data, synthetic data puts these challenges on steroids.

INTRODUCTION 1089

I. TECHNOLOGICAL BACKGROUND 1094
  A. *WHAT IS SYNTHETIC DATA?* 1094
  B. *SYNTHETIC DATA GENERATION* 1095
    1. Generation Based on Transformations of Collected Data 1095
    2. Generation Methods Which Reduce the Need for Collected Data 1098
    3. Synthetic Data Generation Without (Direct) Use of Collected Data 1100
    4. Typology of Synthetic Datasets 1102
  C. *BENEFITS AND COSTS OF SYNTHETIC DATA* 1102

II. EFFECTS ON COMPETITIVE DYNAMICS AND MARKET POWER 1110
  A. *DATA-BASED MARKET POWER VIS-À-VIS COMPETITORS* 1110

B. IMPLICATIONS FOR ANTITRUST AND REGULATION OF PLATFORMS 1115

III. EFFECTS ON DATA PRIVACY 1121
A. THE IMPACT OF SYNTHETIC DATA ON BALANCING OF INTERESTS IN PRIVACY LAWS 1122
B. APPLICATION OF PRIVACY LAWS TO SYNTHETIC DATA 1126
C. ARE DATA PROTECTION LAWS FIT FOR PURPOSE? 1137
1. Challenges Arising from Categorizing Data 1138
2. Limited Ability to Capture Spillover Effects 1139
3. Collective Data Harms 1141

IV. EFFECTS OF INCREASED DATA QUALITY 1143
A. THE EFFECTS OF SYNTHETIC DATA ON DATA QUALITY 1144
B. APPLICATION OF LAWS 1147

CONCLUSION 1154

INTRODUCTION

Data is an essential input in our digital economies and societies.[1] Generally, the better the data (in terms of volume, variety, veracity, and velocity), the better the learning from it (information and knowledge). While algorithms and infrastructure are important elements of artificial intelligence ("AI"), data is a critical element in such value creation.[2]

Data is traditionally collected[3] from the physical world (hereinafter: "collected data").[4] Some types of collected data are relatively abundant and

1. JACQUES CRÉMER, YVES-ALEXANDRE DE MONTJOYE & HEIKE SCHWEITZER, EUR. COMM'N, COMPETITION POLICY FOR THE DIGITAL ERA 73 (2019).

2. *See, e.g.*, Jeremy Kahn, *Deep Learning Pioneer Andrew Ng Says Companies Should Get 'Data-Centric' to Achieve A.I. Success*, FORTUNE (June 21, 2022, 1:44 PM), https://fortune.com/2022/06/21/andrew-ng-data-centric-ai [https://perma.cc/EB29-E73W]. The centrality of data to automated decision-making is also recognized in legal and ethical literature. *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 673–74 (2016); David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 662–64 (2017); Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter & Luciano Floridi, *The Ethics of Algorithms: Mapping the Debate*, BIG DATA & SOC'Y, July–Dec. 2016, at 1, 1–2.

3. The selection of data to be collected and the characteristics of the collection method all affect the collected data, so it constitutes an incomplete representation of the real world. *See generally* LISA GITELMAN ET AL., "RAW DATA" IS AN OXYMORON (2013) (stating data collection is always based on a choice of which data to collect); Marion Fourcade & Kieran Healy, *Seeing Like a Market*, 15 SOCIO-ECON. REV. 9 (2017) (describing how data collection decisions affect the data collected). The definition of collected data used in this Article does not clash with this truism.

4. Different fields use different terms to define such data. These may include, for example, natural or original data.

easy to access and use, such as weather conditions. Yet many types of data are characterized by high access barriers, such as how people reacted to a specific health treatment, or the only set of historical photos of an area that burned down.[5] To be useful, data must be collected, cleaned and prepared, analyzed, and stored—any of which may be prohibitively costly or even impossible, at least for some.[6] Accordingly, those "who possess [such] data . . . may enjoy competitive comparative advantage[s]," potentially providing them with data-based market power which can be exercised or even abused.[7]

But what if some types of data could be created without having to actively collect (all of) it from the real world? To overcome data access hurdles, data scientists make use of synthetic data: artificial data, generally generated by computer simulations or algorithms, which has analytical value.[8] Such techniques use autonomous generation models or inferences from collected data. Synthetic data has numerous uses. First, it often mimics collected data—augmenting or replacing it. Such data is often used for training or testing machine learning AI;[9] prominent examples include the voice recognition algorithm used in Amazon's Alexa and Google's (now Waymo's) autonomous cars.[10] Second, it is used to potentially reduce bias or to overcome statistical imbalance in representative examples; it may thereby increase the quality of decision-making.[11] Third, it is used to increase privacy or cybersecurity, thereby enabling wider use of valuable data for research and decision-making. For example, the U.S. Census Bureau transforms some of its data into synthetic data to enable access.[12] Similarly, synthetic data can increase levels of privacy protection while enabling the operation of smart cities.[13]

---

5. *See* Daniel L. Rubinfeld & Michal S. Gal, *Access Barriers to Big Data*, 59 ARIZ. L. REV. 339, 345 (2017).

6. *See id.* at 353.

7. *Id.* at 342.

8. We adopt the definition used by computer scientists. For the importance of maintaining analytical value, see Donald B. Rubin, *Statistical Disclosure Limitation*, 9 J. OFF. STAT. 461, 462 (1993). Synthetic data is not to be confused with AI. While AI can be used to create synthetic data and can be trained on it, the two are not identical.

9. Elise Devaux, *Types of Synthetic Data and 4 Examples of Real-life Applications (2022)*, STATICE (May 29, 2022), https://www.statice.ai/post/types-synthetic-data-examples-real-life-examples [https://perma.cc/UJJ4-FZGG].

10. *Id.* Other examples involve facial recognition, detection of fraud and money laundering, and prediction of housing markets. *Id.*

11. *See infra* Part I.

12. U.S. CENSUS BUREAU, U.S. DEP'T OF COM., WHAT ARE SYNTHETIC DATA? 1–2 (2021), https://www.census.gov/content/dam/Census/library/factsheets/2021/what-are-synthetic-data/what-are-synthetic-data.pdf [https://perma.cc/X8HN-E2M6].

13. Alex LaCasse, *Synthetic Data a Key to Privacy by Design Practices in New Canadian Smart City Partnership*, IAPP (Nov. 29, 2022), https://iapp.org/news/a/synthetic-data-is-key-to-privacy-by-design-practices-in-new-canadian-smart-city-partnership [https://perma.cc/8SJB-BU8S].

While some forms of synthetic data generation have been around for some time (such as upsampling[14]), the wide use of synthetic data in simulations to train machine learning models is quite new. Yet, as one observer noted, despite the fact that "[s]ynthetic data generation technology is a relatively recent addition to the toolkit of machine learning engineers. . . . [I]t has already evolved from the initially supportive role of augmenting real-world data to one enabling a new wave of AI innovation."[15] This Article focuses mainly on such uses.

Synthetic data has ingrained benefits. It can potentially reduce the costs involved in all stages of the data value chain, obviating the need for excessive data collection, costly cleaning and preparation, and long-term data storage. Generation-for-purpose may also shorten the time necessary to generate useful data. For these reasons, it is forecasted that "[b]y 2024, [sixty percent (!)] of the data used" to train AI systems around the world will be synthetic.[16] "Gartner predict[ed that] by 2030, 'you won't be able to build high-quality, high-value AI models without synthetic data.'"[17] In many ways, synthetic data has the potential to do to data what synthetic threads did to cotton.

The importance of this data-generation revolution cannot be overstated. As we show, synthetic data may change the current balance between data utility and competing considerations. It also affects existing power relationships, both between competitors and between providers/suppliers and users/consumers. Synthetic data is thus poised to affect all spheres of our lives that involve data-based decision-making, including the economic, the social, and the political. Such effects on the welfare of individuals and societies may be both positive and negative. While many of these effects also arise with regard to collected data, synthetic data puts them on steroids.

Given the importance of this revolution, this Article seeks to identify and critically examine the effects of synthetic data on key data governance challenges. In particular, it focuses on three main issues that stand at the basis of our legal data regime, and that are significantly affected by synthetic data: data access, data privacy, and data accuracy. For each, it queries whether the

14. Upsampling is the process of adding more data points of a certain type to the dataset, usually to reduce unbalanced sampling. Nandhini Nallamuthu, *Handling Imbalanced Data – Machine Learning, Computer Vision and NLP*, ANALYTICS VIDHYA (Apr. 4, 2023), https://www.analyticsvidhya.com/blog/2020/11/handling-imbalanced-data-machine-learning-computer-vision-and-nlp [https://perma.cc/S7FX-C42D].

15. Andrey Shtylenko, *The Advantages of Synthetic Data*, LINKEDIN: THE REALITY GAP (Nov. 23, 2022), https://www.linkedin.com/pulse/advantages-synthetic-data-andrey-shtylenko?trk=news-guestshare-article [https://perma.cc/V5NA-VVAM].

16. Andrew White, *By 2024, 60% of the Data Used for the Development of AI and Analytics Projects Will Be Synthetically Generated*, GARTNER (July 24, 2021), https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated (on file with the *Iowa Law Review*).

17. Clayton Nicholas, *Accelerating Innovation with AI Using Synthetic Data*, VIBRONYX, https://vibronyx.com/accelerating-innovation-with-ai-using-synthetic-data [https://perma.cc/KMB7-AMG7].

existing legal regime is fit for purpose and able to address the governance challenges through either current application or a modified interpretation of the law or whether the regulatory toolbox should be updated.

With respect to the first issue, by lowering some access barriers to data, synthetic data affects data access and data flows, potentially changing the competitive dynamics in markets where such access constitutes a significant barrier. Importantly, synthetic data can potentially overcome comparative advantages resulting from data-based network effects and feedback loops, which allow entities that already possess large datasets to keep ahead on the data learning curve, continually improving their decision-making relative to others[18] and entrenching their data-based market power.[19] In some circumstances synthetic data can potentially break such self-perpetuating loops. There is, of course, the countervailing risk that, in cases where collected data is unique and essential for synthetic data generation, the advent of synthetic data could instead increase the comparative advantages of those controlling collected data. However, such instances are becoming less common.[20]

Despite its potential, the effects of synthetic data on competition have not, as of yet, been recognized by academics, legislators, and regulators. Take, for example, calls for stringent regulation of large digital firms. As elaborated below, new regulations are being suggested, and cases are being brought, based on assumptions of data-based market power that are no longer true for some markets.[21] Some proposed regulations that may already be obsolete in certain markets include mandatory data sharing, portability, interoperability, and standardization.[22] Recognizing the effects of synthetic data may lead to a more nuanced, hands-off regulatory approach in this legal realm.

In contrast, the effects of synthetic data on meeting the requirements of data privacy laws have been acknowledged. Yet most of the discussion around this issue so far has been misleading or simplistic. For instance, it is often argued that synthetic data—based on transformations of collected data—can overcome the constraints on data use imposed by privacy laws.[23] Implicit in these claims is the assumption that synthetic data processing does not pose a

18. U.K. DIGIT. COMPETITION EXPERT PANEL, UNLOCKING DIGITAL COMPETITION: REPORT OF THE DIGITAL COMPETITION EXPERT PANEL 33 (2019), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/785547/unlocking_digital_competition_furman_review_web.pdf [https://perma.cc/979X-B7G5].

19. *See* STIGLER COMM. ON DIGIT. PLATFORMS, STIGLER CTR. FOR THE STUDY OF THE ECON. & THE STATE, FINAL REPORT 34–36 (2019), https://www.chicagobooth.edu/-/media/research/stigler/pdfs/digital-platforms—committee-report—stigler-center.pdf [https://perma.cc/B4XU-C4UN].

20. SERGEY I. NIKOLENKO, SYNTHETIC DATA FOR DEEP LEARNING 8 (2021).

21. *See infra* Section II.B.

22. *See infra* Section II.B.

23. STATICE, ANONYMIZATION AND DATA PRIVACY WITH STATICE: GUIDE TO GDPR COMPLIANCE 4, https://privacy.statice.ai/hubfs/Resources/brochures/Anonymization_data_privacy_Statice.pdf [https://perma.cc/2H7T-UU2U].

risk to privacy and related rights, such as autonomy, nondiscrimination, and human dignity. Yet, as we show, this is not the case. To the contrary: synthetic data may sometimes exacerbate such harms, both amplifying traditional harms and creating novel ones. One extreme example involves deep fakes that are presented as real, in which at least part of the data used to create fake images is synthetic.[24] But, more commonly and thus potentially more importantly, synthetic data increases data externalities and collective data harms. In particular, it accentuates the fact that not only is data itself intangible and nonrivalrous,[25] but that the learnings from data also share such characteristics. Yet, as we show below, such effects are often not regulated by privacy laws.

Synthetic data can also potentially increase the quality of a dataset, thereby raising the quality of data-based decision-making. Improved data quality amplifies the benefits, but also the risks, associated with more granular data.[26] The latter include, inter alia, a better ability to profile, nudge, exploit and manipulate individuals, with ramifications for the interpersonal, commercial, social, and political spheres.[27] As we show, the legal framework that currently applies to increased data quality does not create an optimal balance between the competing considerations, especially once synthetic data is added to the mix. Most laws that relate to data quality mandate *increased* accuracy, rather than less, and many harmful uses of accurate data are not regulated.[28]

This reality requires us to reevaluate our regulatory tools, which were designed with collected data in mind. To name but a few challenges, it calls upon us to consider a shift in the focus of data governance models from data collection to its uses and effects, from user consent and control to notions of welfare and well-being, and from private data to inference data and to collective data harms. It also requires us to rethink the current balance between data utility, privacy, security, and human rights, and the tools currently used to protect this balance. For example, it challenges the explainability requirements for AI-based decisions, potentially moving the focus from causality to generation-process reliability,[29] and it affects the

24. Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1760 (2019).

25. Bertin Martens, *An Economic Perspective on Data and Platform Market Power* 21 (JRC Digit. Econ., Working Paper No. 2020-09, 2021), https://joint-research-centre.ec.europa.eu/system/files/2021-02/jrc122896.pdf [https://perma.cc/7DMU-N7E9]. This implies that many users can potentially use the same data at the same time. *See id.*

26. For an excellent discussion of algorithmic accuracy, which has implications for data accuracy, see generally Aileen Nielsen, *The Too Accurate Algorithm* (Ctr. for L. & Econ. Working Paper Series, Working Paper No. 09/2022, 2022), https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/572429/CLE_WP_2022_09.pdf [https://perma.cc/U3CM-Y973] [hereinafter Nielsen, *The Too Accurate Algorithm*].

27. *See infra* Part IV.

28. *See infra* Part IV.

29. *See infra* Part IV.

interpretability of reasonableness principles in contracts and laws pertaining to the collection and use of data. This Article takes a first step in this direction.

We begin in Part I by providing a brief overview of synthetic data: its definition, main production techniques, costs and benefits, and how it is being used in practice. We then focus on the three main effects of synthetic data: data access, data privacy, and data quality. Part II charts its effects on competition and market dynamics, which, in turn, affect the functioning of markets in the data economy. Part III analyzes its impact on data privacy. Part IV focuses on the broader societal implications of more granular, complete, and representative synthetic data. All Parts address the extent to which the existing legal landscape responds adequately to the changes wrought by synthetic data. As such, this Article adds to the growing literature calling into question the ability of existing legal frameworks to respond to the challenges of new data-related techniques. As we show, a new balance is needed, which would allow us to enjoy the significant competition, innovation, privacy, security, human rights, and data quality benefits that this data revolution creates, while addressing the concerns it raises in these spheres. The Conclusion offers avenues for further research.

I. TECHNOLOGICAL BACKGROUND

Opening the synthetic data "black box" and looking under its hood is essential for determining whether our current laws can ensure that its uses increase social welfare. Accordingly, this Part explores what synthetic data is, how it is generated, and its benefits and limitations relative to collected data.

A. WHAT IS SYNTHETIC DATA?

Synthetic data is artificially generated data with analytical value.[30] Its generation can be based on collected data or on assumptions made by the coder (be it a human or AI) about the different variables in the dataset.[31] Accordingly, the main difference between collected and synthetic data is their source: while the former is collected from the real world, via human or technological sensors, the latter is generated artificially.

Synthetic data generation is a general-purpose technology that can be employed in numerous spheres and industries, including health, transportation, and finance.[32] Its flexibility is also reflected in the variety of outputs that it can give rise to, including datasets, images, audio files, and videos. The resulting datasets can be fully or partially synthetic. While synthetic data has been used for a while, it has been significantly developed

30. *See supra* note 8 and accompanying text.

31. Rubin, *supra* note 8, at 465–67.

32. HARVARD BUS. REV. ANALYTIC SERVS., THE EXECUTIVE'S GUIDE TO ACCELERATING ARTIFICIAL INTELLIGENCE AND DATA INNOVATION WITH SYNTHETIC DATA 6 (2021), https://f.hubspotusercontent20.net/hubfs/4408323/HBR%20campaign/HBR%20Analytic%20Services%20Synthetic%20Data.pdf [https://perma.cc/5J82-HG5W].

recently, partly due to advancements in technologies for processing, analyzing, and storing data.[33]

B. SYNTHETIC DATA GENERATION

Numerous methods for generating synthetic data exist. We group these methods into three categories, which are distinguished by their need for collected data (which could be public, private, or a combination thereof) in the generation process. This parameter enables us to explore the extent to which data collection barriers can be overcome, as well as the potential use of private data in the generation process, two conditions that affect our analysis below. The first group requires the same amount of collected data as traditional methods used to create similar outputs. The second reduces the amount or the quality of collected data necessary. The third does not require any direct use of collected data. While the different methods can be combined in the data generation pipeline, here we relate to them in their pure form, to emphasize their distinctive characteristics. In each example, we also explore their main uses and whether new information—defined here as information that could not be directly learned from the input data used in its generation—is created. Our analysis mainly focuses on synthetic data that is generated by machine learning algorithms and that can be created *en masse*.

1. Generation Based on Transformations of Collected Data

Synthetic data can be generated based on transformations of collected data. While such data serves important goals—most importantly enabling better extraction of information from the collected data, or wider sharing of the data by deidentification—it does not significantly reduce the need for collected data or change the features such data must contain. This group of methods have been used for quite some time.[34]

Synthetic data is often generated as a stage in the data flow pipeline, geared toward extracting more information from collected data. Consider two examples. In the first, as collected data is cleaned and prepared for analysis, some of the values in the dataset are replaced with synthetic values to ensure consistency (a process called data curation).[35] For instance, if most data points relate to minutes, those that relate to hours can be replaced by synthetic data to correct the inconsistency. In the second example, the synthetic data results from the analysis performed on the collected data. Take, for example, *collaborative filtering*, which is based on computing similarities between clusters. To illustrate, assume that a seller wishes to create a dataset capturing the

33. *See, e.g.*, KHALED EL EMAM, ACCELERATING AI WITH SYNTHETIC DATA: GENERATING DATA FOR AI PROJECTS 56 (2020).

34. *See, e.g.*, Rubin, *supra* note 8, at 461; NIKOLENKO, *supra* note 20, at 139–59.

35. Mary K. Pratt, *Definition: Data Curation*, TECHTARGET (Jan. 2022), https://www.techtarget.com/searchbusinessanalytics/definition/data-curation [https://perma.cc/ZE8C-4Z9W].

probability that a buyer might be interested in different items.[36] The coder first identifies two types of clusters in the collected data: clusters of relatively similar users and clusters of relatively similar items. He then calculates two types of attributes: attributes of users (e.g., demographic information) and attributes of items (e.g., film genres). The algorithm then determines the distribution of user clusters and item clusters and affiliates them with user/item attributes respectively. These generated correlations create a partially synthetic dataset, where synthetic data fills in the gaps in the collected dataset, enabling learnings from the input data to be more easily extracted or conveyed.

More interesting for our purposes is the generation of new, wholly synthetic datasets, based on transformations of the collected data. Such transformations serve many useful purposes, including transfer or storage of data, where costs depend on the volume of the data, or deidentification of data. The basic idea behind this method is quite simple: computing the (main) statistical characteristics of the original dataset and creating a synthetic one with quite similar characteristics. It involves the following main steps.[37] The first is data preparation: cleaning the collected data to remove errors, ensuring that all fields in the dataset use consistent coding schemes, and confirming that data from multiple sources is mapped into the same data typology.[38] The next step is developing a Data Generator to generate synthetic data based on manipulations of the collected data. The Generator's algorithm computes the metrics for the collected data and then sets the parameters that will be used to generate synthetic data. To maintain logical consistency, some characteristics of the original dataset may need to be checked (for example, no biological male can have a positive value for "pregnant"). The third step is computing metrics for the synthetic data. Finally, the metrics of the collected and the synthetic data are compared using a Discriminator. This step assesses the utility of the synthetic dataset by determining whether its statistical properties are (relatively) similar to those of the original set.[39] If the Discriminator finds that the synthetic data can be distinguished from the collected data, the process adjusts the generation parameters and generates new synthetic data.[40] The process repeats until the Generator produces acceptable synthetic data.[41] "These utility comparisons can be formalized using various similarity metrics so that they are repeatable and automated."[42] An additional, optional step involves a feedback loop which refines the

36. *See* Karen Tso & Lars Schmidt-Thieme, *Attribute-Aware Collaborative Filtering*, *in* FROM DATA AND INFORMATION ANALYSIS TO KNOWLEDGE ENGINEERING 614, 614 (2006).

37. *Synthetic Data*, JPMORGAN CHASE & CO., https://www.jpmorgan.com/technology/artificial-intelligence/initiatives/synthetic-data [https://perma.cc/5WLC-CNQT].

38. *Id.*

39. EL EMAM, *supra* note 33, at 12.

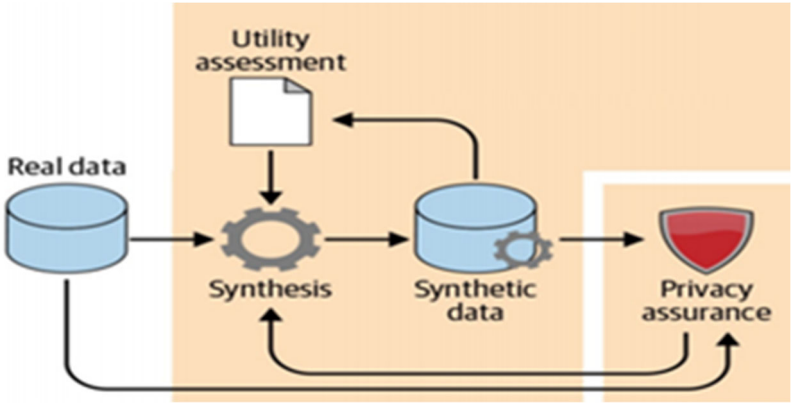
40. *Id.*

41. *Id.*

42. *Id.* at 18–19.

Generator to improve the synthetic data against the relevant comparison metrics.[43] Where privacy concerns arise, a privacy assurance assessment can be added, to ensure that privacy risks remain below a certain benchmark.[44]

Figure 1: Data Generation Process When Collected Data Is Used as an Input[45]



The Data Generator can use different techniques, such as decision trees or deep learning.[46] The exact choice is driven by the characteristics of the collected data, including its complexity and the level of data utility desired.[47] A common model involves variational autoencoders ("VAE").[48] VAE is a two-step unsupervised machine learning method which results in "a meaningful representation of a multidimensional dataset"[49]: First, the original complex distribution is transformed, using the encoder, into a more compact representation with fewer dimensions.[50] Second, the decoder then takes that compressed representation and reconstructs the original input data, generating synthetic data.[51] "The VAE is trained by optimizing the similarity between the [synthetic] data and the input data."[52] VAEs "are relatively *easy to*

43. *Id.* at 12.

44. KHALED EL EMAM, LUCY MOSQUERA & RICHARD HOPTROFF, PRACTICAL SYNTHETIC DATA GENERATION: BALANCING PRIVACY AND THE BROAD AVAILABILITY OF DATA 144 (2020).

45. *Id.* at 39 fig.2.14.

46. NIKOLENKO, *supra* note 20, at 97–102.

47. James Chen, *What Is a Neural Network?*, INVESTOPEDIA (Apr. 30, 2023), https://www.investopedia.com/terms/n/neuralnetwork.asp [https://perma.cc/SH2P-W98P].

48. Christoph Wehmeyer, *How Do You Generate Synthetic Data?*, STATICE (Feb. 11, 2021), https://www.statice.ai/post/how-generate-synthetic-data [https://perma.cc/2XBG-G89U].

49. EL EMAM ET AL., *supra* note 44, at 107.

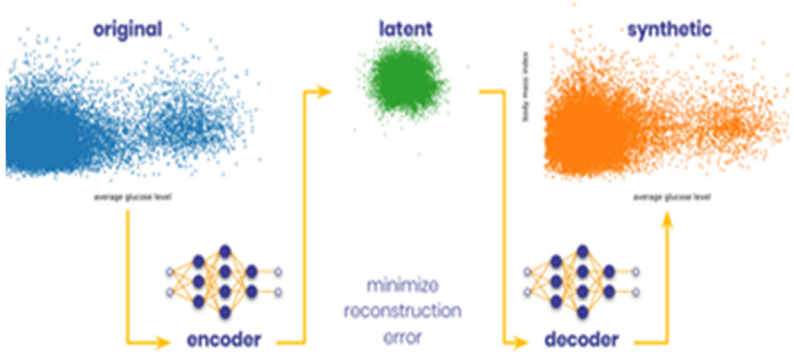
50. Wehmeyer, *supra* note 48.

51. *Id.*

52. EL EMAM ET AL., *supra* note 44, at 107.

implement and to train."[53] Yet as the original "data becomes more heterogeneous, . . . it also becomes more *difficult to formulate a reconstruction* [*method*] *that works well on all data* [*points*]."[54]

Figure 2: Variational Autoencoders[55]



Common to all such techniques is the fact that the information which can be learned from the synthetic dataset is wholly based on the input data.

2. Generation Methods Which Reduce the Need for Collected Data

The second group of generation methods are those which reduce the amount or quality, or change the features, of collected data necessary to achieve a given result. A prominent example of this group is generative adversarial networks ("GANs"), which employs two neural networks (deep learning algorithms) pitted against each other in an adversarial fashion called a zero-sum game.[56] The first network—the Generator—generates synthetic data without directly using the collected data (this process is explained in the next section).[57] The generated data is then sent to the second neural network—the Discriminator—which was trained on collected data.[58] The Discriminator compares the synthetic data with the collected data, creating a propensity score and determining which parts of the data give away its

53. Wehmeyer, *supra* note 48.

54. *Id.*

55. *Id.*

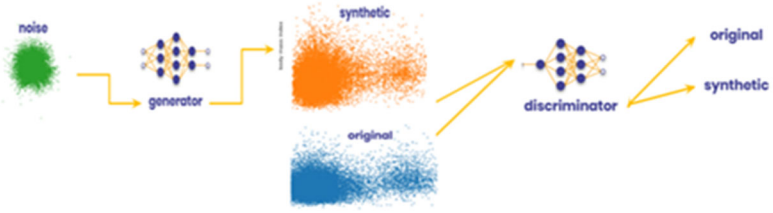
56. *Id.*; *Overview of GAN Structure*, GOOGLE FOR DEVS. (July 18, 2022), https://developers.google.com/machine-learning/gan/gan_structure [https://perma.cc/5DAHZFQA].

57. *Id.*; *see infra* Section I.B.3.

58. *Overview of GAN Structure*, *supra* note 56.

"fakeness."[59] The result is then fed back to the Generator. A good synthetic model is created when the Discriminator is unable to distinguish between the collected and synthetic datasets.[60] The main weakness of GANs is that they are more challenging to train than VAEs.[61]

Figure 3: Generative Adversarial Networks[62]



This approach is particularly useful for synthetic image generation.[63] For example, assume that a Generator has created a fake image of a stop sign. The Discriminator might determine that the image is fake based on mismatches between the fake image and collected data available to the algorithm (e.g., the color of the fake image may not be coherent with the algorithm's learning from the collected data which conveys the actual color of stop signs).[64] Such feedback is fed into the Generator, which generates an updated image. The process repeats until the Discriminator can no longer identify that the image is fake.

Another use is upsampling—a technique for increasing the granularity or quality of an output by adding new data points between existing points.[65] Upsampling can be used to reduce bias that might result from training an automated system on biased or incomplete datasets.[66] To illustrate, consider Amazon's famous attempt to train an algorithm to rate job applicants for technical posts. The algorithm was trained on "resumes submitted to the company" in previous years, which reflected male dominance in the

59. Wehmeyer, *supra* note 48.

60. *Overview of GAN Structure*, *supra* note 56.

61. Wehmeyer, *supra* note 48.

62. *Id.* The term "original" can refer to "collected."

63. For example, the website https://thispersondoesnotexist.com includes images of people who do not exist, created by GAN. It is impossible for a human observer to determine whether the image is real.

64. EL EMAM ET AL., *supra* note 44, at 70.

65. *See supra* note 14 and accompanying text.

66. *See supra* note 14 and accompanying text.

industry.[67] The result was that it judged male applicants as superior, and penalized references in resumes which indicated the applicant was a woman (e.g., women's football captain).[68] Upsampling could counter such bias by adding synthetic data representing the resumes of successful female applicants. Those can be created by a Generator and vetted by a Discriminator which learned from a few real resumes of successful women.

Both uses "extend the . . . available dataset with transformations that do not change the properties that [one wishes] to learn" (data augmentation).[69] Accordingly, the information that can be directly learned from the synthetic data created is wholly based on the input data (e.g., the data does not suggest new characteristics of stop signs). At the same time, new information can be learned from the *process* of creating the synthetic data (such as what elements in the collected data are the most efficient differentiating factors of stop signs, or what elements can "trick" the Discriminator). This new learning can help reduce training and error costs. Such synthetic outcomes are often used as inputs for simulations, to which we turn next.

3. Synthetic Data Generation Without (Direct) Use of Collected Data

Some types of synthetic data can be generated without (direct) use of collected data, implying that in the iteration that created the relevant dataset, the algorithm did not use such data. This is done via a data simulator. Such a simulator generates synthetic data based on a set of rules which determine the relationships between the relevant data attributes.[70] The complexity of the generation method chosen is affected by the potential use of the dataset. As noted, such simulators have recently become the main tool for training and testing machine learning algorithms.[71]

Some simulators do not require collected data at all. Consider a simple example: creating a dataset of numbers to train an algorithm that organizes numbers sequentially. A synthetic data simulator which picks numbers randomly will suffice.

Despite the fact that such simulations do not require collected data, they can create new information. To illustrate, let us contrast two well-known examples. IBM's Deep Blue trained an algorithm to play chess by feeding it numerous examples of winning strategies (collected data).[72] The algorithm

67. Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, REUTERS (Oct. 10, 2018, 6:04 PM), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G [https://perma.cc/S4Z2-XTFB].

68. *Id.*

69. NIKOLENKO, *supra* note 20, at 6.

70. EL EMAM, *supra* note 33, at 8–9, 54.

71. Nicholas, *supra* note 17.

72. Joanna Goodrich, *How IBM's Deep Blue Beat World Champion Chess Player Garry Kasparov*, IEEE SPECTRUM (Jan. 25, 2021), https://spectrum.ieee.org/how-ibms-deep-blue-beat-world-champion-chess-player-garry-kasparov [https://perma.cc/BL9P-TU4C] (chess computer).

famously defeated world chess champion Garry Kasparov in a six-game match.[73] Contrast this with AlphaGo, developed by DeepMind—a subsidiary of Google (now Alphabet)—which was taught to play the board game Go.[74] The algorithm was first fed the rules of the game.[75] It was mainly trained by creating numerous simulations of games where the algorithm played against other instances of itself, using reinforcement learning to improve its play.[76] Each algorithm was, in effect, creating the synthetic dataset of moves from which the other algorithm learned.[77] AlphaGo proved itself by defeating human world champions.[78] But, more importantly, it developed new strategies for playing Go, thereby adding new learning.[79]

Also interesting for our analysis are cases in which collected data is indirectly used, in that the simulation model relies on prior exposure of the coder (human or AI) to such data (i.e., background knowledge). The synthetic data is then based on the coder's assumptions regarding the statistical properties of the relevant data attributes.[80] For example, the coder may base the maximum speed humans are shown reaching in synthetic videos on his real-world observations of human locomotion. Background knowledge can also be based indirectly on collected data, such as when the learning from collected data in another context is embedded in the model (transfer learning). Take an example from retail: companies can "use 3-D simulations to . . . create a synthetic dataset [containing] a thousand images" from "as few as five [actual] images of a product."[81] To do this, they employ existing knowledge about how different shapes look from different angles which was learned from previous tasks.[82]

---

73. *Id.*

74. *AlphaGo*, GOOGLE DEEPMIND, https://deepmind.google/technologies/alphago [https://perma.cc/9P2J-XGCW].

75. *Id.*

76. *Id.* While it was first given examples of moves by expert players from recorded historical games (collected data), this was mainly done to shorten the time the algorithm needed to reach a certain level of proficiency in the game. *Id.*

77. David Silver & Demis Hassabis, *AlphaGo: Mastering the Ancient Game of Go with Machine Learning*, GOOGLE RSCH.: BLOG (Jan. 27, 2016), https://blog.research.google/2016/01/alphago-mastering-ancient-game-of-go.html [https://perma.cc/MPY3-E5JD].

78. *AlphaGo*, *supra* note 74.

79. *Id.*

80. EL EMAM ET AL., *supra* note 44, at 3.

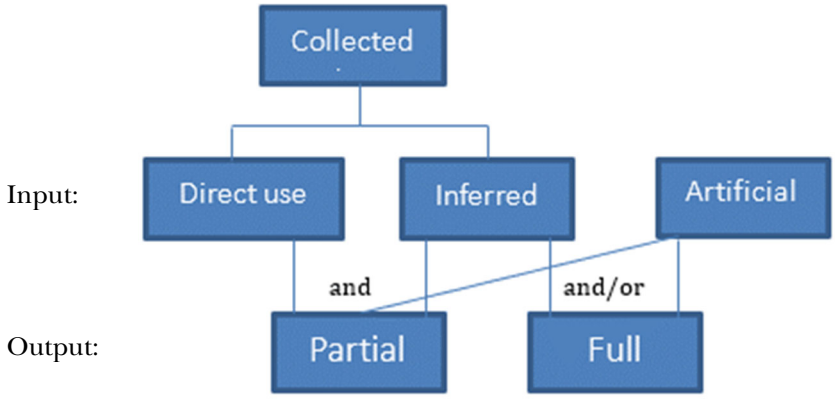
81. Gerard Andrews, *What Is Synthetic Data?*, NVIDIA (June 8, 2021), https://blogs.nvidia.com/blog/2021/06/08/what-is-synthetic-data [https://perma.cc/T8AD-7EPC]. These datasets are part of a range of technologies, including computer vision and geofencing, which underlie smart stores—physical stores where customers pay for goods via an app rather than interacting with a cashier or checkout machine. *Id.*

82. *See id.*

4. Typology of Synthetic Datasets

The analysis above serves as a basis for the following typology of synthetic data, based on the type of input data used and the resultant synthetic data output:

Figure 4: A Typology of Datasets



This typology also echoes the fact that different types of data may relate to different stages in the data processing pipeline, creating a causal chain whereby (temporary) datasets influence each other.[83] For example, collected data can be used as a (partial) basis for creating synthetic data, which can then be used in simulations that create a different form of synthetic data. A good example is training autonomous vehicles: a small set of collected images is used either in the Generator or the Discriminator to create many synthetic images, which are relevant for simulations of road conditions for machine learning training purposes.

C. *Benefits and Costs of Synthetic Data*

Why is artificial data generation used? We identify three main reasons that provide a wider context for the generation methods just explored.

The first reason, which is relevant to the second and third types of generation methods explored above, is that it allows for *replacing collected data characterized by high access barriers*. While synthetic data is not a panacea, it may reduce data-related barriers in all parts of the data value chain: collection–preparation–analysis–storage–use. By enabling generation-for-use, synthetic data can lower the costs and the time involved in collecting the relevant data. This is especially important given that machine learning algorithms are

83. Sebastian Benthall, *Situated Information Flow Theory*, 6 ANN. HOT TOPICS SCI. SEC. 39, 39, 43–44 (2019).

generally data-intensive,[84] and many types of datasets are expensive or hard to find.[85] Also, as Fromer notes, cloud computing, machine learning, and automation of tasks increase the secrecy of input data, thereby making it harder to access indirectly.[86] To overcome such barriers, firms might be able to produce synthetic data internally or obtain it from third parties who specialize in such data production.[87]

Synthetic data also reduces the resources needed for preparation of the raw data for analysis, which involves, inter alia, cleaning, labeling, and organizing the raw data. Such tasks can be complex, laborious, and expensive.[88] In particular, manual labeling "is often costly, generally time-consuming, and error-prone."[89] By labeling and organizing the data automatically during the generation process, synthetic data combines data collection and preparation, creating data that is fit for purpose from the start.[90] This is especially important for machine learning algorithms, where the scale of datasets can reach hundreds of thousands and even millions of data points. One entrepreneur estimated that "[a] single image that could cost [six dollars] from a labeling service can be artificially generated for six cents."[91] In the data analysis stage, synthetic data can be used to make the analysis more efficient. Interestingly, synthetic data can be used to train algorithms so that they "make [synthetic data] more suitable for training."[92]

Synthetic data can also potentially reduce storage costs in four main ways. First, if a synthetic dataset can be easily recreated, its user does not need to store data for future use.[93] Indeed, synthetic data can be generated only when needed (what data scientists call "lazy production"). Second, and relatedly, generation-for-purpose reduces the need to store data for long periods before enough is accumulated for meaningful analysis. Third, generation-for-purpose reduces the amount of redundant data that might otherwise be included in the dataset. This is especially important where the relevant

84. Open data sources are often limited in their availability or utility. Marco Iansiti, *The Value of Data and Its Impact on Competition* 4 (Harvard Bus. Sch., Working Paper No. 22-002, 2021), https://www.hbs.edu/ris/Publication%20Files/22-002submitted_835f63fd-d137-494d-bf37-6ba5695c5bd3.pdf [https://perma.cc/CM3H-TQ4D].

85. NIKOLENKO, *supra* note 20, at 7, 12.

86. Jeanne C. Fromer, *Machines as the New Oompa-Loompas: Trade Secrecy, the Cloud, Machine Learning, and Automation*, 94 N.Y.U. L. REV. 706, 718–25 (2019).

87. Markets exist for several types of synthetic data. *See, e.g.*, Elise Devaux, *List of Synthetic Data Startups and Companies—2021*, MEDIUM (Mar. 23, 2021), https://elise-deux.medium.com/the-list-of-synthetic-data-companies-2021-5aa246265b42 [https://perma.cc/C7AH-46X9].

88. NIKOLENKO, *supra* note 20, at 1–4. While "[t]here exist large open datasets for many" uses (such as ImageNet), many of them are only labeled for certain uses, or may contain inherent labeling biases. *Id.* at 1–3.

89. Shtylenko, *supra* note 15.

90. *Id.*

91. Andrews, *supra* note 81.

92. NIKOLENKO, *supra* note 20, at 13.

93. For such barriers, see Rubinfeld & Gal, *supra* note 5, at 363–64.

information cannot easily be parsed from collected datasets that take up a lot of memory, as in the case of videos. Finally, it can obviate the need for storage in cases where the user does not know a priori precisely what kinds of data might be needed for his analysis, yet such data can be synthetically generated in the future.

The generation of synthetic data might also reduce obstacles to use, in that some limitations—such as legal prohibitions on certain uses of personal data—may not apply to synthetic data, a point we return to later.[94] Furthermore, where it can be internally rather than externally generated, synthetic data can overcome some technical or legal obstacles to data transfers.[95]

Importantly, to replace collected data, synthetic data need not be similar to it. Indeed, for some purposes optimal results might require using synthetic data, which does not reflect actual real-world conditions.[96] For example, a dataset used to train autonomous vehicles may stimulate faster and more effective learning if it contains an outsized proportion of risky situations, such as people jumping into the road.

Second, synthetic data can *enrich the data pool with new or higher quality data*, which can augment or replace collected datasets. Such synthetic data allows analysts to study phenomena for which a sufficient amount of collected data cannot be (affordably) collected or where collected data is not easily available for use (e.g., the collected data is not labeled or is labeled incorrectly).[97] Importantly for our analysis, data augmentation may enable the extension of small, context-specific datasets in a way which does not alter essential underlying features of the data, yet creates new, relevant data (as in the case of 3-D simulations mentioned above).[98]

On its face, generating such data seems like a cheap trick. If you can learn correlations from collected data, then why do you need to apply it to more data? Would this not simply generate similar results? Part of the answer is that even if we know all the theoretical parameters in a dataset, *value can be found in creating a new dataset which makes use of interactions between parameters in simulations.* A notable recent example involves the algorithm-generated solution to the protein-folding problem. In 2020, DeepMind unveiled AlphaFold, which uses computer simulations based on background knowledge about proteins, "to accurately and efficiently predict the 3-D shape of an[y] unknown protein

94. *See infra* Part III. It might be interesting to explore how the possibility of generating synthetic data able to achieve relatively similar results to real personal data would affect data subjects' incentives to share their data in the first place. Such a study is beyond the scope of this paper.

95. *See* Rubinfeld & Gal, *supra* note 5, at 350–61.

96. EL EMAM, *supra* note 33, at 8–9.

97. Andrews, *supra* note 81 ("Because synthetic datasets are automatically labeled and can deliberately include rare but crucial corner cases, it's sometimes better than real-world data.").

98. NIKOLENKO, *supra* note 20, at 6, 88.

using just its DNA or RNA source code."[99] "AlphaFold's predictions are so accurate that the protein-folding problem is considered solved after more than [seventy] years of searching," and its open-access database now contains over two hundred million predicted protein structures.[100] The tool and database together comprise a significant scientific breakthrough in the understanding and treatment of human disease.[101] Another example involves the use of synthetic data, based on various real sociodemographic conditions, for modelling micropopulations to evaluate the potential impact of different events, such as the spread of a pandemic.[102]

Another reason synthetic data generation is more than a "cheap trick" is that, as noted above, synthetic data can *lead to new (or more accurate) learnings*. AlphaGo discovering new move sets is a case in point. Or consider the following: some computer scientists are using synthetic data to fix highly accurate but overconfident AI models, which are especially problematic for use in critical applications such as cybersecurity.[103] By adding synthetic data to create counterfactual explanations[104] for points that are not captured by the model's training distribution, or for novel cases that were not included in the collected data, they can test the accuracy of the model in such cases, potentially lowering the level of uncertainty of the model, while retaining and even increasing its predictive value.[105] In a similar way, synthetic data can also be used to *enable new products or services*, such as creating virtual spaces in the metaverse.[106]

Another part of the answer is that *machine learning algorithms can be trained on synthetic data* to increase their accuracy before using them in the real world.[107] For example, Nvidia uses synthetic data to train robots in warehouses to recognize objects of different shapes and sizes in different conditions to

99. Bryan McMahon, *AI Is Ushering in a New Scientific Revolution*, GRADIENT (June 4, 2022), https://thegradient.pub/ai-scientific-revolution [https://perma.cc/B26P-DP9U].

100. *Id.*; Demis Hassabis, *AlphaFold Reveals the Structure of the Protein Universe*, GOOGLE DEEPMIND (July 28, 2022), https://deepmind.google/discover/blog/alphafold-reveals-the-structure-of-the-protein-universe [https://perma.cc/VS6Y-LPKJ].

101. Hassabis, *supra* note 100.

102. *See* NIKOLENKO, *supra* note 20, at 281–82.

103. Sumedha Singla, Nihal Murali, Forough Arabshahi, Sofia Triantafyllou & Kayhan Batmanghelich, *Augmentation by Counterfactual Explanation—Fixing an Overconfident Classifier*, 2023 IEEE/CVF WINTER CONF. ON APPLICATIONS COMPUT. VISION (WACV) 4709, 4710.

104. Counterfactual explanations are a machine learning technique that "describe[] a causal situation in the form: 'If X had not occurred, Y would not have occurred.'" CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING: A GUIDE FOR MAKING BLACK BOX MODELS INTERPRETABLE 240 (2020). They enable learning via consideration of hypothetical scenarios that contradict the observed facts. *Id.*

105. Singla et al., *supra* note 103, at 4709, 4715–16.

106. Victor Dey, *Why the Metaverse Needs Synthetic Data*, VENTUREBEAT (Sept. 30, 2022, 9:05 AM), https://venturebeat.com/ai/deep-dive-how-synthetic-data-can-enhance-ar-vr-and-the-metaverse [https://perma.cc/8QEX-X6M3].

107. NIKOLENKO, *supra* note 20, at 13.

make production lines more efficient.[108] A synthetic dataset was generated using artificial images, based on background knowledge incorporated into the generating algorithm about how lighting affects the appearance of images.[109] The synthetic dataset was then used to train the robot.[110] In another prominent example, Amazon used synthetic data to train Alexa, its digital assistant, to apply voice recognition in Hindi, U.S. Spanish, and Brazilian Portuguese, for which it faced a shortage of collected data.[111] Similar processes and rationales apply in the use of synthetic data to train autonomous vehicles under different road conditions, as discussed above.[112] Indeed, "acquiring and storing [such] data from live tests of real cars on real roads would" be prohibitively expensive.[113] These examples also illustrate why we cannot always simply incorporate previous knowledge into a new algorithm without generating synthetic data: existing knowledge might be both too limited and too complex for efficient application. In the example above, the algorithm operating the robot does not need to account for the myriad of parameters describing how the same image looks under different lighting conditions. Rather, it needs to learn to react to the contours of *any* object under *any* lighting. For this purpose, synthetic images comprised a useful intermediate step in the algorithm's learning.

Such uses build upon many of the comparative advantages of synthetic data noted above, including lower costs and increased speed of data acquisition and preparation; facilitating the creation of more representative datasets that include rare events and edge-case scenarios, thereby reducing bias in predictive models; and allowing for automatic and almost costless high-quality labeling without the need for manual annotation.[114] In addition, they allow seamless experimentation with different situations.[115] For example, the performance of computer vision models is dependent on the quality of the camera used to collect the training data and how well it matches the camera to be used in the final product (e.g., the quality of the lens).[116] Synthetic data that accommodates a wide variety of interchangeable cameras limits the risks that might otherwise arise every time a product's cameras are modified. It is

108. EL EMAM, *supra* note 33, at 34. Such services are being used by companies such as Amazon Robotics and PepsiCo. Andrews, *supra* note 81.

109. EL EMAM, *supra* note 33, at 34.

110. *Id.*

111. Janet Slifka, *Tools for Generating Synthetic Data Helped Bootstrap Alexa's New-Language Releases*, AMAZON SCI. (Oct. 11, 2019), https://www.amazon.science/blog/tools-for-generating-synthetic-data-helped-bootstrap-alexas-new-language-releases [https://perma.cc/FDY9-ALRG].

112. Javier Tordable, *Synthetic Data Creates Real Results*, FORBES (Aug. 26, 2020, 1:10 PM), https://www.forbes.com/sites/googlecloud/2020/08/26/synthetic-data-creates-real-results/? [https://perma.cc/VE4F-7X6U].

113. *Id.*

114. *See id.*

115. *See id.*

116. Shtylenko, *supra* note 15.

thus not surprising that synthetic data has already become the main data resource to train machine learning systems for "experimentation and prototyping of product[s] and [of] algorithm[ic] hypotheses."[117]

Such training may also enable *faster learning*. To illustrate, the U.S. Food and Drug Administration is collaborating with researchers to explore the use of virtual patients "in [medical] drug and device developments."[118] Such patients can be duplicates of real medical profiles ("virtual twins") which are used to test different, mock conditions in simulations.[119] Alternatively, they can be completely virtual, "computer-generated [patients that] represent the range of human variables" used to replace or augment real patients. [120] Such methods can potentially reduce human testing and shorten testing times.[121] As one of the leading researchers in this area argues:

> We should not limit ourselves by how the real world limits us. We can't create a person that represents more than that person, but we can create a model that represents more than one person. Why not take advantage of that? . . . Once you understand the diversity [of patients], you can build that into the [future, virtual] patient population.[122]

Such tools are especially important "where delays and costs can impede patient access to novel treatments."[123] Of course, they may only be used where appropriate levels of safety and efficacy are ensured. This discussion also illustrates how synthetic data simulations can add value when it is *too risky to test different scenarios in the real world.*

Synthetic data can also be used to *ensure that learning is not focused on irrelevant, immoral, or illegal parameters.* A well-known example involves an algorithm that was trained to distinguish between images of husky dogs and wild wolves.[124] While the algorithm succeeded most of the time, the separating principle it adopted was insufficiently representative: it focused on the background, having learned that a white background (snow) signifies a wolf.[125] To teach the algorithm not to focus only on the background, the

117. *Id.*

118. Allison Proffitt, *The Role of Virtual Twins in Clinical Trials*, CLINICAL RSCH. NEWS (July 13, 2020), https://www.clinicalresearchnewsonline.com/news/2020/07/13/the-role-of-virtual-twins-in-clinical-trials [https://perma.cc/L8KJ-337B] (describing an interview with "Steve Levine, the senior director of virtual human modeling at Dassault Systèmes").

119. *Id.*

120. *Id.*

121. *Id.*

122. *Id.*

123. *Id.*

124. Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, 2016 PROC. 22ND ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING 1135, 1142.

125. *Id.*

coders applied a method called data perturbation, mixing and matching parts from different datasets.[126] This process generated synthetic data in which pictures of wolves appeared against a range of backgrounds.[127] The algorithm then learned to distinguish the two animals based on more relevant parameters.[128] In another example, synthetic data was used to create a melanoma detection model that works on dark skin, where collected data repositories mostly contained images of lighter skin.[129] Such benefits are not limited to imagery. One example is the introduction of artificial sentences to address bias "in toxic sentence classification systems."[130]

For the same reason, synthetic data can also assist analysts in *solving class imbalance problems* which arise when two groups of users are widely imbalanced, making it difficult to train algorithms on collected data.[131] Amazon's resume rating algorithm described above, which learned from its training dataset to discriminate against female applicants, offers a case in point.[132] Another example involves JPMorgan's use of synthetic data to create an algorithm to detect money laundering. Given the scarcity of collected relevant data points, they used an automated simulator, based on examples of known behavior, to generate synthetic data points that provided a richer representation of the information a financial institution can observe when money laundering takes place.[133] The coders then inserted the simulated data into a real dataset, according to predetermined probability distributions, and trained the algorithm to detect these instances.[134] Similar methods can also serve to *artificially overcome bias*: where a dataset correctly replicates existing societal prejudices and produces illegal discriminatory decisions that reduce social welfare,[135] synthetic data can be used to reflect norms of equality rather than real-world inequalities.[136]

---

126. *Id.* at 1137, 1142.

127. *Id.* at 1142.

128. *Id.*

129. Timo Kohlberger & Yuan Liu, *Generating Diverse Synthetic Medical Image Data for Training Machine Learning Models*, GOOGLE RSCH.: BLOG (Feb. 19, 2020), https://ai.googleblog.com/2020/02/generating-diverse-synthetic-medical.html [https://perma.cc/4TZT-8XBV].

130. AGATHE BALAYN & SEDA GÜRSES, BEYOND DEBIASING: REGULATING AI AND ITS INEQUALITIES 45, 46 n.62 (2021).

131. *See* Charitos Charitou, Simo Dragicevic & Artur d'Avila Garcez, Synthetic Data Generation for Fraud Detection Using GANs 1 (Sept. 26, 2021) (unpublished manuscript) (on file with the *Iowa Law Review*).

132. Dastin, *supra* note 67.

133. *See* Samuel A. Assefa et al., *Generating Synthetic Data in Finance: Opportunities, Challenges and Pitfalls*, PROC. FIRST ACM INT'L CONF. ON AI FIN., Oct. 2020, at 1, 3–4.

134. *Id.*

135. Barocas & Selbst, *supra* note 2, at 729–32.

136. *See, e.g.*, David Leslie, Anjali Mazumder, Aidan Peppin, Maria K. Wolters & Alexa Hagerty, *Does "AI" Stand for Augmenting Inequality in the Era of COVID-19 Healthcare?*, 372 B.M.J., no. 304, Mar. 16, 2021, at 1, 1–4.

A final potential use of synthetic data is in *overcoming business constraints (such as trade secrets or data security) or legal ones (such as privacy regulation) on inter and intra-firm data transfers.*[137] That is, synthetic data can potentially serve as a "smokescreen" for sensitive variables or for variables that are key identifiers, while preserving the (main) statistical attributes of the collected data.[138] This type of protection is used, for example, by the U.S. Federal Reserve and the U.S. Census Bureau in some of their published datasets.[139] It is also widely used in healthcare, especially where the focus is on populations of patients, rather than individual patient records.[140] In addition, synthetic data can help firms meet updated privacy and security requirements while still making use of learnings from personal data which they were permitted to use in the past.

Of course, synthetic data also creates harms and risks. Most importantly, synthetic data models can be *inaccurate.*[141] This can result from incorrect or misrepresentative input data or background information. In particular, the coder can make erroneous assumptions about the distributions and correlations among the variables involved, which might arise when the process is new or the coder lacks sufficient experience with that system.[142] This problem can be partially addressed by testing synthetic data against real-world outcomes, to determine and increase its accuracy. Perhaps more difficult to address is inaccuracy that results from data curation. Consider deidentification of collected data. Because privacy considerations imply that synthetic data cannot necessarily capture all the statistical relationships in the original data, such data manipulations can come at a cost: they create an inherent trade-off between data protection and data utility (measured by the accuracy of the synthetic data as compared to collected data). However, although the field is still in its infancy, current methods for generating synthetic data already show

137. Steven M. Bellovin, Preetam K. Dutta & Nathan Reitinger, *Privacy and Synthetic Datasets*, 22 STAN. TECH. L. REV. 1, 7 (2019) ("[S]ynthetic data allows us to step away from the deidentification–reidentification arms race and focus on what really matters: useful data.").

138. J.P. Reiter, *Using CART to Generate Partially Synthetic Public Use Microdata*, 21 J. OFF. STAT. 441, 441–42, 450 (2005); Bellovin et al., *supra* note 137, at 2–4, 15.

139. Reiter, *supra* note 138, at 442.

140. *See, e.g.*, *Synthetic Data*, NAT'L INST. FOR HEALTH & CARE RSCH. (July 27, 2023), https://www.cprd.com/content/synthetic-data [https://perma.cc/H3TL-M92S] (examples include a "cardiovascular disease synthetic dataset" and a "COVID-19 symptoms and risk factors synthetic dataset").

141. Theresa Stadler, Bristena Oprisanu & Carmela Troncoso, Synthetic Data – Anonymisation Groundhog Day 1, 15 (Jan. 24, 2022) (unpublished manuscript) (on file with the *Iowa Law Review*).

142. This, however, does not necessarily imply that the synthetic data might not still be useful. A synthetic dataset based on inaccurate correlations can be used, for example, for debugging a data analysis program, or for some types of performance testing of software applications. *See, e.g.*, *How to Use Synthetic Data to Maximize Test Coverage*, GENROCKET (Aug. 12, 2020), https://www.genrocket.com/blog/how-to-use-synthetic-data-to-maximize-test-coverage [https://perma.cc/VK6D-6HTZ]; *Datasets_for_Debugging: Synthetic Datasets, for Demo's and Debugging 'abrem,'* RDRR (May 2, 2019, 4:49 PM), https://rdrr.io/rforge/abrem/man/datasets_for_debugging.html [https://perma.cc/WX2F-CQZS].

promise in achieving high accuracy levels while lowering the risks of reverse engineering that would expose protected variables.[143]

In some situations, adding synthetic data increases the *risk of duplicating bias or errors*. Consider again Amazon's recruitment algorithm, which inadvertently perpetuated gender bias.[144] Augmenting such a dataset with synthetic data could address this sort of bias by adding synthetic counterexamples (synthetic resumes of successful women). However, this sort of social engineering requires awareness and know-how. It is easy to imagine that adding synthetic data could perpetuate other kinds of biases of which coders are less aware.

Another potential concern with respect to synthetic data is that it may lead to complacency about the *risks of exposure* of the collected data used in its generation. While deidentification offers some protection, the continuing development of efficient algorithms, as well as synthetic datasets based on collected data, raises the prospect that such datasets could be analyzed together, leading to reidentification of data that would be impermeable to separate analysis. This is particularly true given advances in quantum computing, which can more easily break encryption methods.[145]

A final concern is that synthetic data could *change the power relationships* between players. By reducing barriers to data access, it can increase competition among suppliers/providers and facilitate data-based innovation. By potentially increasing the accuracy of the information held by suppliers/providers about consumers/users, it opens up more opportunities for beneficial use, but also for exploitation, manipulation, and abuse. This is true even if the supplier/provider does not possess significant market power as long as he possesses relative bargaining power toward (some) consumers/users. The next Part focuses on how synthetic data might affect power dynamics among suppliers/providers, while the following two focus on the effects on consumers/users.

II. EFFECTS ON COMPETITIVE DYNAMICS AND MARKET POWER

A. DATA-BASED MARKET POWER VIS-À-VIS COMPETITORS

Data-based advantages play an important role in the competitive dynamics of digital markets. This is because data is often the raw material for generation of information and knowledge, which enable better-informed

143. VANESSA AYALA-RIVERA, PATRICK MCDONAGH, THOMAS CERQUEUS & LIAM MURPHY, UNIV. COLL. DUBLIN, SYNTHETIC DATA GENERATION USING BENERATOR TOOL 1–2, 8 (Nov. 6, 2018) (on file with the *Iowa Law Review*); *see also infra* Section III.A (discussing the balance between data quality and privacy protection).

144. Dastin, *supra* note 67.

145. *See, e.g.*, Tammy Xu, *What Are Quantum-Resistant Algorithms—and Why Do We Need Them?*, MIT TECH. REV. (Sept. 14, 2022), https://www.technologyreview.com/2022/09/14/1059400/explainer-quantum-resistant-algorithms [https://perma.cc/T2SF-YSC5].

decisions.[146] The growth in machine learning applications, "especially data-hungry deep" learning techniques, is "pushing the boundaries of what is economically feasible and physically possible."[147] In such a setting, data-based advantages may not only strengthen the market power of some firms but also make it more durable.[148] As the Organization for Economic Co-operation and Development ("OECD") observed:

> [D]ata can give rise to self-perpetuating feedback loops, network effects and economies of scale that enhance the first-mover advantage of incumbent firms. Further, data access can be leveraged across multiple markets. . . . [E]vidence suggests that market power may be on the rise, and that it may be becoming more durable, particularly in digital-intensive sectors.[149]

Accordingly, competition in data and data-based markets is shaped by the height of access barriers to data.[150] When such barriers are high, potential entrants might not be able to challenge incumbents who enjoy data-based advantages because they cannot provide users with the utility that stems from better datasets.[151] As a result, some data-based markets are characterized by limited contestability.[152]

Such durable market power enables incumbents to enjoy high profit margins, which may lead to loss of allocative efficiency. But more importantly, productive and dynamic efficiency could be harmed because firms with potential cost or quality advantages might not be able to enter the market, and the incentives of incumbents to develop consumer-welfare-enhancing

146. Of course, this is not always the case. Take, for example, autonomous reinforcement learning models, such as the model used to train AlphaGo by DeepMind, which beat one of the top human players. AlphaGo needed almost zero training data, but much computational power. NIKOLENKO, *supra* note 20, at 9–11.

147. *Id.* at 1.

148. *See, e.g.*, STIGLER COMM. ON DIGIT. PLATFORMS, STIGLER CTR. FOR THE STUDY OF THE ECON. & THE STATE, *supra* note 19, at 40.

149. OECD, DATA PORTABILITY, INTEROPERABILITY AND DIGITAL PLATFORM COMPETITION 7 (2021), https://web-archive.oecd.org/2021-10-31/591383-data-portability-interoperability-and-digital-platform-competition-2021.pdf [https://perma.cc/RMP8-B2W6] [hereinafter OECD, DATA PORTABILITY] (citation omitted); *see also* Frederic Jenny, *Competition Law Enforcement and Regulation for Digital Ecosystems: Understanding the Issues, Facing the Challenges and Moving Forward*, CONCURRENCES, Sept. 2021, at 38, 44–56 (discussing the importance of data in some digital markets).

150. *See* OECD, DATA-DRIVEN INNOVATION: BIG DATA FOR GROWTH AND WELL-BEING 391–92 (2015) (describing how data now drives all aspects of innovation in the economy and society); *see also* MAURICE E. STUCKE & ALLEN P. GRUNES, BIG DATA AND COMPETITION POLICY 79 (2016) (describing how Facebook's and WhatsApp's privacy policies create difficulties for other firms to access user data).

151. *See* STIGLER COMM. ON DIGIT. PLATFORMS, STIGLER CTR. FOR THE STUDY OF THE ECON. & THE STATE, *supra* note 19, at 40.

152. *See id.* at 9, 34.

innovations could be suppressed.[153] These negative welfare effects are strengthened by the lack of any guarantee that products offered by incumbents are the best of their kind. Rather, first-mover advantages can lock a market into a suboptimal technological equilibrium.[154] Furthermore, in recent years "vertically integrated or conglomerate business models" have become more commonplace in the digital marketplace, leading to the formation of data-based ecosystems—further raising entry barriers.[155]

Synthetic data can help change such market dynamics. By introducing an alternative to some types of collected data or by lowering the amounts of collected data needed, it can potentially *reduce obstacles in any part of the data value chain.*[156] Furthermore, given that synthetic data can be used to augment collected datasets which are otherwise too small to be useful,[157] firms with small datasets could compete with firms that possess much more collected data. Furthermore, in some industries synthetic data could even reduce the benefits of indirect network effects[158] for suppliers and the resulting market structure of data-based ecosystems. Where, for example, synthetic data reduces the marginal benefit to a supplier from aggregating collected data from different sources (e.g., Facebook and Instagram), the comparative advantages firms can gain from such aggregations are reduced. This, in turn, could lead to more competition, more consumer choice (in both specific products and product bundles), and less concentrated market structures.

Synthetic data can also *change competitive dynamics via its effects on data sharing*. Where collected data no longer confers a significant comparative advantage on the collector, their willingness to share it is increased. The market price of such data will be capped by the costs of generating comparable synthetic data. The ability to share data will also rise. To illustrate, should synthetic data—based on collected private data—not fall within the scope of privacy laws,[159] it need not comply with legal requirements that relate to data privacy (e.g., obtaining the consent of data subjects, developing and applying internal systems to limit the exposure of private data, or storing

153. *Id.* at 8 ("[M]arket power may manifest itself through lower quality, lower privacy protection, . . . less variety of political viewpoints, and, importantly, less investments in innovation.").

154. *See* W. BRIAN ARTHUR, INCREASING RETURNS AND PATH DEPENDENCE IN THE ECONOMY 13–15 (Timur Kuran ed., 1994). *See generally* W. Brian Arthur, *Competing Technologies, Increasing Returns, and Lock-In by Historical Events*, 99 ECON. J. 116 (1989) (explaining how inferior technology affects market returns).

155. OECD, DATA PORTABILITY, *supra* note 149, at 7.

156. For such obstacles, see Rubinfeld & Gal, *supra* note 5, at 350–68.

157. *See* NIKOLENKO, *supra* note 20, at 12.

158. STIGLER COMM. ON DIGIT. PLATFORMS, STIGLER CTR. FOR THE STUDY OF THE ECON. & THE STATE, *supra* note 19, at 38. "[N]etwork effects are . . . mediated by a 'complement' to the network," such as when a large number of sellers on a platform increases the number of consumers interested in using this platform. *Id.*

159. *See infra* Section III.B.

private data only in the jurisdiction where data subjects reside).[160] This, in turn, makes it easier to share such data, both internally and externally.[161] While synthetic data might dampen firms' motivation to collect collected data, thereby obviously reducing the ability to share it, this dynamic may reduce wasteful collection of unneeded private data. Similar dynamics to those analyzed in this paragraph also apply to the sharing of synthetic data once created.

Synthetic data may also potentially *reduce the collective action problem in data markets*, especially where it can be used to supplement and expand existing datasets. As Heller recognized in his seminal article *The Tragedy of the Anticommons*, a breakdown in coordination may prevent the emergence of a commons even when general access to resources or infrastructure would be a social good.[162] For example, Heller and Eisenberg showed that in the biomedical context, a patent thicket over otherwise synergetic pieces of information can stifle the creation of life-saving innovations that build on such information.[163] This dynamic is, of course, not limited to medical information. Rather, data collection dynamics involving transaction costs (such as harms to privacy) and strategic behavior (such as free riding) can affect the incentives of data subjects to provide their data, even when they know it will be put to good use which might benefit themselves as well as others.[164] Consider, for example, a situation where each data subject assumes that the data of others will have the same effect on the production of a good from which she can eventually benefit. If parting with such data involves even a small potential loss of privacy, she might not provide it. If many data subjects act in the same fashion, the collective good will not be created, reducing social welfare. By potentially lowering the number of collected data entries needed to make an informed decision, synthetic data can indirectly overcome this collective action problem.

Accordingly, the increased possibilities for internal data generation and internal and external data sharing may facilitate cumulative and synergetic knowledge production, which may stimulate competition and *generate new and better products or services.*[165] Most fundamentally, access to data can shape both

160. Commission Regulation 2016/679 of Apr. 27, 2016, General Data Protection Regulation, 2016 O.J. (L 119) 1, 3, 6, 15 (EU) [hereinafter GDPR]; Rubinfeld & Gal, *supra* note 5, at 364.

161. HARVARD BUS. REV. ANALYTIC SERVS., *supra* note 32, at 3, 6.

162. Michael A. Heller, *The Tragedy of the Anticommons: Property in the Transition from Marx to Markets*, 111 HARV. L. REV. 621, 676–678 (1998).

163. Michael A. Heller & Rebecca Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCI. 698, 698–701 (1998).

164. Michael J. Madison, Brett M. Frischmann, Madelyn R. Sanfilippo & Katherine J. Strandburg, *Too Much of a Good Thing? A Governing Knowledge Commons Review of Abundance in Context*, FRONTIERS IN RSCH. METRICS & ANALYTICS, July 13, 2022, at 1, 4–6 (2022).

165. This was recognized by the European Commission. *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of*

the level and direction of innovative activity[166] by increasing the diversity (of competitors and inputs) that is essential to creativity.[167] Furthermore, by moving comparative advantages away from data collection, synthetic data may encourage greater investment in data analysis and innovative applications to improve knowledge. Importantly, the positive effects of better knowledge may extend beyond the market for which the data are immediately relevant due to transfer learning. The potential benefits of widening the use of most types of data are numerous.

Overcoming data-based comparative advantages is important not only for strengthening competition within a jurisdiction, but also for *overcoming comparative advantages of other jurisdictions*. To illustrate, consider the comparative advantage the Chinese government created for Alibaba by enabling it to test its algorithms for smart cities in several locations.[168] It would be very difficult for any other firm to accumulate so much collected data, especially given privacy concerns. If, however, this can be partly overcome by a synthetic dataset, then synthetic data can increase competition.

Of course, synthetic data does not necessarily reduce all data access barriers. Access to collected data is still needed where the creation or utility of synthetic data is based on collected data or where access to collected data can make the production of synthetic data cheaper. For example, Meta recently bought the image rights to all cricket matches in India, giving it a large image bank on which to train its algorithm to create synthetic images of cricket players for the metaverse.[169] A competing firm might buy a similar image bank for another sport. Yet where access to the necessary collected data is characterized by high barriers, control of such data could strengthen a firm's market power for two main reasons. First, its potential use to create more and better synthetic data increases the comparative advantages it offers. In that sense, less is more. Second, by reducing obstacles to internal data flows, including the costs of adopting data governance systems that comply with legal data privacy and security requirements, synthetic data enables

the Regions: A Digital Single Market Strategy for Europe, at 14–15, COM (2015) 192 final (May 6, 2015); *see also* HARVARD BUS. REV. ANALYTIC SERVS., *supra* note 32, at 6 (synthetic data enables data sharing that spurs innovation).

166. *See, e.g.*, Jeffrey L. Furman & Scott Stern, *Climbing Atop the Shoulders of Giants: The Impact of Institutions on Cumulative Research*, 101 AM. ECON. REV. 1933, 1936 (2011); Heidi L. Williams, *Intellectual Property Rights and Innovation: Evidence from the Human Genome*, 121 J. POL. ECON. 1, 1–2 (2013).

167. Wolfgang Kerber & Simonetta Vezzoso, Dow/Dupont: Another Step Towards a Proper Assessment Concept of Innovation Effects of Mergers 21 (June 24, 2019) (unpublished manuscript), https://ssrn.com/abstract=3856885 [https://perma.cc/2ZL7-DN39].

168. Tamar Giladi Shtub & Michal S. Gal, *The Competitive Effects of China's Legal Data Regime*, 18 J. COMPETITION L. & ECON. 936, 954–55 (2022).

169. Manish Singh, *Facebook Secures Exclusive Digital Rights for ICC Cricket Events*, TECHCRUNCH (Sept. 26, 2019, 3:19 AM), https://techcrunch.com/2019/09/26/facebook-secures-exclusive-digital-rights-to-stream-icc-global-events-in-indian-sub-continent/? [https://perma.cc/5238-3CTW].

(better) enjoyment of internal data-based economies of scale and scope by those who already possess large quantities of collected data. To illustrate, consider the case of JPMorgan, where data anonymization through the generation of a synthetic dataset, based on private data, enabled it to share data on risk analysis across its departments.[170] While such increased internal data flows raise a firm's utility from the data, they could also strengthen the market power of digital ecosystems, increasing both horizontal and vertical concentration, and heightening entry barriers.

Where the generation of synthetic data requires specific expertise, including methods based on protected intellectual property, this could create another entry barrier. However, projects like the Synthetic Data Vault, released in 2020 by MIT researchers, reduced such obstacles by creating external synthetic data generators.[171] The project uses synthesizers which allow users to upload their data and receive a new dataset with the same statistical properties as their original data, which can be made public without infringing privacy.[172] Firms may also outsource the creation of synthetic data, making it affordable at smaller scales. Finally, open-source communities, such as Open Synthetics, develop and offer the use of some synthetic data generators.[173]

B. IMPLICATIONS FOR ANTITRUST AND REGULATION OF PLATFORMS

Competitive dynamics form an important basis for the application of laws designed to ensure that, where possible, competition can take its course, so that consumers can enjoy what markets have to offer. Such laws include, inter alia, antitrust and platform regulations. They are based on assumptions regarding the operation of markets, which determine when regulation is justified. Given that synthetic data can potentially affect the competitive dynamics in data-based markets, it may change the applicability of such laws. Here we explore several examples.

Let us first focus on antitrust, which aims to prevent the erection of artificial barriers to competition. *Market power* is a foundational prerequisite for most antitrust prohibitions, including monopolization and merger regulation. This is because absent market power, competitive concerns are

170. *Synthetic Data*, *supra* note 37.

171. *See, e.g.*, Lab'y for Info. & Decision Sys., *The Real Promise of Synthetic Data: MIT Researchers Release the Synthetic Data Vault, a Set of Open-Source Tools Meant to Expand Data Access Without Compromising Privacy*, MIT NEWS (Oct. 16, 2020), https://news.mit.edu/2020/real-promise-synthetic-data-1016 [https://perma.cc/HL84-63A5].

172. *SDV: The Synthetic Data Vault*, DATACEBO (Mar. 28, 2023), https://sdv.dev/SDV [https://perma.cc/FQ38-DLTH].

173. *See, e.g.*, OPENSYNTHETICS, https://opensynthetics.com [https://perma.cc/U4YW-VE5P].

limited.[174] It is well established that collected data can create significant market power in some digital markets.[175]

Synthetic data may affect findings of market power. To illustrate, suppose a newcomer wishes to compete in the market for autonomous trains. To do so, they need numerous videos of images seen from the train. Such videos are generally the proprietary information of incumbents. But if they can generate such images synthetically, by using a small number of real images, this would significantly reduce entry barriers. Put differently, synthetic data may break down economies of scale and scope in the use of collected data in some markets.

These potential changes in data-based market power have implications for all parts of antitrust law. Below we focus on two types of effects: instances in which the possible generation of synthetic data affects the *need* for regulatory intervention in the marketplace, and those in which it affects *findings* of anticompetitive conduct. While such effects will generally be case-specific, in some industries they may be generalized.

Let us start with several examples of the first effect. Take the regulation of *mergers and acquisitions*. The digital era is characterized by high-profile megamergers that involve vast amounts of data relating to user conduct on digital platforms. These include, inter alia, Google/DoubleClick, Microsoft /Yahoo!, Microsoft/Skype, Microsoft/LinkedIn, and Facebook/Instagram.[176] Strong criticisms have been voiced against the clearance of such data-based mergers by antitrust authorities.[177] Enter synthetic data. Where data-based comparative advantages can be (partly) overcome, the negative welfare effects of changes in market structure will be lower. Accordingly, in industries where firms will be able to compete with smaller quantities of collected data and where the collection of such data does not involve insurmountable barriers, more mergers would be benign.

---

174. For an argument that challenges this foundational principle, see generally Noga Blickstein Shchory & Michal S. Gal, *Market Power Parasites: Abusing the Power of Digital Intermediaries to Harm Competition*, 35 HARV. J.L. & TECH. 73 (2021).

175. *See, e.g.*, Filippo Lancieri & Patricia Morita Sakowski, *Competition in Digital Markets: A Review of Expert Reports*, 26 STAN. J.L., BUS. & FIN. 65, 82–88 (2021). It is also suggested that data can give rise to forms of power not neatly captured by the concept of market power, see, for example, Orla Lynskey, *Grappling with "Data Power": Normative Nudges from Data Protection and Privacy*, 20 THEORETICAL INQUIRIES L. 189, 192–97 (2019); and more broadly, see generally JULIE E. COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM (2019) (arguing that the legal frameworks and structures that govern information and data in our contemporary digital society are not neutral but rather serve to consolidate power and shape the dynamics of informational capitalism).

176. Anca D. Chirita, *Data-Driven Mergers Under EU Competition Law*, *in* THE FUTURE OF COMMERCIAL LAW: WAYS FORWARD FOR HARMONISATION 147, 147 (2019); Jörg Hoffmann & Germán Oscar Johannsen, *EU-Merger Control & Big Data on Data-Specific Theories of Harm and Remedies* 12 (Max Planck Inst. for Innovation & Competition, Research Paper No. 19-05, 2019).

177. *See, e.g.*, Chirita, *supra* note 176, at 147 ("No merger should be unconditionally cleared if it involves a large amount of users' data.").

At the same time, as not all data could be created artificially, some data-based mergers should still be prohibited. This can be exemplified by the case brought against Facebook by the Federal Trade Commission in December 2020.[178] One of the main allegations concerned Facebook's acquisition of the social networking applications WhatsApp and Instagram: these acquisitions strengthened Facebook's market power, on the grounds that, inter alia, "[o]ver time, users of a personal social network build more connections and develop a history of posts and shared experiences, which they would lose by switching to another personal social networking provider."[179] Such data could not be artificially recreated, at least not easily.[180] Furthermore, when synthetic data strengthens comparative advantages, as in the case of JPMorgan elaborated above,[181] the need for merger review is increased.

Moreover, the application of antitrust should also be sensitive to the effects of internal transfers of deidentified data. Consider the following example, offered by Stavroulaki[182]: the Affordable Care Act prevents health insurers from discriminating against citizens based on their preexisting conditions.[183] One way to circumvent this prohibition is to offer less coverage in geographic areas where consumers are more prone to suffer from certain medical conditions. The data required for such a strategy can be based on patients' medical records. While such records are protected under privacy laws, their transfer or sale is allowed if the data is deidentified.[184] Such strategic use of data could result, it is argued, from "the recent merger between UnitedHealth Group, a giant health insurer, and Change Healthcare, the largest U.S. electronic data interchange . . . clearinghouse."[185] This is because the approved merger will give UnitedHealth access to deidentified health data on millions of patients, enabling it to cherry-pick the most

178. Complaint for Injunctive and Other Equitable Relief at 1–3, FTC v. Facebook, Inc., No. 20-cv-03590 (D.D.C. Dec. 9, 2020).

179. *Id.* at 19.

180. For an additional example, see the Ticketmaster/Live Nation merger that was conditionally approved in January 2010 (data on ticketing data—including data on number of tickets sold, proceeds from those sales, ticket inventory, pricing, marketing, and corresponding sales—and ticket buyer data—including nonpublic identifying information for ticket buyers—could not be artificially created). *See, e.g.*, Final Judgment, United States v. Ticketmaster Ent., Inc., No. 10-cv-00139 (D.D.C. July 30, 2010); Amended Final Judgment, United States v. Ticketmaster Ent., Inc., No. 1:10-cv-00139 (D.D.C. Jan. 28, 2020).

181. *Synthetic Data*, *supra* note 37.

182. Theodosia Stavroulaki, *How the Wrong Presumptions Led to the Wrong Conclusions in the United/Change Healthcare Merger*, PROMARKET (Nov. 11, 2022), https://www.promarket.org/2022/11/11/how-the-wrong-presumptions-led-to-the-wrong-conclusions-in-the-united-change-healthcare-merger [https://perma.cc/LD7D-9V2K].

183. Patient Protection and Affordable Care Act, Pub. L. No. 111-148, § 1201, 124 Stat. 119, 154–55 (2010).

184. Health Insurance Portability and Accountability Act of 1996, 42 U.S.C. § 1320d (1996).

185. Stavroulaki, *supra* note 182.

profitable geographic areas.[186] This, in turn, "would give United[Health] a . . . competitive advantage, especially if [its] rivals are deprived of access to a similar [type and] range . . . of data."[187]

Synthetic data can also *reduce the need and justification for data sharing arrangements* in order to realize data synergies. If such synergies can be realized via internal generation of synthetic data, then such arrangements, especially between competitors, have weaker justification. This observation has implications for many aspects of antitrust, as well as for ex-ante regulation of the digital economy which seeks to advance more efficient use of data.

Take, for example, *mandatory data access*. Given the importance of data for the efficient operation of many markets, several laws mandate data sharing, including the Essential Facilities Doctrine in antitrust.[188] The doctrine mandates sharing on fair and nondiscriminatory terms of facilities whose use is essential for competition in an adjacent market.[189] While the wings of the doctrine were clipped in the Supreme Court's *Trinko* decision,[190] the digital economy has revived calls for its application to data which is essential for competition. Graef, for example, suggests recognizing some types of data as "essential data."[191] She is not alone.[192] Such calls advance the claim that, despite data's nonrivalrous nature, some forms of data are indispensable, given that obstacles make it impossible or extraordinarily difficult to collect or to recreate it.[193] Synthetic data may challenge such essentiality in some

186. *Id.*

187. *Id.*

188. MCI Commc'ns Corp. v. Am. Tel. & Tel. Co., 708 F.2d 1081, 1132–33 (7th Cir. 1983) (Terms for application include: "(1) control of the essential facility by a monopolist; (2) a competitor's inability practically or reasonably to duplicate the essential facility; (3) the denial of the use of the facility to a competitor; and (4) the feasibility of providing the facility").

189. *See id.*

190. Verizon Commc'ns Inc. v. Law Offs. of Curtis V. Trinko, LLP, 540 U.S. 398, 409 (2004) (interpreting the duty to deal articulated in *Aspen Skiing* as confined to a setting in which "[t]he unilateral termination of a voluntary *(and thus presumably profitable)* course of dealing suggested a willingness to forsake short-term profits to achieve an anticompetitive end"). The doctrine has been criticized based on concerns about incentives for dynamic innovation, trust in the self-correcting mechanisms of markets, denials of the very existence of incentives to monopolize adjacent markets, and dire assessments of the ability of the courts and agencies to replace market mechanisms took center stage. *See generally* Zachary Abrahamson, Comment, *Essential Data*, 124 YALE L.J. 867 (2014) (describing criticisms of the Essential Facilities Doctrine).

191. *See generally* INGE GRAEF, EU COMPETITION LAW, DATA PROTECTION AND ONLINE PLATFORMS: DATA AS ESSENTIAL FACILITY (Alastair Sutton ed., 2016). In some rare cases an obligation to share data has been imposed through antitrust laws in order to advance competition. *See* Vikas Kathuria & Jure Globocnik, *Exclusionary Conduct in Data-Driven Markets: Limitations of Data Sharing Remedy*, 8 J. ANTITRUST ENF'T 511, 519–20 (2020).

192. *See* CRÉMER ET AL., *supra* note 1, at 98 ("[T]he 'classical' EFD may not be the right framework to handle refusal of access to data cases, as the doctrine has been developed with a view to access to 'classical' infrastructures and later expanded to essential IPRs.").

193. For criticism and highlights of the difficulties in applying the doctrine to data, see, for example, Giuseppe Colangelo & Mariateresa Maggiolino, *Big Data as Misleading Facilities*, 13 EUR. COMPETITION J. 249, 264–77 (2017).

markets: if collected data can be replaced by synthetic data, then the justifications for requiring firms to share it are weakened. This does not imply, of course, that no dataset will ever be deemed indispensable, but it may significantly reduce the number of such instances. The same logic applies to "mandatory data sharing as a remedy" for anticompetitive conduct.[194] Sharing will be necessary for reintroducing competition only if there are no alternative cost-effective ways of duplicating ill-gotten data-based advantages. At the same time, where sharing can assist in reintroducing competition, the ability to deidentify private data via synthetic data increases such a possibility.

Synthetic data also affects the need for *mandated data portability, interoperability, and standardization*. All three involve technical standards that affect the ability to learn from the data and are "often mentioned as . . . key [elements] of a digital competition policy reform agenda."[195] Data portability "seek[s] to reduce user switching costs," for example by mandating that the user be able to transfer his personal data to another provider.[196] Interoperability "focus[es] on allowing systems to communicate with one another," increasing "the ability of digital services to incorporate data . . . or functionality from [another data] provider."[197] Data standardization ensures that data is understandable by the receiver and is in useable format.[198] All three "measures have [already] been implemented through [antitrust] enforcement, . . . sector-specific regulation and other broad-based regulation[s]" around the world.[199] For example, data portability in banking was incorporated through the Dodd–Frank Act,[200] and "[s]everal new interoperability measures have been proposed . . . [for] the digital sector."[201] Synthetic data reduces the need for such policy tools in some markets. This is important given that such measures carry their own costs.[202] A recent OECD study emphasized that developing "legal, technical and procedural aspects of these measures may be particularly complex, as will monitoring."[203] Such policy

194. *See generally* Kathuria & Globocnik, *supra* note 191 (describing "the viability of mandatory data sharing" as a solution for affected markets).

195. OECD, DATA PORTABILITY, *supra* note 149, at 8, 20, 24 ("For providers of complementary services, linkages with a central ecosystem platform may also be its primary means of attracting users, and a way of leveraging existing functionality, such as account authentication and sign-in functions. . . . [I]f an incumbent makes small changes to [application programming interfaces (APIs)] or layers on additional procedures, it could have a fatal effect on the business model of firms relying on the API.").

196. *Id.* at 8.

197. *Id.* at 8, 12.

198. *See* Michal S. Gal & Daniel L. Rubinfeld, *Data Standardization*, 94 N.Y.U. L. REV. 737, 749–50 (2019).

199. *See* OECD, DATA PORTABILITY, *supra* note 149, at 27.

200. 12 U.S.C. § 5533 (2018).

201. OECD, DATA PORTABILITY, *supra* note 149, at 24–26, 42.

202. *Id.* at 15.

203. *Id.* at 15, 28, 48.

tools might also inadvertently create negative effects. For example, data standardization can entrench inefficient collection standards.[204] Accordingly, synthetic data could reduce the need to use second-best solutions for expanding data use.

Let us now turn to the second type of effect, which *changes the content of antitrust prohibitions*. Synthetic data may change, for instance, optimal burdens of proof. Take, for example, a case where a monopolist limits the data portability or interoperability associated with its product. Determining whether such a change constitutes an illegal refusal to deal is based, inter alia, on the assessment of harm to competition.[205] Should synthetic data generally limit the need for interoperability, harm to competition will be reduced, and a higher burden of proof might be justified across all cases.

Another example relates to the substance of laws. To illustrate, consider the prohibition of *cartels*, which is based on the existence of an "[agreement] in restraint of trade."[206] Synthetic data on market conditions and rivals' actions can help train algorithms to reach a coordinated equilibrium without such an agreement,[207] and the algorithm can then be applied to real-world conditions. In other words, synthetic data can be used to circumvent the existing law, in a way that can only be addressed by reformulating "the content of the prohibition."[208]

Similar considerations to those explored in this section should affect a host of additional regulatory tools which are grounded in the assumption that data-based advantages are significant and cannot be easily overcome. Take, for example, the European Digital Markets Act,[209] one of the first regulations in the world geared specifically toward digital platforms, which also applies to U.S.-based firms in their dealings in the European Union ("EU").[210] Some of its prohibitions were designed to limit comparative advantages based on collected data. For example, it mandates the gatekeeper platform to provide competing providers of online search engines with access on fair, reasonable, and nondiscriminatory terms to ranking, query, click, and view data generated

204. *Id.* at 11, 14; Gal & Rubinfeld, *supra* note 198, at 762.

205. OECD, DATA PORTABILITY, *supra* note 149, at 15–16, 19–22, 30.

206. The Sherman Antitrust Act of 1890 § 1, 15 U.S.C. § 1.

207. Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò & Sergio Pastorello, *Artificial Intelligence, Algorithmic Pricing, and Collusion*, 110 AM. ECON. REV. 3267, 3267–72, 3295 (2020).

208. *See, e.g.*, ARIEL EZRACHI & MAURICE E. STUCKE, VIRTUAL COMPETITION: THE PROMISE AND PERILS OF THE ALGORITHM-DRIVEN ECONOMY 61–70 (2016); Michal S. Gal, *Algorithms as Illegal Agreements*, 34 BERKELEY TECH. L.J. 67, 84–88, 112 (2019).

209. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828, 2020/1828 (Digital Markets Act) 2022 O.J. (L 265) 1 [hereinafter Digital Markets Act].

210. *See* Kevin E. Davis & Florencia Marotta-Wurgler, *Contracting for Personal Data*, 94 N.Y.U. L. REV. 662, 667 (2019).

by searches on its engines, subject to anonymization of personal data.[211] This regulation, which makes use of synthetic data to enable data transfers, is based on the assumption that such data creates significant data-based comparative advantages that could entrench the market power of the platform. A similar assumption stands at the basis of some recent U.S. legislative proposals. To illustrate, the Augmenting Compatibility and Competition by Enabling Service Switching (ACCESS) Act would mandate dominant platforms to maintain certain standards of data portability and interoperability.[212] It might be better, however, to limit such regulations to those types of data in which data-based advantages cannot easily be overcome once the option of synthetic data is introduced.

Many of the examples explored above suggest that the introduction of synthetic data will lead to a less interventionist approach to data-based advantages that affect competition, especially those that result from the natural conditions of the market rather than anticompetitive conduct. At the same time, antitrust and platform regulation have an important role to play in ensuring that artificial barriers to the creation of synthetic data are not erected and that where synthetic data increases market power, such power is not abused.

Synthetic data alters not only the power relationships between competitors in data-driven markets, but also between those who use data for decision-making and those who are affected by such decision-making, whether individuals or groups. Indeed, synthetic data can *affect the bargaining power of those that possess data*, even if they do not possess market power in the antitrust sense. The next section reviews effects on privacy, while the following one focuses on the effects of increased data quality on users/consumers.

III. EFFECTS ON DATA PRIVACY

Privacy laws govern the handling of personal information. At their essence, they "constrain the power that human information confers."[213] As

211. Digital Markets Act, *supra* note 209, at 6, 11.

212. H.R. 3849, 117th Cong. (2021). Congress did not enact the bill in 2021. *Id.* On May 25, 2022, the bill was reintroduced and was again not enacted. *Lawmakers Reintroduce Bipartisan Legislation to Encourage Competition in Social Media*, MARK R. WARNER: U.S. SEN. FROM THE COMMONWEALTH OF VA. (May 25, 2022), https://www.warner.senate.gov/public/index.cfm/2022/5/lawmakers-reintroduce-bipartisan-legislation-to-encourage-competition-in-social-media [https://perma.cc/CJG3-ETF4]. On July 26, 2023, the bill was reintroduced again. *Warner, Colleagues Reintroduce Bipartisan Legislation to Encourage Competition in Social Media*, MARK R. WARNER: U.S. SEN. FROM THE COMMONWEALTH OF VA. (July 26, 2023), https://www.warner.senate.gov/public/index.cfm/2023/7/warner-colleagues-reintroduce-bipartisan-legislation-to-encourage-competition-in-social-media [https://perma.cc/75V9-W9MV].

213. NEIL RICHARDS, WHY PRIVACY MATTERS 39 (2022); *see, e.g.*, ORLA LYNSKEY, THE FOUNDATIONS OF EU DATA PROTECTION LAW 77–78 (2015) (discussing the use of regulation to correct for market failures); DANIELLE CITRON, THE FIGHT FOR PRIVACY PROTECTING DIGNITY, IDENTITY AND LOVE IN OUR DIGITAL AGE 105–30 (2022) (arguing that the recognition of privacy as a civil liberty would rectify some of the power asymmetries she identifies).

such, they constitute essential vehicles for the protection of human rights, such as autonomy and self-definition, and act as facilitators of social institutions, including democracy and trust.[214] They seek to strike a social-welfare-enhancing balance between various interests, including fundamental human rights and data utility, both of which may benefit data holders, individuals, and society at large.[215] In this section we identify the impact of synthetic data on this balance and examine the challenges of applying existing privacy laws to such data.

Technological change, in particular the increased possibility of reidentifying deidentified data, has already interrupted the calibration of interests promoted by privacy laws.[216] Synthetic data further disrupts this balance. Moreover, as we show, the lens of synthetic data illuminates many of the problems that exist with our current legal conceptions of privacy harms.

A. THE IMPACT OF SYNTHETIC DATA ON BALANCING OF INTERESTS IN PRIVACY LAWS

Common to all data privacy regimes, albeit to differing extents, are principles that promote the fairness of data processing.[217] The requirement of lawful processing, frequently reflected in consent requirements, is well-known.[218] Additional principles include, among others, requirements regarding data security, data minimization, and data quality.[219] Synthetic data can *potentially promote these latter principles*. Most importantly, by replacing collected data with artificially generated data, or by adding to a dataset synthetic data that screens the outlier data points while retaining the statistical properties of the data, synthetic data offers an additional layer of security to personal data.[220] Data minimization requires that only the minimum amount

214. GLOB. PRIV. ASSEMBLY POL'Y STRATEGY WORKGROUP THREE, PRIVACY AND DATA PROTECTION AS FUNDAMENTAL RIGHTS: A NARRATIVE 5 (2022).

215. *See generally* Madison et al., *supra* note 164 (discussing the benefits and potential harms of having access to an abundance of data).

216. *See generally* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010) (explaining how anonymizing data fails to protect privacy); Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239 (2013) (arguing that the rapid development of data accessibility requires lawmakers to revisit the fundamentals of privacy law and draft new legislation).

217. LEE A. BYGRAVE, DATA PRIVACY LAW: AN INTERNATIONAL PERSPECTIVE 146–47 (2014). In the United States, these are the so-called Fair Information Practice Principles ("FIPPS"). *See, e.g.*, Woodrow Hartzog, *The Inadequate, Invaluable Fair Information Practices*, 76 MD. L. REV. 952, 959 (2017).

218. *See* Hartzog, *supra* note 217, at 959.

219. *Id.* at 954–60.

220. We follow Maynard-Atem, who considers synthetic data to be "a subset of anonymisation created by an automated process such that it holds similar statistical patterns as the original dataset." Louise Maynard-Atem, *The Data Series—Solving the Data Privacy Problem Using Synthetic Data*, IMPACT, Autumn 2019, at 11, 11–12.

of personal data be processed for a specified purpose.[221] As synthetic data is generated on demand, often with a specific use envisaged, it reduces the incentives for data holders to gather excess data.[222] Finally, by augmenting datasets where collected data is unavailable, synthetic data can promote data accuracy.[223] In light of these advantages, data privacy regulators have recommended the use of synthetic data as a substitute for collected data in certain contexts. For example, the Norwegian Confederation of Sports unintentionally shared the personal data of 3.2 million Norwegians online when testing solutions for moving a database from a physical server to the cloud.[224] The Norwegian Data Protection Authority, a Norwegian data privacy regulator, emphasized that this could have been avoided had synthetic data been used to test the migration.[225]

Moreover, some computer scientists argue that synthetic data can, in some situations, *improve the privacy-utility tradeoff relative to other privacy-enhancing techniques.*[226] Utility requires retaining in the dataset the relationship between different features and the distribution of values for each feature.[227] An optimal balance would preserve privacy while also allowing the data to be analyzed for socially valuable ends. For instance, in the medical context, it should be possible to process patient information so as to protect privacy while still enabling providers to mine the data for insights into illnesses and new treatments. Many existing techniques achieve anonymization by introducing noise into a dataset, thus disturbing the relationship between attributes or the distribution of values for attributes, or by stripping the dataset of some meaningful data points.[228] Such techniques reduce the accuracy, and therefore the utility of data. Some computer scientists claim that synthetic

221. *See, e.g.*, Hartzog, *supra* note 217, at 957 (calling the principle the "Collection Limitation Principle").

222. Note, however, that this does not require, or imply, that the overall amounts of data produced will be lower.

223. *See infra* Part IV.

224. *Norwegian DPA: Norwegian Confederation of Sport Fined for Inadequate Testing*, EUR. DATA PROT. BD. (June 15, 2021), https://edpb.europa.eu/news/national-news/2021/norwegian-dpa-norwegian-confederation-sport-fined-inadequate-testing_en [https://perma.cc/H3NV-FSMW].

225. *Id.*

226. *See supra* text accompanying notes 220–24.

227. Features or attributes are the characteristics of an object studied (e.g., age; height), while variables are the values assigned to these attributes (e.g., young/old; 18–35).

228. *See, e.g.*, *Opinion 05/2014 on Anonymisation Techniques*, WORKING PARTY ON THE PROTECTION OF INDIVIDUALS WITH REGARD TO THE PROCESSING OF PERS. DATA, at 3 (Apr. 10, 2014), https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf [https://perma.cc/U7PM-B9F2] (discussing "noise addition, permutation, differential privacy, aggregation, k-anonymity, l-diversity and t-closeness"). "A common challenge for state-of-the-art methods such as differential privacy is mainly on how to balance the added noise with the utility of the data. More noise usually means less privacy risk, but also means less utility on the data." Zhenchen Wang, Puja Myles & Allan Tucker, *Generating and Evaluating Cross-Sectional Synthetic Electronic Healthcare Data: Preserving Data Utility and Patient Privacy*, 37 COMPUTATIONAL INTEL. 819, 825 (2020).

data may significantly recalibrate this inverse relationship, by "maintain[ing] the majority of the valuable information and statistical integrity of the original data but eliminat[ing] the risk of re-identification."[229] Accordingly, it can potentially act as a relatively accurate proxy for collected data, but with lower privacy risks. Notably, however, privacy gains might not be evenly distributed: as discussed below, synthetic data may offer better privacy protection for some records.[230] Some individuals might therefore be disadvantaged relative to others. Still, the result may be Pareto-optimal relative to the state without synthetic data, meaning that it is better for all as compared with the previous situation. As elaborated below, some synthetic data proponents make an even stronger claim: that by combining synthetic data with differential privacy (another privacy-preserving technique) the resulting dataset retains both privacy and utility.[231]

Even if this is the case, and we do not take a stance on this point, an interpretability challenge may ensue.[232] "[A] machine learning model is interpretable if you can inspect the actual model and understand why it got a particular answer for a given input, and how the answer would change when the input changes."[233] Put differently, interpretability involves the ability to understand the algorithm's functioning and its outputs, rather than providing an answer to why the specific model was chosen.[234] When differential privacy and synthetic data are combined, it is not possible to know which patterns of the original dataset are retained and which are lost.[235] This has led some to conclude that "it is neither possible to anticipate the minimum gain in privacy from synthetic data publishing nor its utility loss."[236]

---

229. Maynard-Atem, *supra* note 220, at 13; *see* Fida K. Dankar & Mahmoud Ibrahim, *Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation*, 11 APPLIED SCIS., Mar. 1, 2021, at 1, 2; Bellovin et al., *supra* note 137, at 4–5. For a different view, see, for example, Stadler et al., *supra* note 141, at 1, 15.

230. *See infra* Section III.B.

231. Stadler et al., *supra* note 141, at 8–9. For an overview of some claims with regard to privacy, see, for example, Dankar & Ibrahim, *supra* note 229, at 2.

232. Dara Hallinan & Frederik Zuiderveen Borgesius, *Opinions Can Be Incorrect (in Our Opinion)! On Data Protection Law's Accuracy Principle*, 10 INT'L DATA PRIV. L. 1, 4 n.18 (2020) ("[G]iven the lengthy use of the concepts of quality and accuracy in computer science, it is somewhat of a peculiarity that there has not been more cross-fertilization of ideas with the accuracy concept in data protection law.").

233. STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 729 (4th ed. 2022). Interpretability (and explainability) is not always defined similarly by all researchers. *See, e.g.*, Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NATURE MACH. INTEL. 206, 206 (2019) ("[A]n interpretable machine learning model is constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, causality, structural (generative) constraints, additivity or physical constraints that come from domain knowledge." (citations omitted)).

234. MOLNAR, *supra* note 104, at 19–24.

235. Stadler et al., *supra* note 141, at 2, 15.

236. *Id.* at 15.

Yet, as explained in Part II above, both parameters can be tested against the statistical properties of the collected data. We analyze the consequences of this reduction of interpretability further below. However, for now it suffices to note that synthetic data might introduce another cost into the equation.

Furthermore, where collected data is used to generate a synthetic dataset, there is always some risk that the model itself or the dataset might indirectly leak some of the original personal data.[237] The attacker would essentially be reverse engineering the collected data from the model. Most notably, synthetic data that attempts to retain all the statistical properties of the collected data also retains the risk of linking the data to a specific person in outlier situations. Writing in the context of medical data, Wang et al. warn that the reidentification of outliers in the dataset could be an issue in cases where the synthetic dataset is "very similar . . . in terms of aggregated characteristics to real-world data."[238] This problem is also recognized in the industry literature. As a Gartner executive observes, "[i]f you are creating data for a rural area and it's one person per [one hundred] miles, even though I can create a synthetic person, it doesn't hide anything."[239] Solving the outlier risk problem often requires a reduction in the utility of the data.[240] Note, though, that in some situations it is possible to add to the dataset synthetic data which conceals the association of the data to a specific person by enlarging the group of data subjects which exhibit the outlier features. Thus, while synthetic data can potentially significantly reduce the risks of reidentification, it is not always the game-changer for the privacy-utility tradeoff that some suggest. This also points to the dangers of treating synthetic data as distinct from other anonymization techniques.

More fundamentally, synthetic data could create privacy harms even when a direct link does not exist between the individual and the personal data protected. This direct link, which is required in most privacy laws,[241] is challenged in the context of horizontal data relations, where data about individual A (or individuals in group A) is used in relation to, or to learn about, individual B (or individuals in group B). While it has already been recognized that modern data analytics techniques put pressure on this relational dimension of privacy laws with regard to collected data,[242] as

237. *Id.* at 1. An empirical study conducted by Stadler et al. found that "synthetic data does not protect all records in the original data from linkage and attribute inference." *Id.*; *see also* Khaled El Emam, Lucy Mosquera & Jason Bass, *Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation*, J. MED. INTERNET RSCH., Nov. 2020, at 736, 737 (finding that fully synthetic data still presents identity disclosure risks).

238. Wang et al., *supra* note 228, at 821.

239. HARVARD BUS. REV. ANALYTIC SERVS., *supra* note 32, at 4–5.

240. Wang et al., *supra* note 228, at 826.

241. *See infra* Sections III.B–.C.

242. *See, e.g.*, Christopher Jon Sprigman & Stephan Tontrup, *Privacy Decisions Are Not Private: How the Notice and Choice Regime Induces Us to Ignore Collective Privacy Risks and What Regulation Should Do About It* 45–56 (N.Y.U. L. & Econ. Rsch. Paper Series, Working Paper No. 23-22, 2023).

elaborated below synthetic data places further strain on their capacity to tackle data externalities.

Synthetic data therefore accentuates existing challenges to the effectiveness of data privacy laws, calling into question their rationale and the assumptions in which they are grounded. With this in mind, we focus on two main questions. We ask, doctrinally, whether synthetic data is captured under privacy laws. Normatively, we ask whether existing laws are fit to protect against privacy harms, while not harming data flows to an unnecessary extent, in a world in which synthetic data is widely used.

B. APPLICATION OF PRIVACY LAWS TO SYNTHETIC DATA

Is synthetic data captured under privacy laws?[243] Anonymous data escapes the application of data privacy frameworks worldwide.[244] This has led to some categorical claims that synthetic data is anonymous data and thus is not captured under privacy laws. For instance, a report commissioned by Mostly AI, a synthetic data provider, and published by Harvard Business Review Analytics Services, contains the following: "With . . . emerging privacy regulations around the world making the sharing of personal information so complicated, if not impossible, synthetic data is vital to support collaboration. As it is fully anonymous, it is exempt from these rules."[245]

Such categorical claims should be rejected. As with other anonymization techniques, a determination of whether synthetic data infringes data privacy laws requires a context-specific assessment against existing legal standards.

---

243. *See generally* Bellovin et al., *supra* note 137 (discussing how synthetic data is similar to raw data in the eyes of the law); César Augusto Fontanillo López & Abdullah Elbi, *On Synthetic Data: A Brief Introduction for Data Protection Law Dummies*, EUR. L. BLOG (Sept. 22, 2022), https://europeanlawblog.eu/2022/09/22/on-synthetic-data-a-brief-introduction-for-data-protection-law-dummies [https://perma.cc/7Y8S-5A82] (discussing how once data becomes synthetic, it will circumvent European data protection law).

244. For instance, the GDPR defines "personal data [as] any information relating to an identified or identifiable natural person." GDPR, *supra* note 160, art. 4(1). Therefore, any information that has been deidentified to a sufficiently robust standard (where reidentification will not be possible using means reasonably likely to be used by any person), falls outside its scope. Most other international instruments define personal data in a similar way. *See, e.g.*, African Union Convention on Cyber Security and Personal Data Protection, art. 1, June 27, 2014, https://au.int/sites/default/files/treaties/29560-treaty-0048_-_african_union_convention_on_cyber_security_and_personal_data_protection_e.pdf [https://perma.cc/4CKQ-C3RC]; Org. for Econ. Coop. & Dev. [OECD], Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, at 3, Doc. 0188 (2013) [hereinafter, OECD, Privacy Guidelines], https://legalinstruments.oecd.org/public/doc/114/114.en.pdf [https://perma.cc/W6Q3-3P3Q] (requiring that this data "pose a risk to privacy and individual liberties").

245. HARVARD BUS. REV. ANALYTIC SERVS., *supra* note 32, at 3. The report does not relate to specific conditions but includes general statements. *See id.* Similarly, Nvidia-supported research suggests that as "synthetic data would not be considered identifiable personal data, privacy regulations would not apply, and obligations of additional consent to use the data for secondary purposes would not be required." EL EMAM, *supra* note 33, at 6.

Our analysis confirms that whether synthetic data is captured by privacy laws requires us to consider, first, whether it falls within the material scope of a privacy law to begin with, and second, whether it can be brought outside of the scope of this legal framework by being deidentified in a way that satisfies the law's requirements. Data processing is permissible unless specifically regulated.[246] Both the sectoral focus of many privacy laws and the way in which they define the term "personally identifiable information" ("PII") imply that even their application to collected data may be limited. As we illustrate, the capacity of these laws to capture synthetic data, which is one step further removed from the individual, is more doubtful. To illustrate this, we first provide a brief introduction to the key terms defining the scope of application of data privacy laws, most notably the legal definition of PII and what constitutes deidentified data. We then apply this analysis to two types of synthetic data. The first is a simple example where, in an attempt to anonymize the data, a dataset containing data collected directly from individuals is used to generate a synthetic dataset.[247] For our second example, an entirely synthetic dataset (not based on the direct processing of any collected data describing specific individuals) is used to make inferences about an individual. We query in both cases whether the resulting synthetic data falls within the scope of privacy laws.

The United States does not tie itself to the mast of a single privacy law.[248] Rather, privacy regulations comprise a tapestry of legal provisions at the federal and state levels. Such laws have traditionally been mainly concerned with the liberty of citizens vis-à-vis the state, with data processing operations by private entities being given comparatively wide latitude.[249] Moreover, the First Amendment constrains the development of some privacy laws.[250] Accordingly, personal data processing is largely permissible, subject to compliance with some limited sectoral legal frameworks[251] and, more

246. As Schwartz and Solove state, "[t]he general approach to information flow in the United States is a 'Schillerian' one . . . ('What is not forbidden is allowed')." Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814, 1868 (2011).

247. This is sometimes referred to as a linear approach. See, for example, Solow-Niederman who writes that "[t]he linear approach assumes that the individual who cedes control of their data is the same individual potentially affected by the information collection, processing, or disclosure." Alicia Solow-Niederman, *Information Privacy and the Inference Economy*, 117 NW. U. L. REV. 357, 404 (2022).

248. Several proposals are currently seriously discussed, most notably the American Data Privacy and Protection Act, H.R. 8152, 117th Cong. (2022).

249. James Q. Whitman, *The Two Western Cultures of Privacy: Dignity Versus Liberty*, 113 YALE L.J. 1151, 1160–64 (2004).

250. *See infra* note 313.

251. For examples, see *infra* Section III.B.

recently, state data privacy laws.[252] For the moment, "the United States is the great international outlier in [Western societies] for data privacy."[253]

The lack of a single privacy law in the United States means that, unlike other significant privacy frameworks like the EU's General Data Protection Regulation ("GDPR")[254] or the OECD Privacy Guidelines,[255] there is no unifying concept that defines its applicability. The closest equivalent is PII, a notion that defines the scope and boundaries of numerous federal and state privacy statutes.[256] Since these laws differ in how they define PII and equivalent terms,[257] we use a categorization suggested by Schwartz and Solove to structure our analysis. They document three (sometimes inconsistent, sometimes overlapping) ways in which the term PII is defined: tautological (PII is "information which identifies a person"), "nonpublic" data, and through a bright-line rule which enumerates types of protected data (the "specific-types" approach).[258]

Recent developments point to the emergence of a fourth approach[259] to PII.[260] The Californian Consumer Privacy Act of 2018 ("CCPA") (as amended in 2021) defines personal information in an open-ended way as information that is "capable of being associated with, or could reasonably be linked,

252. *See infra* Section III.B.

253. Anupam Chander & Paul Schwartz, *Privacy and/or Trade*, 90 U. CHI. L. REV. 49, 86 (2023). Interestingly, international firms are starting to voluntarily adopt stricter provisions, in line with the EU's GDPR, *supra* note 160. *See, e.g.*, Davis & Marotta-Wurgler, *supra* note 210, at 667–99.

254. *See generally* GDPR, *supra* note 160.

255. *See generally* OECD, Privacy Guidelines, *supra* note 244.

256. Such sectoral legislation applies in health care, financial services, certain educational contexts, and credit reporting, among others, and is typically grounded more in consumer protection than in fundamental rights concerns. Anupam Chander, Margot E. Kaminski & William McGeveran, *Catalyzing Privacy Law*, 105 MINN. L. REV. 1733, 1748 (2021). Chander, Kaminski, and McGeveran note that "[a]s a final backstop, general-purpose consumer protection regulators, such as the Federal Trade Commission (FTC) and state attorneys general, address a subset of cases falling outside any sectoral rules, again largely following a consumer protection model." *Id.*

257. Schwartz & Solove, *supra* note 246, at 1816, 1827. The emergence of PII as a front-line legal concept and its significance is documented. *Id.* at 1819–28.

258. *Id.* at 1828–35. Some laws may fall into more than one category.

259. Schwartz and Solove proposed the adoption of a similar approach (which they termed PII 2.0) which incorporates a category of identifiable information, which is "when there is some non-remote possibility of future identification" based on the data. *Id.* at 1877–78. This was further developed in Paul M. Schwartz & Daniel J. Solove, *Reconciling Personal Information in the United States and European Union*, 102 CALIF. L. REV. 877, 907–08 (2014).

260. Not all recently adopted laws represent a rupture with earlier approaches. Indeed, Virginia's Consumer Data Protection Act, VA. CODE ANN. § 59.1-575 (West 2023), and the Colorado Privacy Act, COLO. REV. STAT. ANN. § 6-1-1303 (West 2022), remain firmly in the "publicly available" approach camp. For a comparison of these laws and the CCPA, see Cathy Cosgrove & Sarah Rippy, *Comparison of Comprehensive Data Privacy Laws in Virginia, California and Colorado*, TECKEDIN (July 9, 2021), https://docs.teckedin.info/docs/comparison-of-comprehensive-data-privacy-laws-in-virginia-california-and-colorado [https:// perma.cc/57N3-SU9D].

directly or indirectly, with a particular consumer or household."[261] This definition is accompanied by a nonexhaustive list of examples of such personal information.[262] While the CCPA continues to exclude publicly available information from its scope,[263] it represents a break from the other approaches to PII "by using the real-world potential for identifiability as the touchstone."[264] The draft proposal of a federal privacy bill, the American Data Privacy and Protection Act ("ADPPA"), goes along the same path by applying to "information that identifies or is linked or reasonably linkable, alone or in combination with other information, to an individual or [their] device."[265] Both laws also incorporate inferences, at least to some extent, within their ambit, as discussed below.[266] We take the CCPA as our exemplar of this new generation of U.S. privacy laws.

Let us now apply these laws to our first case, where synthetic data is generated using collected data about an individual as an input, either in the data generator or as a comparator, to refine the quality of the resulting synthetic dataset. Will the synthetic data produced in such circumstances be captured by privacy laws? This depends on which of the four definitions of PII apply.

At first glance, the CCPA's definition of PII[267] is most likely to bring such synthetic data within its scope. At the same time, the CCPA excludes

---

261. California Consumer Privacy Act of 2018, CAL. CIV. CODE § 1798.140(v)(1) (West 2022).

262. *Id.*

263. Pursuant to the changes introduced by the California Privacy Rights Act of 2020:

> "Personal information" does not include publicly available information or lawfully obtained, truthful information that is a matter of public concern. . . . "Publicly available information" means: information that is lawfully made available from federal, state, or local government records, or information that a business has a reasonable basis to believe is lawfully made available to the general public by the consumer or from widely distributed media; or . . . by the consumer . . . or information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience.

Id. § 1798.140(v)(2).

264. Chander et al., *supra* note 256, at 1750.

265. American Data Privacy and Protection Act, H.R. 8152, 117th Cong. § 2(8)(A) (2022). The ADPPA defines "covered data" as "information that identifies or is linked or reasonably linkable, alone or in combination with other information, to an individual or a device that identifies or is linked or reasonably linkable to an individual, and may include derived data and unique persistent identifiers." *Id.*

266. The ADPPA excludes from the definition of "'covered data' . . . inferences made exclusively from multiple independent sources of publicly available information that do not reveal sensitive covered data with respect to an individual." *Id.* This suggests that sensitive inferences based on publicly available information as well as other inferences derived from nonpublicly available information remain within the scope of the law. Yet unlike the CCPA, *id.* § 1798.140(v)(1), it does not capture household data—rather than personal data—within its scope.

267. *Id.* § 1798.140(v)(1).

deidentified data from its application. Deidentified data was previously defined as information that cannot reasonably be connected to a particular consumer, provided that business has implemented technical and organizational safeguards to prevent reidentification.[268] In 2023, a new definition came into effect, defining such data as "information that cannot reasonably be used to infer information about, or otherwise be linked, to a particular consumer."[269] Accordingly, what is crucial is whether the synthetic data can reasonably be connected to a specific individual or household. The amendment clarifies that this connection exists when one can infer information about a certain consumer or household from the data.[270]

Let us delve deeper into the reasonableness requirement. As noted above, the technical literature suggests that where synthetic data is generated using collected data as an input or comparator, a risk remains that it can be linked back to an individual through inference[271] or by linking the synthetic data with other datasets.[272] Yet a question arises as to how much time and effort would be considered reasonable to invest in preventing reidentification. This might require considering factors such as who might reasonably be thought to engage in reidentification, for what purpose, and what assumptions should be made about their ability to engage in reidentification. As these factors suggest, what might be deemed a "reasonable" risk threshold might differ from one type of data or one size of dataset to another and might change over time. Accordingly, how low this threshold should be is a legal and policy question that will have to be determined by courts and legislators. It is noteworthy that in the EU, where a similar test has been in place since 1995,[273] there remains much confusion and disagreement regarding the acceptable level of risk and how this may be quantified.[274] Note that the level of utility from the dataset and who it benefits does not come into the equation. In addition to this reasonableness requirement, the CCPA now also requires organizational safeguards such as a public commitment not to reverse the

268. *Id.* § 1798.140(h) (amended 2023, current version available at CAL. CIV. CODE § 1798.140(m) (West 2023)).

269. *Id.* § 1798.140(m).

270. *Id.*

271. Inference is understood here as the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

272. *See infra* Section III.C.1.

273. Identifiability is assessed by reference to the means "reasonably likely" to be used by a controller or another third party to reidentify the individual. Regulation 2016/679, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, 2016 O.J. (L 119) 26.

274. *See, e.g.*, Michèle Finck & Frank Pallas, *They Who Must Not Be Identified—Distinguishing Personal from Non-Personal Data Under the GDPR*, 10 INT'L DATA PRIV. L. 11, 31 (2020).

deidentification process unless it is necessary to verify deidentification.[275] It is apparent therefore that synthetic datasets generated using an individual's collected data can still fall within the scope of the CCPA.

The application of the other statutory definitions of PII to our first type of synthetic data is even more questionable (assuming that the data was not sufficiently deidentified). Take, for example, the definition of "personal information" found in the Children's Online Privacy Protection Act ("COPPA").[276] COPPA defines "personal information" as "individually identifiable information about an individual collected online."[277] It provides a list of examples, including, among others, an individual's first and last name, and a home or other physical address.[278] The list of examples—all specific identifiers—suggests that personal information is unlikely to capture our first type of synthetic data.[279] However, the law also enables the Federal Trade Commission ("FTC") to adopt implementing provisions.[280] Furthermore, in 2019, the FTC launched a consultation suggesting that it was willing to consider what is deemed PII, specifically querying whether it should be revised to include information that is "inferred about, but not directly collected from children" and "other data that serve as proxies for personal information covered under this definition."[281] Changes along these lines would more readily bring synthetic data generated using collected data within the law's scope. Yet this consultation closed the same year it opened without any substantive amendments having been proposed, much less adopted.[282] Furthermore, even with an expanded definition of personal information, COPPA applies only to personal information that is collected online by websites and online services.[283]

Other laws that adopt this "specifics" approach are also unlikely to encompass synthetic data generated using collected data. The Health

275. California Consumer Privacy Act of 2018, CAL. CIV. CODE § 1798.100. The business must also specifically prohibit reidentification and prevent the inadvertent release of the deidentified dataset. *Id.*

276. 15 U.S.C. 91 §§ 6501–6506.

277. *Id.* § 6501(8).

278. *Id.*

279. Bellovin et al., *supra* note 137, at 48 ("[B]ecause privacy statutes do not speak to 'fake' data, a door is left open, for better or worse.").

280. 15 U.S.C. § 6502.

281. Request for Public Comment on the Federal Trade Commission's Implementation of the Children's Online Privacy Protection Rule, 84 Fed. Reg. 35842 (July 25, 2019) (seeking comments for a new COPPA rule, including a question about whether to include inferred information in the definition of personal information).

282. FTC Commissioner Bedoya suggested that the FTC's preference is for Congress to amend the law, whether through COPPA or by introducing a federal privacy law. Andrea Vittorio, *FTC's Bedoya Calls for Congress to Update Kids' Privacy Law*, BL (Sept. 20, 2022, 2:58 PM), https://news.bloomberglaw.com/tech-and-telecom-law/ftcs-bedoya-calls-for-congress-to-update-kids-privacy-law?context=article-related (on file with the *Iowa Law Review*).

283. 15 U.S.C. § 6501.

Insurance Portability and Accountability Act of 1996 ("HIPAA") adopts an inverse-specifics approach.[284] It applies a safe harbor for the sharing of protected health information as long as seventeen specified identifiers are removed from the dataset, including, inter alia, names, e-mail addresses, Social Security numbers, and medical record numbers.[285] Synthetic data is unlikely to contain these real identifiers if the entire dataset has been replaced by artificially generated data. Therefore, even if it might be possible to infer the identity of an outlier individual from the shared dataset, it would be considered sufficiently private for HIPAA. In this sense, the specifics approach could lead to underinclusive protection.[286] Given that this approach does not allow for a contextual assessment of privacy loss, the protection offered will not be sufficiently calibrated to the risks that our first form of synthetic data might pose.

The tautological approach to defining PII might also not apply to our first case of synthetic data. This can be exemplified by the Video Privacy Protection Act ("VPPA"), which prohibits "video tape service provider[s]" from "knowingly disclos[ing] PII to third parties (with certain exceptions)."[287] It defines PII as "includ[ing] information which identifies a person as having requested or obtained specific video materials or services from a video tape service provider."[288] This definition has been interpreted narrowly, to exclude individuals who are identifiable, rather than identified, from its scope. In the *Hulu Privacy Litigation*, the U.S. District Court for the Northern District of California assessed whether Hulu, a provider of online access to prerecorded content, infringed the VPPA when it provided the URL addresses of content viewed by Hulu users to Facebook.[289] Despite the obvious risk that Facebook could connect this information with information it already held about its users in order to identify them, the court held that there was no disclosure of PII as Hulu did not have actual knowledge that Facebook would identify individuals on the basis of the disclosed data.[290] In the absence of such actual knowledge, this interpretation enables synthetic data providers to assume that such data would not constitute PII. This assumption could be strengthened by, for

284. 42 U.S.C. § 1320d(6).

285. *Id.*

286. Bellovin et al., *supra* note 137, at 42–45 (arguing that it might also be overinclusive as the risk of linkage across datasets is reduced, as we would be cross-referencing collected data with synthetic data).

287. 18 U.S.C. § 2710(b)(1). These exceptions apply when the provider is compelled to provide the information by a warrant, when the provider only discloses names and addresses for the purpose of direct marketing, or when the provider makes the disclosure "incident to the ordinary course of business." *Id.* § 2710(b)(2).

288. *Id.* § 2710(a)(3). This includes "prerecorded video cassette tapes or similar audio visual materials" and therefore captures the services of content streaming services. *Id.* § 2710(a)(4).

289. *In re* Hulu Priv. Litig., 86 F. Supp. 3d 1090, 1091 (N.D. Cal. 2015).

290. *Id.* at 1105.

instance, introducing contractual restrictions on connecting the synthetic data with individuals in the collected data used to generate it.

The third category of privacy laws, such as the Gramm–Leach–Bliley Act ("GLBA"), requires financial institutions to respect the privacy of financial information pertaining to their clients,[291] and exclude publicly available information from the definition of PII (the "nonpublic" category).[292] The GLBA defines "publicly available information" as "any information . . . lawfully made available to the general public from . . . Federal, State, or local government records[,] [w]idely distributed media [(including the Internet),] or [d]isclosures to the general public that are required to be made by Federal, State, or local law."[293] Data about an individual scraped from publicly available websites and subsequently used by financial institutions would therefore not constitute PII.[294]

What emerges from the application of these various laws to our first category of synthetic data is that the newer generation of privacy laws, like the CCPA, are less rigid and more adaptive to technological change than earlier privacy laws. They achieve this, in part, through their broad and general reach, a facet of the law that is not without controversy or complication. In contrast, the more specific sectoral privacy laws apply to this form of synthetic data in much the same way as to collected data. They suffer from underinclusivity and failure to capture relevant privacy risks, as well as overinclusivity when they apply irrespective of such risk.

---

291. 15 U.S.C. § 6801(a).

292. Charles M. Horn, *Financial Services Privacy at the Start of the 21st Century: A Conceptual Perspective*, 5 N.C. BANKING INST. 89, 107 (2001).

293. 16 C.F.R. § 314.2(o)(1) (2022). The GLBA defines personally identifiable financial information as "information [a] consumer provides . . . to obtain a financial product or service from" a provider; information about them resulting from such a transaction, or information "otherwise obtain[ed] about a consumer in connection with [the provision of] a financial product or service to" them. *Id.* § 314.2(n)(1). "Information that does not identify a consumer," including "aggregate information or blind data that does not contain personal identifiers," is excluded. *Id.* § 314.2(n)(2)(ii)(B).

294. Contrast this with the California Consumer Protection Act, where the definition of publicly available information is narrower:

> "[P]ublicly available" means: information that is lawfully made available from federal, state, or local government records, or information that a business has a reasonable basis to believe is lawfully made available to the general public by the consumer or from widely distributed media; or information made available by a person to whom the consumer has disclosed the information if the consumer has not restricted the information to a specific audience. "Publicly available" does not mean biometric information collected by a business about a consumer without the consumer's knowledge.

California Consumer Privacy Act of 2018, CAL. CIV. CODE § 1798.140(v)(2) (West 2022). The application of CCPA to scraping is contested. Brian Stuenkel, *Personal Information and Artificial Intelligence: Website Scraping and the California Consumer Privacy Act*, 19 COLO. TECH. L.J. 429, 445–50 (2021).

Let us now apply the legal definitions to the second type of synthetic data: replicating collected data by using assumptions rather than the direct processing of collected data. Here the application of privacy laws is even more questionable, despite the fact that the same knowledge about an individual that is otherwise protected by privacy laws could still be derived. Of the laws discussed above, in their current format, only the CCPA and ADPPA potentially apply to such data.

To show this, it is useful to think of personal data as on a spectrum of proximity to an individual:

Figure 5: Spectrum of Proximity of Personal Data to an Individual

Closest to the individual is data that identifies them, followed by data from which they can be identified with additional effort. The next three categories focus on inferences, which we define, following Sandra Wachter and Brent Mittelstadt, as information relating to a "natural person created through deduction or reasoning rather than mere observation or collection from the data subject."[295] These categories relate to the *source* of the inferences. The first includes inference-based synthetic data derived from data pertaining to a specific individual (inferences about Ann based on data describing Ann); the second includes inference-based synthetic data derived from data pertaining to a third party (inferences about Ann based on data describing Barry). The last category relates to inferences about an individual derived from a synthetic dataset pertaining to a group. In the last category, it is useful to separate two cases: inferences about Ann based on an effectively deidentified dataset, which was previously based on collected data (pertaining to Ann or others) and inferences about Ann based on a dataset which was generated by a model based on assumptions about how individuals behave, which never related to any specific individual. In all cases, the inference is linked back to Ann based on some known fact about her.

The legal treatment of inferences has, until recently, received relatively little attention in privacy scholarship and doctrine.[296] The main question

295. Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI*, 2019 COLUM. BUS. L. REV. 494, 515. For an alternative definition of inference, see Hallinan & Borgesius, *supra* note 232, at 1 ("[P]ersonal data constituting an assertion about a data subject, built on the back of facts about that subject, subjected to some interpretative framework to produce new, probable facts about that data subject.").

296. The legal classification and governance of inferences is foregrounded in the work of Salomé Viljoen, *A Relational Theory of Data Governance*, 131 YALE L.J. 573 (2021), and Solow-

contested was whether an inference deduced about an individual from their own personal data constitutes PII.[297] To the extent that the law has been attentive to inferences to date, this has been primarily to consider whether inferences about a person deduced from their own personal data constitute personal data. Most famously, the CCPA gives consumers the right to know specific pieces of personal information that a business has collected about them. The Attorney General ("AG") in California was asked to opine on whether this right applied to internally generated inferences from either internal or external information sources held by the business about the consumer.[298] The AG determined that the CCPA applies to such inferences.[299] The definition of "personal information" in the Act includes a subdivision listing the types of information that constitute personal information, such as personal identifiers, consumer information, and online transactions.[300] It also includes "[i]nferences,[301] drawn from any of the information identified in this subdivision," used to create consumer profiles.[302] The AG reasoned that the source of personal information was irrelevant when responding to a request to know,[303] as long as it was of the kind listed in the Act.[304] Accordingly, even

---

Niederman, *supra* note 247. In the EU, the judgment of the Court of Justice of the European Union ("CJEU") in Case C-434/16, *Nowak v. Data Protection Commissioner*, ECLI:EU:C:2017:994, brought inferences within the material scope of EU data protection law, however the consequences of this classification remain disputed. *See, e.g.*, Wachter & Mittelstadt, *supra* note 295, at 499–500.

297. Viljoen, *supra* note 296, at 585.

298. California Consumer Privacy Act of 2018, 20-303 Op. Att'ys Gen. 1 (2022).

299. *Id.*

300. The definition of personal information specifically includes personal identifiers (such as name, date of birth, and SSN) in addition to information about education, employment, travel, health, credit, banking, Internet Protocol addresses, online transactions, online searches, biometric data, or geolocation data. It also includes "inferences drawn from any of the information identified in this subdivision to create a profile about a consumer reflecting the consumer's preferences, characteristics, psychological trends, predispositions, behavior, attitudes, intelligence, abilities, and aptitudes." CAL. CIV. CODE § 1798.140(v)(1)(K) (West 2022).

301. Inferences are defined as "the derivation of information, data, assumptions, or conclusions from facts, evidence, or another source of information or data." *Id.* § 1798.140(r).

302. A consumer profile is "a profile about a consumer reflecting the consumer's preferences, characteristics, psychological trends, predispositions, behavior, attitudes, intelligence, abilities, and aptitudes." *Id.* § 1798.140(v)(1)(K).

303. California Consumer Privacy Act of 2018, 20-303 Op. Att'ys Gen. 11 (2022). The opinion points to CCPA subd. (o), which "draws no distinction between public and private sources." *Id.*

> It follows that, for purposes of responding to a request to know, it does not matter whether the business gathered the information from the consumer, found the information in public repositories, bought the information from a broker, inferred the information through some proprietary process of the business's own invention, or any combination thereof.

Id.

304. The definition of inferences refers to inferences "drawn from 'information identified in this subdivision.'" *Id.* This definition suggests that the inference must be based on one of the

inferences based wholly or partly on public records would need to be disclosed, so long as they fall within the list and were used to create a profile of the consumer.[305] Based on this interpretation of the law, as the list of identifiers is merely an indicative list, it remains possible that inferences about Ann, derived from a broad range of information about Ann, constitute her personal information.

It is unclear whether a causal connection is required between the data on Ann and the inference about Ann. What if, for example, Ann's data has only a marginal effect, and the same inference could be drawn without Ann's data, by using Barry's data? [306] Can the "identified or identifiable" individual be *any* individual who is subject to a data-informed inference? The text of the CCPA leaves this possibility open. In particular, its definition of PII includes information that *relates to* "a particular consumer or household."[307] In the EU, the term "relates to" has been interpreted to mean that the content of the data provides some information about a person, or that it relates to them in terms of the purposes of its processing or its effects.[308] If a similar interpretation were applied under the CCPA, it would capture inferences about an individual derived from the data of others. Indeed, while the inference should be derived from the list of examples of personal information found in the CCPA, the term "inference" is defined broadly as "the derivation of information, data, assumptions, or conclusions from facts, evidence, or another source of information or data."[309] This suggests that inferences regarding a person generated from entirely artificial data (with no personal data used as a direct input) might still be captured under the law.[310]

The challenge with this interpretation is, of course, that it may cast the net of privacy laws too wide. Moreover, it could well raise First Amendment challenges. The First Amendment has been interpreted to protect the content

already specified examples of personal information. *See id.* One might, however, argue for a broader construction of this provision as it refers to the "subdivision" (subd. (o)) and the subdivision includes the definition of personal information. *See id.* Moreover, the list of personal information examples provided is not exhaustive. *Id.*

305. This interpretation of inferences "rules out situations where a business is using inferences for reasons other than predicting, targeting, or affecting consumer behavior." *Id.* at 12.

306. As Solow-Niederman explains, machine learning models can aggregate the data of individuals to identify patterns, which are subsequently used to make inferences about other individuals. Solow-Niederman, *supra* note 247, at 361–62.

307. CAL. CIV. CODE § 1798.140 (v)(1) (West 2022).

308. Case C-434/16, Peter Nowak v. Data Prot. Comm'r, ECLI:EU:C:2017:994, ¶¶ 34–35 (Dec. 20, 2017).

309. CAL. CIV. CODE § 1798.140(r).

310. *See* Case C-184/20, OT v. Vyriausioji tarnybinės etikos komisija, ECLI:EU:C:2022:601, ¶ 120 (Aug. 1, 2022) (holding that "data that are capable of revealing the sexual orientation of a natural person by means of an intellectual operation involving comparison or deduction" are in fact sensitive data protected by "Article 9(1) of the GDPR"). There, the publication of the name of a spouse or partner amounted to the processing of sensitive data because it could reveal sexual orientation, even if no such inference was indeed made. *Id.* ¶ 119.

of speech between parties, and an inference constitutes such speech.[311] In particular, the speech of private commercial actors—which might include companies making inferences—has historically been treated as protected speech.[312] The extent to which states can limit the free speech rights of digital platforms is currently the subject of disagreement between states.[313]

This analysis demonstrates why any categorical claim that data privacy laws do not apply to synthetic data must be rejected. At the same time, the applicability of rule-based concepts of PII largely fail to incorporate synthetic data processing. Principle-based approaches, like the CCPA and COPPA, offer much more scope for capturing synthetic data, but much depends on how they will be interpreted.

C. ARE DATA PROTECTION LAWS FIT FOR PURPOSE?

The analysis above raises fundamental questions about what we seek to protect through information privacy laws, and whether our methods are fit for purpose. In this section we explore three main normative challenges: the focus on categories of information, the limited ability to capture spillover effects from data on others, and collective data harms. While such challenges are not unique to synthetic data, the rise of synthetic data pushes these tensions and challenges into the spotlight and may exacerbate them. As a result, synthetic data affects the balance between privacy and data utility on which privacy laws are based.

Another way to frame this challenge is through the lens of Nissenbaum's "context-relative information norms," which prescribe or proscribe actions relating to the flow of information about an information subject from one actor to another.[314] Synthetic data highlights the importance of considering indirect data relations within data flows. In particular, it challenges what Nissenbaum calls "transmission principles," which are "constraint[s] on the flow . . . of information from [one] party to [another] in a [specific] context."[315] This is because common transmission principles (such as consent and anonymization) might be insufficient to protect privacy in the age of synthetic data. Indeed, if synthetic data can preserve the utility of a dataset while dispensing with the need for direct use of collected data to generate it, this

311. *See, e.g.*, Reno v. ACLU, 521 U.S. 844, 864 (1997).

312. *See, e.g.*, *id.* at 869–70; Packingham v. North Carolina, 582 U.S. 98, 108–09 (2017).

313. Both Florida and Texas sought to introduce laws which prohibited digital platforms from restricting content based on the viewpoint of the user or another person. The Florida law was deemed an unconstitutional interference with the free speech rights of private platforms by the Eleventh Circuit, while the Texas law was upheld by the Fifth Circuit Court of Appeals. Tom Jowitt, *Florida Seeks Supreme Court Ruling on Social Media Law*, SILICON (Sept. 23, 2022, 12:54 PM), https://www.silicon.co.uk/e-management/social-laws/florida-supreme-court-social-media-law-477250 [https://perma.cc/NN67-UEHE].

314. HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE 141–43 (2010) (emphasis omitted).

315. *Id.* at 145–47.

may serve to circumvent rather than promote the objectives of existing privacy rules.

1. Challenges Arising from Categorizing Data

Privacy laws often protect categories of information-sensitive data, nonpublic data, or specific types of data.[316] As Ohm has written, "[t]his approach assumes that lawmakers can evaluate the inherent riskiness of data categories, assessing with mathematical precision whether or not a particular data field contributes to the problem enough to be regulated."[317] Yet, as our discussion indicates, this approach fails to capture all types of data that might create privacy harms.

Furthermore, this approach gives insufficient weight to the fact that deidentified data—such as synthetic data generated either by using collected data directly or by using inferences based on real-world observations—might be linked back to individuals.[318] The unreliable boundary between deidentified and identified data is widely recognized in the legal and computer science literatures. For example, already in 2009, Acquisti and Gross showed how it was possible to predict an individual's Social Security number from only publicly available data about their place and date of birth.[319] As the accumulated knowledge mined for data grows and better data-mining techniques are created, this reverse-engineering problem will only increase.[320] Accordingly, the risk of reidentification will grow along with the availability of more collected data and of synthetic datasets based on such data. Cloud

316. For example, the ADPPA refers to "sensitive covered data." American Data Privacy and Protection Act, H.R. 8152, 117th Cong. § 2(8)(A) (2022).

317. Ohm, *supra* note 216, at 1734. Zarsky notes that "the rise of Big Data substantially undermines the logic and utility of applying a separate and expansive legal regime to 'special categories.'" Tal Z. Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 SETON HALL L. REV. 995, 1014 (2017). De Gregorio notes that "the rationale behind the distinction between 'ordinary' and 'particular' categories of data tends to be nullified by the way in which the data are processed." GIOVANNI DE GREGORIO, DIGITAL CONSTITUTIONALISM IN EUROPE: REFRAMING RIGHTS AND POWERS IN THE ALGORITHMIC SOCIETY 243 (2022).

318. *See* JOSEF DREXL, DATA ACCESS AND CONTROL IN THE ERA OF CONNECTED DEVICES 48 (2019). "Given the potentials of big data analytics, which allows to draw probability conclusions from correlations between different pieces of information, it is no longer possible to neatly distinguish between non-personal and personal data." *Id.*

319. Alessandro Acquisti & Ralph Gross, *Predicting Social Security Numbers from Public Data*, 106 PROC. OF THE NAT'L ACAD. OF SCIS. 10975, 10975 (2009) ("The inferences are made possible by the public availability of the Social Security Administration's Death Master File and the widespread accessibility of personal information from multiple sources, such as data brokers or profiles on social networking sites.").

320. As Schwartz and Solove state, "[t]he public or private status of data often does not match up to whether it can identify a person or not," and individuals may not want their publicly available data aggregated with nonpublicly available information. Schwartz & Solove, *supra* note 246, at 1830. *See generally* Helen Nissenbaum, *Protecting Privacy in an Information Age: The Problem of Privacy in Public*, 17 LAW & PHIL. 559 (1998) (discussing how new surveillance methods are gathering more personal data that is being willingly shared by individuals).

storage, which significantly reduced the costs of data storage, further magnifies this possibility.

In light of such limitations of a category-based approach to information privacy, one might query whether an approach that focuses on the final, synthetic dataset rather than the category of data it includes (such as an individual's social security number) might be superior. This proposition faces two main obstacles. From a practical perspective, this would require an overhaul of some existing statutory instruments, changing both the designated entity (such as videotape service providers in the VPPA or financial institutions in the GLBA) and the focus of the limitations on PII. Second, the risk of reidentifiability may vary with conditions that the generator of the dataset might not be privy to (for example, how much more collected data the user has or is likely to have). This risk could also change over time. In line with this challenge, the UK data protection regulator argued that it is almost impossible to predict the risk of reidentification through data linkage, as "it can never be assessed with certainty what data is already available or what data may be released in the future."[321] Accordingly, imposing such a requirement might significantly harm the ability to use any synthetic data relating to people, unless applied in a manner which is sensitive to the effects of such changes on incentives for beneficial data mining.

Another suggestion involves imposing statutory obligations on the downstream operators (such as Facebook in the *Hulu Privacy Litigation*) which can turn identifiable data into identified information. The fact that the probability of identifiability may change over time strengthens this proposition. Indeed, to limit overbroad chilling effects on the use of data, such limitations should potentially balance the probability of detection at the time the dataset was internally created or purchased, the costs and financial risks involved in creating it, the observable level of harm at the time of use, and the social utility from that can be gained by data flows via synthetic datasets.

2. Limited Ability to Capture Spillover Effects

The next two challenges raise more fundamental conceptual questions regarding the connection between the data governed by data privacy laws and the objectives of these laws. Most obviously, are we concerned with protecting the data of an individual as such (such as the health information of Ann), or are we concerned with the impact that data processing might have on power dynamics between the data holder and the individual? As our analysis demonstrates, data privacy laws tend to emphasize the personal nature of the information processed—focusing primarily on the *source* or *type* of data at stake—rather than on the nature of the harms that might flow from data processing. They are predicated on an implicit assumption that by protecting

321. INFO. COMM'R'S OFF., ANONYMISATION: MANAGING DATA PROTECTION RISK CODE OF PRACTICE 18 (2012).

PII, we are protecting the privacy rights of each individual. More importantly, for harm to arise, there must be a direct connection between the individual and the personal information protected, in the form of direct or indirect identifiers or inferences. Synthetic data challenges this assumption, requiring us to reassess the nature of the connection between the data processed and the individual. Synthetic data can increase such concerns in two main ways: through the ability to learn via inferences based on data of others (spillover effects) and by strengthening collective data harms.[322]

Synthetic data increases spillover data privacy harms, which are beginning to be recognized, and which further challenge existing assumptions regarding the source of data.[323] Such harms result from externalities in data analysis. For example, an individual's data (provided or observed) may enable the accumulation of data about others as well. Put differently, maintaining the confidentiality of Ann's data will not necessarily prevent those details from being inferred from (public) data on Barry. For instance, if we know from data on others that eating a high-fat diet increases the risk of heart conditions, and that Ann is part of a community that eats a high-fat diet, we may infer that Ann is at higher risk of heart disease. As Viljoen argues, this relationality, or ability to make inferences about others from data, is not simply a negative externality of current business models; rather, it is integral to them "and constitutes much of what makes data production economically valuable in the first place."[324] Solow-Niederman takes this point one step further: "Contemporary information privacy protections do not grapple with the way that machine learning facilitates an inference economy in which organizations use available data collected from individuals to generate further information about both those individuals and about other people."[325] Accordingly, as Tontrup and Sprigman argue, data analysis "externalities strip the individual of the power to protect her privacy alone."[326]

Synthetic data increases these risks. Take, for example, collaborative filtering, which is based on correlations in data groups (for example, consumers and products).[327] Data on previous purchases and the features of consumers creates data spillovers among data subjects by enabling the

322. On the distinction between group privacy and collective privacy, see Alessandro Mantelero, *From Group Privacy to Collective Privacy: Towards a New Dimension of Privacy and Data Protection in the Big Data Era*, *in* GROUP PRIVACY: NEW CHALLENGES OF DATA TECHNOLOGIES 139, 148 (Linnet Taylor, Luciano Floridi & Bart van der Sloot eds., 2017).

323. *See, e.g.*, Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 96–105 (2014); DREXL, *supra* note 318, at 48 ("Given the potentials of big data analytics, which allows to draw probability conclusions from correlations between different pieces of information, it is no longer possible to neatly distinguish between non-personal and personal data."); Madison et al., *supra* note 164, at 7.

324. Viljoen, *supra* note 296, at 611.

325. Solow-Niederman, *supra* note 247, at 360.

326. Sprigman & Tontrup, *supra* note 242, at 29.

327. *See supra* Part II.

algorithm to make inferences about Ann based on collected data relating to Barry. Furthermore, even if a synthetic dataset anonymizes information about each individual but enables the algorithm to learn about groups, once the algorithm can connect an individual to a group, it can make informed inferences about her preferences. This begs the question of whether the concept of identifiability is sufficient to prevent harm to individuals and whether it can capture linkages or inferences on which synthetic data might be based. This also casts further doubt on the utility of individual control over one's privacy. Accordingly, lawmakers and courts face a dilemma: to define or interpret the scope of data privacy laws more broadly, thereby loosening the link between collected data and the individual and capturing more data flows under their scope, or to see some of the values that data privacy laws promote undercut by synthetic data processing.

3. Collective Data Harms

For similar reasons, synthetic data also increases collective data harms. Such harms arise when the analysis of data leads to decisions that might affect a group of individuals (such as residents of a town or a country) whose data may or may not constitute part of the dataset.[328] A well-known example involves the Facebook/Cambridge Analytics debacle,[329] where granting access to data on one individual led to revealing collected data related to others.[330] The data accumulated might have potentially led to manipulations that affected the political system, thereby indirectly impacting a group of individuals (all U.S. citizens),[331] with no need to include or tie each individual in this group to the dataset. The rise of synthetic data highlights the fact that even if anonymization of a dataset can be increased, privacy harms might also increase given that more nonpersonal data may now be available for use.

Accordingly, synthetic data challenges the effectiveness of existing data privacy laws in significant ways, calling into question their rationale: whether their goal is to protect the data of an individual[332] or to protect society from (certain) information-induced harms.[333] It also challenges the efficiency, and

328. *See generally* Solow-Niederman, *supra* note 247 (suggesting a new regulatory framework to contend with the impact of data on human lives in the inference economy).

329. Alvin Chang, *The Facebook and Cambridge Analytica Scandal, Explained with a Simple Diagram*, VOX (May 2, 2018, 3:25 PM), https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram [https://perma.cc/U23T-K4KA].

330. BRITTANY KAISER, TARGETED: THE CAMBRIDGE ANALYTICA WHISTLEBLOWER'S INSIDE STORY OF HOW BIG DATA, TRUMP, AND FACEBOOK BROKE DEMOCRACY AND HOW IT CAN HAPPEN AGAIN 217–37 (2019).

331. *Id.*

332. "This conceptualization of 'data as an individual medium' . . . privileges data processing's capacity to transmit knowledge about the data subject over its capacity to transmit knowledge about others." Viljoen, *supra* note 296, at 594.

333. Among the values we might protect through data protection are autonomy, dignity, identity, freedom of conduct, and democracy.

thus the continued relevance, of current data privacy laws in achieving an overall positive social-welfare balance once the effects of synthetic data are added to the equation.

The need to reexamine current data privacy laws is heightened by the fact that findings regarding the (in)applicability of privacy laws to synthetic datasets could affect informed data subjects' incentives to share or protect their data. Synthetic data could increase rational apathy toward data protection: if the law does not protect against the use of synthetic data in ways that might negatively affect citizens or consumers, and if data subjects have only a marginal ability to affect the collection of data that could serve as the basis for synthetic data generation, then individuals have no incentive to invest in data protection. This could increase individual and collective harms.

Indeed, if we assume that a core concern of privacy law is to act as a constraint on informational power—the power that an entity derives from having significant knowledge about an individual or group—then it matters little whether the source of the information is a direct identifier or a proxy for it. This leads us to ask, what should be the appropriate relational link with an individual for information to fall within the protective ambit of our laws? If we would like to protect individuals' privacy, these harms need to be addressed. Indeed, if the synthetic generation process is successful, then the dataset generated will constitute a convincing replica of a dataset about real-world people. If this replica dataset can be used to impact individuals, then irrespective of the precise data used to draw this inference, the threat to individuals' fundamental rights will be the same. Put differently, "it does not matter *who* the data 'came' from, but *what* such data says about [a person], and *how* such meaning is used to act upon [a person]."[334] The European Court of Justice, reasoning along these lines, adopted a purposive interpretation of its privacy regulation, the GDPR. By this logic, "the boundaries of the concept of information in data protection eventually stretch to encompass whatever substance engages the types of harms dealt with by data protection."[335] The Court could therefore find that when an inference is brought to bear on an individual, it does not matter whether that inference was derived from other personal data, the data of a third party, or synthetic data.[336]

Broadening privacy laws to capture these types of datasets might have intuitive appeal. Yet, whether the net of privacy law should be cast so widely remains highly contested.[337] Capturing entirely artificial data under data protection rules would potentially give individuals rights, such as rights to

334. Viljoen, *supra* note 296, at 608.

335. Dara Hallinan, *Data Protection Without Data: Could Data Protection Law Apply Without Personal Data Being Processed?*, 5 EUR. DATA PROT. L. REV. 293, 297 (2019).

336. For an excellent overview of the developments leading to international data privacy norms, see GLORIA GONZÁLEZ FUSTER, THE EMERGENCE OF PERSONAL DATA PROTECTION AS A FUNDAMENTAL RIGHT OF THE EU 75–107 (2014).

337. We reserve this exploration of the role of inference in data protection law for future work.

access and delete, in relation to such data—a possibility far removed from the image of a personal information dossier that underlies the idea of individuals' rights over their data. Furthermore, casting the net too widely would significantly restrain the utility of data. Finding an optimal balance is further complicated by the fact that inferences might involve minimal privacy loss or harm in some contexts (e.g., if Ann's colleagues use Ann's lunch choices to make inferences about her diet, and thus her health), but very significant harm in others (if an insurance company uses the contents of Ann's shopping trolley to do the same).[338]

The complex relationship between personal data protection and goals like guaranteeing functioning markets and enhancing innovation has long been recognized.[339] On the one hand, as Drexl notes, privacy protection can be regarded "as a condition for the functioning of markets . . . as well as a driver of innovation."[340] This is because consumers will be less willing to provide their data, and even to buy goods which they assume will affect their personal profile, if the law does not guarantee certain levels of data protection.[341] On the other hand, restrictions on the use of data may limit firms' ability to develop innovative products and processes. This has led to a balance whereby certain collection and use restrictions are imposed on PII.[342] But this balance is uprooted once synthetic data enters the equation, as it potentially changes both sides of the scale: it limits protections for individuals, and it has the potential to increase competition and innovation.

Until we find the correct balance in the scope of our privacy laws, a task which is beyond the scope of this Article, other laws may address some of these concerns indirectly. We turn to some relevant examples in the next Part.

IV. EFFECTS OF INCREASED DATA QUALITY

Synthetic data can potentially increase the quality of some datasets. Here we briefly explore how synthetic data does this, as well as the effects of such increased quality on data power and social welfare. In particular, we focus on its capacity to alter, in unique and fundamental ways, the relationship between data-based decision-makers and those affected by such decisions. We then analyze the ways in which the law affects the motivation to create more

338. *See generally* NISSENBAUM, *supra* note 314 (explaining that data effects are context-specific).

339. DREXL, *supra* note 318, at 7.

340. *Id.*

341. *Id.*; *see also* Niva Elkin-Koren & Michal S. Gal, *The Chilling Effects of Governance-by-Data on Data Markets*, 86 U. CHI. L. REV. 403, 417–18 (2019) (analyzing how the use of private data for governmental purposes affects motivations of data subjects to provide it).

342. For example, the CCPA contains a do-not-sell rule which enables Californian residents to opt out of the "selling, renting, releasing, disclosing, disseminating, making available, transferring or otherwise communicating" of their data; and a right to opt out. California Consumer Privacy Act of 2018, CAL. CIV. CODE §§ 1798.140(ad)(1), 1798.120(d), 1798.120(a) (West 2022).

accurate and complete datasets as well as the ability to use them. As we show, while the law has long been attentive to the fact that higher quality data can increase the accuracy of decision-making, not enough consideration has been given to higher quality data as a source of power.[343] In particular, the law currently places only limited restrictions on the ability to exploit or manipulate high-quality data in ways which negatively affect individuals or groups.[344] As such, synthetic data strengthens the need to consider the effects of increased data quality on data governance. While our analysis applies to both personal and nonpersonal synthetic data, and to its use in the private and the public spheres, it focuses mainly on the effects on individuals.

A. THE EFFECTS OF SYNTHETIC DATA ON DATA QUALITY

Data quality has multiple related, yet distinct, dimensions, of which two fundamental elements are completeness and accuracy. Completeness ensures that certain data features are not *unrepresented* in a dataset, while accuracy ensures that they are not *mispresented* in the dataset.[345]

Synthetic data offers the potential to strengthen both dimensions.[346] It does so by plugging gaps in datasets that result from difficulties in gaining access to collected data, which might emerge when collected data is rare or too costly or impractical to collect. Take, for example, completeness. During the COVID-19 pandemic, social distancing restrictions placed limits on data gathering for autonomous vehicle training. Google's Waymo responded by using synthetic data simulations of road conditions, based on data already collected, to continue training during this period.[347] Another example centers on efforts for debiasing machine learning algorithms.[348] Kate Crawford describes "dark zones or shadows [in datasets] where some citizens and

343. *See also* Nielsen, *The Too Accurate Algorithm*, *supra* note 26, at 45–47.

344. *See infra* Section IV.B.

345. Accuracy is defined differently in different fields. For our purposes, we use the definition used in machine learning, which is based on "the fraction of outputs of a model that are correct." Aileen Nielsen, Accuracy Bounding: A Regulatory Solution for the Algorithmic Society 6–9 (2022) (unpublished manuscript) (on file with authors) [hereinafter Nielsen, Accuracy Bounding]. Such a definition, however, is blind to the different (social) weights of different data points. Furthermore, it might mask poor decision-making, if the given dataset is unbalanced. For that, upsampling might be useful.

346. Data may never reach a complete level of accuracy as a picture of the real world, given data collection or generation limitations. Yet it might be accurate with regard to the specific features it includes.

347. Kyle Wiggers, *The Challenges of Developing Autonomous Vehicles During a Pandemic*, VENTURE BEAT (Apr. 28, 2020, 7:00 AM), https://venturebeat.com/2020/04/28/challenges-of-developing-autonomous-vehicles-during-coronavirus-covid-19-pandemic [https://perma.cc/49SW-5NR4].

348. There is an extensive literature on bias in automated decision making. *See, e.g.*, VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 81–83, 190–93 (2018); Barocas & Selbst, *supra* note 2, at 677–93.

communities are overlooked or underrepresented."[349] By augmenting collected datasets in a targeted way, synthetic data can be used to make datasets more representative, reducing the harms stemming from sampling or statistical bias.[350] The potential to synthesize less-biased data to train unbiased or minimally biased models in a variety of contexts, ranging from uneven gender and age distribution to regional biases, is emphasized in the literature on synthetic data.[351] The example of Amazon's algorithm for vetting job applicants is a case in point.[352] Such bias, which stems from training an automated system on a biased or incomplete dataset, can be addressed by adding synthetic data so that the dataset more accurately reflects either the real world, or the actual parameters that satisfy the needs of the decision-maker (e.g., choosing the best workers). In the same way, synthetic data can also be used to potentially correct overrepresentation (e.g., the overpolicing of certain communities, leading to their overrepresentation in criminal conduct datasets).

Now consider accuracy. As elaborated in Part I above, by acting as a (partial) replacement for missing data, synthetic data can potentially create more representative datasets. Furthermore, synthetic data can increase accuracy by verifying the correctness of the analysis performed on collected data, as exemplified by the use of synthetic data to create counterfactuals to fix overconfident AI models.[353]

Accordingly, synthetic data is a potentially useful technological tool to increase data quality. Yet in doing so, synthetic data draws to the fore a fundamental tension.[354] On the one hand, as noted, it may increase the quality of decisions and reduce bias based on misrepresentation. Furthermore, by potentially enabling more players to enter the market, it can indirectly increase quality by strengthening market-based motivations to provide higher-quality products and services.[355] On the other hand, a high-quality dataset could become a double-edged sword, as more accurate decisions might not

349. Kate Crawford, *Think Again: Big Data*, FOREIGN POL'Y (May 10, 2013, 12:40 AM), https://foreignpolicy.com/2013/05/10/think-again-big-data [https://perma.cc/ULU3-M9NJ].

350. Assefa et al., *supra* note 133, at 1–2. "Realistic synthetic data along with appropriate data imputation techniques offer a promising approach to tackle this challenge." *Id.*

351. HARVARD BUS. REV. ANALYTIC SERVS., *supra* note 32, at 6 ("It's not only bias in people—gender, race, and so on—but business biases, such as 'we will pay more attention to this region versus that region because we have more records from this region.'").

352. *See supra* Section I.B.2.

353. Singla et al., *supra* note 103, at 2.

354. For such tension with regard to collected data, see generally Nielsen, *The Too Accurate Algorithm*, *supra* note 26.

355. Shoshana Zuboff, *Big Other: Surveillance Capitalism and the Prospects of an Information Civilization*, 30 J. INFO. TECH. 75, 80 (2015).

always increase social welfare.[356] This is a nontrivial claim.[357] As Chen notes, overly accurate information can enable new forms of differentiation and categorization, which might have negative welfare effects on individuals and groups through exploitation or manipulation.[358] Accurate data can also give rise to a "loss of manoeuvre space" for individuals.[359] Likewise, it makes individuals and society more "readable," potentially reducing individuals' capacity for self-development and change, while exacerbating power and information asymmetries between those who process data and those who are subject to this data processing. Nielsen's thoughtful taxonomy of algorithmic accuracy harms is also relevant to more accurate data. She relates to three categories: "accuracy directly creating harms" (such as undermining human autonomy), "behavior associated with the pursuit of accuracy causing harms (side effects," such as privacy incursions in the acquisition and use of data), "and strategic responses to algorithms driven by" potentially mistaken perceptions (such as automation bias based on assumptions of algorithmic superiority).[360] More accurate data can increase all three. Such dangers are, of course, not unique to synthetic data. However, synthetic data exacerbates the regulatory challenge, bringing it to another level which might require a new balance between the competing considerations.

The potential harms of higher-quality data are best illustrated by synthetic data's potential contribution to the creation of more accurate digital profiles, which, in turn, enable more personalized treatment of individuals.[361] In the economic sphere, an individual may receive microtargeted offers for products that better fit their preferences, but possibly at higher, discriminatory prices that reflect their elasticity of demand.[362] In the social sphere, they may

356. Furthermore, as Nielsen argues, "the current overemphasis . . . on algorithmic accuracy is itself a form of welfare loss, in which practitioners put their efforts exclusively on a proxy that could at times be anticorrelated with social welfare." Nielsen, Accuracy Bounding, *supra* note 345, at 14.

357. The limited legal academic treatment of the principle of data quality so far has mainly focused on accuracy—that data should be correct and precise. *See* sources cited Section IV.A. However, data may be accurate but still incomplete. Rachel Levy Sarfin, *5 Characteristics of Data Quality*, PRECISELY (Nov. 2, 2022), https://www.precisely.com/blog/data-quality/5-characteristics-of-data-quality [https://perma.cc/5P5D-DHV7].

358. Jiahong Chen, *The Dangers of Accuracy: Exploring the Other Side of the Data Quality Principle*, 4 EUR. DATA PROT. L. REV. 36, 40 (2018).

359. *Id.*

360. Nielsen, *The Too Accurate Algorithm*, *supra* note 26, at 64.

361. *See* EXEC. OFF. OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES 7, 44 (2014) (outlining the type of data contained within a profile and the method by which profiles are created). *See generally* EXEC. OFF. OF THE PRESIDENT, BIG DATA: A REPORT ON ALGORITHMIC SYSTEMS, OPPORTUNITY, AND CIVIL RIGHTS (2016) (addressing harms that can result from supplying data to algorithmic profiling software).

362. *See* FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? 9–12 (2016) (outlining ways in which the use of big data can generate harmful consequences for low-income groups).

receive suggestions for connections (e.g., via LinkedIn) and content that better cater to their prior interests, but could potentially limit their viewpoints.[363] In the political sphere, their personalized digital feed could be designed to strengthen certain opinions and affect their political choices.[364] In the legal sphere, digital profiles could inform decisions made by law enforcement or judicial bodies (e.g., based on a suspect's presumed flight risk), and even lead to the creation of personalized laws.[365] While the ability to create a personal profile from a mosaic of data points has long been acknowledged,[366] synthetic data can increase it dramatically by overcoming barriers in the data value chain. These examples also illustrate that the same data can be used in both welfare-enhancing and welfare-reducing ways.[367]

Synthetic data also raises the opposite concern: it might reduce data quality when an analyst bases the generated synthetic data on incorrect assumptions. While many such instances can be addressed by technological means—including by testing the data against collected data or against counterfactuals,[368] in others the concern that the quality of data used for decision-making might be reduced, is a real one.

B. APPLICATION OF LAWS

The potential effects of synthetic data on data quality require us to determine to what extent current data governance laws that relate to data quality apply and whether such application furthers our social goals. Optimally, the law should encourage those instances in which more accurate data-based decisions increase welfare, while prohibiting those in which it significantly reduces it. This, in turn, requires the design of rules that can separate the two types in a cost-effective way—a tall order. Nonetheless, at least in some instances this may be achievable. As we show below, while some laws are fit for purpose, others might need to be changed to incorporate this new data generation reality.

We identify five challenges. In the first, laws relating to data quality that are conditioned on the provenance and nature of data—such as data privacy laws—might not apply to synthetic data, despite the fact that their application

363. *See* Christoph B. Graber, *The Future of Online Content Personalisation: Technology, Law and Digital Freedoms* 6–8 (Univ. of Zurich, J-Call Working Paper No. 2016/01, 2016) (discussing online content personalization and popular criticisms of it).

364. *See, e.g.*, Emma Graham-Harrison, Carole Cadwalladr & Hilary Osborne, *Cambridge Analytica Boasts of Dirty Tricks to Swing Elections*, GUARDIAN (Mar. 19, 2018), https://www.theguardian.com/uk-news/2018/mar/19/cambridge-analytica-execs-boast-dirty-tricks-honey-traps-elections [https://perma.cc/EBE8-BSMX].

365. *See generally* Omri Ben-Shahar & Ariel Porat, *Personalizing Negligence Law*, 91 N.Y.U. L. REV. 627 (2016) (arguing that courts can and should use data to create personalized reasonable person standards for negligence inquiries).

366. Yuichiro Tsuji, *Medical Big Data in Japan*, 8 J.L. & CYBER WARFARE 153, 154 (2020).

367. *See* Nielsen, *The Too Accurate Algorithm*, *supra* note 26, at 50–54.

368. *See supra* Part II.

may have increased social welfare.[369] This leads to the second challenge: where the provenance or nature of the data, including its method of collection or generation, is the distinguishing parameter for the application of a law, synthetic data increases enforcement challenges. This is because if both real and synthetic data can lead to similar decisions, the enforcer might not be able to distinguish which type of data (such as private data) was used, based on the observed outcome alone. Furthermore, firms might falsely claim to be using a synthetic data generator—a claim which might be hard to contradict without sophisticated reverse engineering.

The third challenge relates to whether and when the law should incentivize or mandate the use of synthetic data through legal requirements (such as reasonableness or risk reduction requirements) where such data can further legal goals or requirements. As noted, synthetic data widens the scope of possible options: it is no longer necessary to choose whether or not reliance on the collected data accumulated is sufficient to meet the legal requirements. Whether and when these additional options should be taken into account in a legal analysis depends, in part, on synthetic data's technical capabilities.

Most evidently, synthetic data should not be treated as a quick or even the most efficient fix for all illegal data-based decisions. Take, for example, illegal discrimination. The introduction of bias through the data sample used to train a model is just one way to create bias. It might equally be introduced, for example, at the point at which the target variable—the objective of the data mining, such as finding a creditworthy borrower—is defined, or when the characteristics associated with that variable (the class labels) are chosen. Another issue involves measurement limitations. To illustrate, the current state of the art of "fairness" in automated decision-making often equates it to parity between two groups. This implies that complex issues such as intersectional discrimination are treated in a simplistic manner that flattens the interests at stake.[370] More fundamentally, treating synthetic data as a means to tackle discrimination maintains a legal and policy focus on technical fixes.[371] In so doing, we delegate to technologists the task of determining what counts as discrimination and what constitutes a representative dataset. Furthermore, we view the question of bias through a narrow lens.[372] As Balayn & Gürses suggest, "[f]raming the debate around technical responses will obscure the complexity of the impact of AI systems in a broader political economy and ringfence the potential responses to the technical sphere."[373] In short, using synthetic data to correct a biased dataset is better conceived as a

369. *See supra* Part III.

370. BALAYN & GÜRSES, *supra* note 130, at 121.

371. Julia Powles & Helen Nissenbaum, *The Seductive Diversion of "Solving" Bias in Artificial Intelligence*, MEDIUM (Dec. 7, 2018), https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53 (on file with the *Iowa Law Review*).

372. BALAYN & GÜRSES, *supra* note 130, at 118–24.

373. *Id.* at 9.

minimum patch to address a flaw, rather than a holistic response to a complex legal and societal problem.

Yet, where synthetic data can be relatively easily and cost effectively used to reduce illegal harms, it should be taken into account by courts when considering the reasonableness of the conduct (such as reducing harms or self-help in tort law), or meeting quality-related requirements. Furthermore, it should also affect data governance requirements where accuracy is an important parameter, such as in content moderation practices.[374] In such instances, using synthetic data to increase quality, where such use is possible, is not only responsible but should also be mandatory.

The fourth challenge focuses on legal requirements of explainability and interpretability,[375] designed to further legal norms of transparency and reason-giving,[376] which in turn increase accountability. Generally, the more sophisticated the synthetic data generator, the more difficult it becomes to explain correlations and—even more strongly—causality in the data generated.[377] Some correct explanations might even be nonintuitive when compared to commonsense understandings of how the world works.[378] As a result, synthetic data could strengthen the transparency deficit, reducing the ability of third parties to separate high-quality data from entropy.

Where explainability is a mandatory legal requirement, the use of synthetic data might be limited, reducing the ability to enjoy its benefits. To illustrate, empirical research suggests that when synthetic data is combined with differential privacy, it may offer better privacy protection than traditional sanitization methods for some datasets.[379] However, it is not possible to predict which patterns in the dataset will be preserved, which could also lead to poor interpretability.[380]

---

374. Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401, 415–19 (2017); Danielle Keats Citron, *How to Fix Section 230*, 103 B.U. L. REV. 713, 753 (2023).

375. Russell and Norvig consider a machine learning model to be interpretable "if you can inspect the actual model and understand why it got a particular answer for a given input, and how the answer would change when the input changes." RUSSELL & NORVIG, *supra* note 233, at 729. For a different definition, see, for example, Rudin, *supra* note 233, at 206.

376. For some of the justifications for such requirement, see, for example, Jonathan Zittrain, *Intellectual Debt: With Great Power Comes Great Ignorance*, MEDIUM (July 24, 2019), https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c [https://perma.cc/4CUZ-NA4U].

377. *See, e.g.*, DAVID FREEMAN ENGSTROM, DANIEL E. HO, CATHERINE M. SHARKEY & MARIANO-FLORENTINO CUÉLLAR, GOVERNMENT BY ALGORITHM: ARTIFICIAL INTELLIGENCE IN FEDERAL ADMINISTRATIVE AGENCIES 75 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3551505 [https://perma.cc/G3B6-YFTS].

378. *Id.*

379. See studies cited in Stadler et al., *supra* note 141, at 1 (the article then challenges these researches).

380. *Id.* at 2.

To ensure accountability while enabling socially beneficial uses of synthetic data, we propose instead that for most uses of synthetic data for which explainability or interpretability are highly costly or impractical, accountability should relate to the data generation process, including a quality assurance component. Put differently, accountability and transparency should build more on the "ingredients and machinations" of the processes leading to the data-based decision, rather than the exact internal processes. Accordingly, accountability should relate to the choice of tool that created the synthetic data, the data inputs and assumptions used in the process, and the coder's modelling choices. It should thus be determined by technically informed experts and should not be assessed in the abstract but rather should relate to the specific use of the data. Similar proposals were made in the context of AI.[381] This may lead to what Sanfilippo, Frischmann, and Strandburg call procedural legitimacy.[382] Such a focus will also reduce exposure of trade secrets or privacy concerns.

To achieve procedural legitimacy, additional tools might need to be developed. These include standards for synthetic data generation and strengthening accuracy by design. Naturally, creating such standards carries some costs. While some elements of data generation standards will be relevant for many contexts, others might differ from one context to another due to differences in the necessary data and levels of risk. Other costs may arise from the standardization process itself, such as setting suboptimal standards and capture by strong players.[383] Another tool, suggested by Engstrom, Ho, Sharkey, and Cuéllar in the context of AI, requires users to engage in prospective "benchmarking" of full or partial datasets "by reserving a random hold-out sample of cases for human decision, thus providing critical information to smoke out when an algorithm has gone astray."[384] The use of contrarian algorithms, which test the robustness of the explanation of the generation method, may serve similar purposes. At the same time, procedural legitimacy might not be fit for all contexts in which synthetic data can be used. For instance, we might insist that where explainability is essential, synthetic data should only be used as training data but not as test data.[385]

The final challenge relates to laws that apply regardless of the provenance or nature of the data (such as those conditioned on the use of

381. Rita Matulionyte, Paul Nolan, Farah Magrabi & Amin Beheshti, *Should AI-Enabled Medical Devices Be Explainable?*, 30 INT'L J.L. & INFO. TECH. 151, 151 (2022).

382. Madelyn Sanfilippo, Brett Frischmann & Katherine Strandburg, *Privacy as Commons: Case Evaluation Through the Governing Knowledge Commons Framework*, 8 J. INFO. POL'Y 116, 118 (2018).

383. *See, e.g.*, Gal & Rubinfeld, *supra* note 198, at 762–63.

384. ENGSTROM ET AL., *supra* note 377, at 7.

385. Wang et al., *supra* note 228, at 825. Synthetic data must not be used to make clinical decisions. But, because the structure of the data is the same as the collected data, it can be used to plan and refine analyses before making a formal request to Public Health England's Office for Data Release to conduct the same analysis on the collected data.

the data, the resultant outcome, or data quality), yet whose assumptions and internal balances do not necessarily fit the effects of synthetic data.

Let us separate the different cases. Some legal prohibitions, which are based on the assumption that a certain conduct is always welfare reducing, do not require a different balance. Take, for example, consumer protection laws which prohibit data-based deceptive practices,[386] or laws that prohibit certain types of data-based bias,[387] whether directly or through fairness requirements.[388] Such laws apply based on the outcome and thus capture both real and synthetic datasets. Prohibiting such conduct increases social welfare, regardless of the nature of the data.

Next, let us consider laws which apply regardless of the provenance of the data, yet synthetic data can change the overall effects on social welfare. We illustrate such challenges by laws that focus on data quality as a requirement for decision-making. For instance, some types of health data are subject to extensive quality standardization.[389] Data quality is also one of the core principles found in data privacy frameworks.[390] For example, the Federal Privacy Act requires regulatory agencies to ensure that all records which are used in making any determination about an individual are made "with such accuracy . . . and completeness as is reasonably necessary to assure fairness to the individual in the determination."[391] Likewise, the accuracy of data is an important component of the Fair Credit Reporting Act ("FCRA"). Any entity providing data about its customers to consumer reporting agencies ("CRAs") for inclusion in a consumer report must provide accurate information.[392] In addition, CRAs are under an obligation to "follow reasonable procedures to assure maximum possible accuracy of the information concerning the individual about whom the [consumer] report relates."[393]

Such laws are based on the epistemic concern that where data is used as the basis for decision-making, the reliability of the decision will be affected by

386. Section 5(a) of the FTC Act declares "unfair or deceptive" acts unlawful. 15 U.S.C. § 45(a). The FTC has suggested that it may use this provision in order to sanction deceptive data-based practices such as dark patterns. BUREAU OF CONSUMER PROT., FED. TRADE COMM'N, BRINGING DARK PATTERNS TO LIGHT 3, 34 n.2 (2022), https://www.ftc.gov/system/files/ftc_gov/pdf/P214800%20Dark%20Patterns%20Report%209.14.2022%20-%20FINAL.pdf [https://perma.cc/FFQ9-AVFF].

387. *See supra* Section IV.A.

388. See, for example, the requirement of fairness in GDRP Article 5(1)(a) which has been interpreted to prohibit bias. SEBASTIÃO BARROS VALE & GABRIELA ZANFIR-FORTUNA, FUTURE OF PRIV. F., AUTOMATED DECISION-MAKING UNDER THE GDPR: PRACTICAL CASES FROM COURTS AND DATA PROTECTION AUTHORITIES 36, 39–40 (2022), https://fpf.org/wp-content/uploads/2022/05/FPF-ADM-Report-R2-singles.pdf [https://perma.cc/T47P-UKVB].

389. Gal & Rubinfeld, *supra* note 198, at 740.

390. *See, e.g.*, OECD, Privacy Guidelines, *supra* note 244, at 7.

391. 5 U.S.C. § 552a(e)(5).

392. 15 U.S.C. § 1681.

393. *Id.* § 1681e(b).

the quality of the underlying data.[394] It is mostly assumed that improved data quality will increase social welfare. As Barocas observes, when speaking of decisions that affect individuals, "[a]t issue is the simple fact that certain individuals may be subject to erroneous inferences" based on their data doubles.[395] Of concern are harms and rights breaches, like denials of credit, social welfare rights, bail, or employment opportunities, as well as societal harms, such as entrenching existing misrepresentations and stereotypes, scaling miscarriages of justice, and exacerbating information and power asymmetries.[396] As both state and private actors increasingly resort to data-informed decision-making across almost all areas of human activity, we might expect to see such accuracy requirements proliferate. Misrepresentation, underrepresentation (including by omission), or overrepresentation in a dataset are often viewed as examples of accuracy errors: i.e., what happens when things go wrong. As noted, synthetic data might improve compliance with such legal requirements.[397]

Yet, as noted above, not in all situations is increased quality welfare-enhancing. Indeed, the law does not promote data accuracy as an absolute value.[398] Some laws already recognize that, in some contexts, better data accuracy may have negative welfare effects. Consider, for instance, medical insurance. While an insurer might wish to have the most granular information possible about individuals seeking insurance in order to accurately assess the firm's risk of insuring them, this more granular profiling will work to the detriment of some individuals (e.g., those who are predisposed to certain illnesses). The law often recognizes the merits of broad insurance coverage and limits the information that can be relied upon by insurers to calculate premiums.[399] In this sense, the law acts as a constraint on accuracy, to promote a better power balance between the relevant parties and to achieve broader social goals. Such laws apply to both real and synthetic data.[400] The level of

394. Mittelstadt et al., *supra* note 2, at 5 ("[C]onclusions can only be as reliable (but also as neutral) as the data they are based on.").

395. Solon Barocas, *Data Mining and the Discourse on Discrimination*, PROC. DATA ETHICS WORKSHOP, 2014, at 1, 2.

396. Wachter & Mittelstadt, *supra* note 295, at 506–10.

397. *See* Tordable, *supra* note 112 (noting that a "central goal of synthetic data" is "to overcome the limitations [and] restrictions [on] obtaining . . . real-world data" by "us[ing] artificially generated data—which is similar to real-world data in a meaningful way").

398. OECD, Privacy Guidelines, *supra* note 244, at 7. There is a difference between not requiring a firm to reach a high level of accuracy and mandating it to artificially reduce the level of accuracy (e.g., by adding randomness to the dataset). The latter might also reduce accountability. Nielsen, Accuracy Bounding, *supra* note 345, at 58–59.

399. Louis DeNicola, *Which States Restrict the Use of Credit Scores in Determining Insurance Rates?*, EXPERIAN (Sept. 23, 2020), https://www.experian.com/blogs/ask-experian/which-states-prohibit-or-restrict-the-use-of-credit-based-insurance-scores [https://perma.cc/L6F2-43K2].

400. The findings of the U.K. Supreme Court in the case of *PJS* could be read in this way as providing some practical obscurity to the claimants by preventing further publication of private

data quality required by law may also affect such a balance. For example, completeness of a dataset is often required only to the extent necessary for the purposes of its processing.[401] While such requirements may be based on cost-benefit efficiency considerations, they might also implicitly recognize that there is merit in obfuscation and incomplete datasets in some circumstances.[402] Yet in most cases, the law does not mandate data holders to limit or reduce the quality of their datasets.

In light of the above, supplementary data governance tools are required. A first step is to determine in which contexts the costs of higher accuracy for social welfare outweigh its benefits. This involves not only the identification of specific products and services, but also the level of accuracy at which the balance will tip in each case. The second stage focuses on determining which, if any, regulatory tools might best achieve such a balance, based on a comparative analysis of the costs and benefits of applying different tools, informed by the enforcement of accuracy-limiting tools that are already in place.[403] Existing tools that can be used as potential sources of legal power include, inter alia, core principles found in data privacy law such as data security[404]; antitrust prohibitions that regulate the ability to collect or generate data[405]; or the use of the fair trade requirements included in the FTC Act to set bounds on the accuracy of predictive analytics.[406] Yet, we may need to adopt additional legal measures. Along these lines, Ohm suggests the creation of "throttling metrics," by which friction in the algorithm might protect important human values,[407] and Nielsen proposes that the accuracy of automated decision-making systems may be bounded where the output is too accurate for the context and leads to social harms.[408] The challenge lies in

information even though it was already circulating in the public domain. PJS v. News Grp. Newspapers Ltd. [2016] UKSC 26, [26] (appeal taken from EWCA).

401. OECD, Privacy Guidelines, *supra* note 244, at 7.

402. *See also* Genetic Information Nondiscrimination Act of 2008, 42 U.S.C. § 20009ff (prohibiting "discrimination on the basis of genetic information" in health insurance and employment scenarios, even if such data exists).

403. See examples throughout this Section. For a more complete list, see Nielsen, Accuracy Bounding, *supra* note 345, at 44–52.

404. OECD, Privacy Guidelines, *supra* note 244, at 7.

405. *See, e.g.*, Dina Srinivasan, *The Antitrust Case Against Facebook: A Monopolist's Journey Towards Pervasive Surveillance in Spite of Consumers' Preference for Privacy*, 16 BERKELEY BUS. L.J. 39, 97 (2019) (arguing that the monopoly rents that Facebook may have inflicted on consumers are a form of pervasive surveillance).

406. Dennis D. Hirsch, *From Individual Control to Social Protection: New Paradigms for Privacy Law in the Age of Predictive Analytics*, 79 MD. L. REV. 439, 497–502 (2020).

407. *See generally* Paul Ohm, *Throttling Machine Learning*, *in* LIFE AND LAW IN THE ERA OF DATA-DRIVEN AGENCY 214 (Mireille Hildebrandt & Kieron O'Hara eds., 2020) (arguing for the adoption of a machine-to-human performance ratio and a completeness quotient as throttle mechanisms). *See also* Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459, 466–73 (2019) (arguing that input exclusion is an inappropriate mechanism for regulation).

408. Nielsen, *The Too Accurate Algorithm*, *supra* note 26, at 15.

identifying those specific contexts in which such measures are justified to ensure a welfare-increasing balance between accuracy and other societal goals.

Unfortunately, several recently proposed laws exemplify missed opportunities to acknowledge and take account of the need for such balancing.[409] The Algorithmic Accountability Act, proposed in 2019, which was designed to require assessments of the costs and benefits of high-risk automated systems, focuses on the privacy and security of personal information.[410] Likewise, the Algorithmic Justice and Online Platform Transparency Act of 2021 incorporates an extremely narrow menu of tools to address algorithmic harms, which is limited to transparency, a right to data portability, and nondiscrimination.[411]

CONCLUSION

Synthetic data created a revolution in data generation. Its techniques have advanced to the point that in some instances it can replace collected datasets with fully or partially synthetic datasets characterized by a similar or even higher level of utility. Synthetic data has also brought about a qualitative shift, where fewer bits of collected data need to be combined to facilitate learning. While it is not a panacea, in some contexts it can significantly reduce access barriers to data, extend the scope of use of collected data, increase data quality, and reduce privacy and data security breaches. As such, it can be seen as a technological method for self-improvement of data-related decisions and for overcoming some obstacles to the collection and use of data. It is thus not surprising that the use of synthetic data is becoming commonplace. Indeed, as noted above, most of the data used to train automated systems will soon be synthetic.

At the same time, synthetic data creates data governance challenges. This Article focused on challenges resulting from three main effects of synthetic data: data access, data privacy, and data quality. As shown, synthetic data requires us to rethink the current legal status quo between data utility and data harms. Depending on the context, it could help alleviate or reinforce data-related social challenges, including fairness, equality, transparency, trust, and democracy, thereby illuminating many of the existing challenges in contemporary data governance.[412] Some challenges are not new: synthetic data reflects or strengthens issues that are also likely to arise with regard to collected data. Yet synthetic data could significantly increase their prevalence. For example, by increasing data externalities and data-based collective harms, synthetic data strengthens the case for regulation which focuses on usage

409. Nielsen, Accuracy Bounding, *supra* note 345, at 74–79.

410. Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019).

411. Algorithmic Justice and Online Platform Transparency Act of 2021, S. 1896, 117th Cong. §§ 4–5 (2021).

412. *See, e.g.*, Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 534 (2015).

rather than on data provenance. Or it further blurs the lines between personal and nonpersonal data, calling into question the utility of this binary divide.[413] Other challenges are unique to synthetic data. For instance, synthetic data challenges the ingrained assumption in some laws that firms need collected data to affect welfare. Yet in both cases, the challenges created by synthetic data often go to the core of the legal data regime, mandating answers to questions like what are we protecting and why. Given its nature as a general-purpose technology, such effects are relevant across numerous industries.

Some legal challenges identified are cross-sectional and pertain to all three areas analyzed in this Article, such as the need to reevaluate the level of risk to some rights (such as privacy, security, or nondiscrimination) or motivating factors (such as dynamic efficiency), once synthetic data is used. As a result, some laws are mismatched with legal challenges. In others, a nuanced application might be required. One example pertains to the interpretation of reasonableness requirements. Synthetic data can raise the benchmark (e.g., where bias can be reduced by adding synthetic data) or lower it (such as a decision by a monopolist not to share collected data where synthetic data is comparable). Other challenges may pertain to one area affected by synthetic data.

As the synthetic data train has already left the station, it is surprising and even disconcerting that almost no attempts have been made in the legal literature to deal with such challenges beyond the effects of synthetic data on privacy protection[414] or in the context of deep fakes.[415] Furthermore, it seems that there is a disconnect between legal requirements and what firms in the industry, and academics writing in the business context, believe such requirements to be, and both sides currently disregard some of the most pertinent relevant legal challenges.[416] This Article attempts to partially fill this gap. By doing so, it hopefully increases legal certainty for firms wishing to use synthetic data and for decision-makers applying laws to synthetic data. It also hopefully paints a picture of requirements for legal change, to potentially increase social welfare.

We leave for future research challenges which arise when synthetic data is regulated in a dissimilar fashion in different states or jurisdictions. Indeed, synthetic data provides a good example of the stark chasm that exists between U.S. statutory data privacy law and the EU data privacy framework.[417] We also

413. *See, e.g.*, Nadezhda Purtova, *The Law of Everything: Broad Concept of Personal Data and Future of EU Data Protection Law*, 10 LAW, INNOVATION & TECH. 40, 41, 73–75 (2018).

414. *See supra* Part III.

415. *See* Chesney & Citron, *supra* note 24, at 1771–86 (focusing on the negative effects). The use of deep fakes is not always negative. *See* Katrina G. Geddes, *Ocularcentrism and Deepfakes: Should Seeing Be Believing?*, 31 FORDHAM INTELL. PROP., MEDIA & ENT. L.J. 1042, 1044–45, 1060–61 (2021).

416. *See supra* Part III.

417. The EU and the United States seek to protect data privacy in different ways. *See* Schwartz & Solove, *supra* note 246, at 1872–77; Chander et al., *supra* note 256, at 1746–62.

leave for future research some areas of law which were not covered in this Article. These include, inter alia, whether certain property rights (such as copyrights)[418] limit the use of collected data as a basis for creating synthetic data; to what extent does the First Amendment apply to inference data;[419] how should deep fake images and fake profiles, based on synthetic data, be regulated;[420] who owns computer creations (such as synthetic images); is it ethical to use one's medical profile to create a "virtual twin" to be used in virtual clinical trials;[421] who is legally liable for harmful synthetic data; whether risk-based liability is better suited for synthetic data uses;[422] and how synthetic data affects those laws that can otherwise help in reducing negative data externalities, such as contracts and disclosure law. While canvassing the effects of synthetic data on all areas of the law is beyond the scope of this Article, these harms share many of the considerations elaborated above. Accordingly, we hope this Article has provoked thought in these areas as well.

418. *See* Chloe Xiang, *AI Is Probably Using Your Images and It's Not Easy to Opt Out*, VICE (Sept. 26, 2022, 6:00 AM), https://www.vice.com/en/article/3ad58k/ai-is-probably-using-your-images-and-its-not-easy-to-opt-out [https://perma.cc/CRU3-HW6E]. Some interesting issues include how fair use will be applied where images are used to train an AI system and what should the remedy be when only a small part of the images used to train the algorithm are real and their use is illegal.

419. *See* Paul Ohm, *How to Regulate Harmful Inferences*, JOTWELL (Dec. 22, 2021), https://cyber.jotwell.com/how-to-regulate-harmful-inferences [https://perma.cc/Z82B-Q72S].

420. *See generally* Yitzchak Besser, *Web of Lies: Hate Speech, Pseudonyms, the Internet, Impersonator Trolls, and Fake Jews in the Era of Fake News*, 17 OHIO ST. TECH. L.J. 233, 265–75 (2021) (discussing possible solutions to "impersonator trolls"); Solomon E. Asch, *Opinions and Social Pressure*, SCI. AM., Nov. 1955, at 31 (demonstrating that convergence to a singular point is human nature); Rex D. Glensy, *The Right to Dignity*, 43 COLUM. HUM. RTS. L. REV. 65 (2011) (discussing the various approaches to dignitary rights and calling for the affirmation for the right to dignity).

421. Proffitt, *supra* note 118.

422. *See generally* Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703, 707 (2016) (suggesting a focus on "minimizing the risk of reidentification and sensitive attribute disclosure" rather than on preventing harm).