SYNTHETIC DATA AND THE FUTURE OF AI

Peter Lee[†]

The future of artificial intelligence (AI) is synthetic. Several of the most prominent technical and legal challenges of AI derive from the need to amass huge amounts of real-world data to train machine learning (ML) models. Collecting such real-world data can be highly difficult and can threaten privacy, introduce bias in automated decision making, and infringe copyrights on a massive scale. This Article explores the emergence of a seemingly paradoxical technical creation that can mitigate—though not completely eliminate—these concerns: synthetic data. Increasingly, data scientists are using simulated driving environments, fabricated medical records, fake images, and other forms of synthetic data to train ML models. Artificial data, in other words, is training artificial intelligence. Synthetic data offers a host of technical and legal benefits; it promises to radically decrease the cost of obtaining data, sidestep privacy issues, reduce automated discrimination, and avoid copyright infringement. Alongside such promises, however, synthetic data offers perils as well. Deficiencies in the development and deployment of synthetic data can exacerbate the dangers of AI and cause significant social harm.

In light of the enormous value and importance of synthetic data, this Article sketches the contours of an innovation ecosystem to promote its robust and responsible development. It identifies three objectives that should guide legal and policy measures shaping the creation of synthetic data: provisioning, disclosure, and democratization. Ideally, such an ecosystem should incentivize the generation of high-quality

---

† Martin Luther King Jr. Professor of Law and Director, Center for Innovation, Law, and Society, UC Davis School of Law. I would like to thank Elizabeth Joh, Mark Lemley, Sarah Polcz, and workshop participants at the Lewis & Clark Fall Forum, the UC Davis-Jindal Global Law School symposium, the Transatlantic Tech Exchange Roundtable hosted by the German Marshall Fund, the UC Davis School of Law Schmooze, the Intellectual Property Scholars Conference at UC Berkeley School of Law, and the University of Texas School of Law for very helpful comments. I would also like to thank Dean Kevin Johnson and Senior Associate Dean Afra Afsharipour for providing generous institutional support for this project. This research was supported by a grant from the UC Davis Academic Senate Committee on Research. My thanks as well to McKenzie Deutsch and the UC Davis School of Law Library staff for excellent research assistance. I would also like to thank the outstanding editors of the *Cornell Law Review*.

synthetic data, encourage disclosure of both synthetic data and processes for generating it, and promote multiple sources of innovation. This Article then examines a suite of "innovation mechanisms" that can advance these objectives, ranging from open source production to proprietary approaches based on patents, trade secrets, and copyrights. Throughout, it suggests policy and doctrinal reforms to enhance innovation, transparency, and democratic access to synthetic data. Just as AI will have enormous implications for law, legal regimes can play a central role in shaping the future of AI.

INTRODUCTION

In Waabi World, there are lots of automobile accidents, but nobody actually gets hurt. Here, autonomous vehicles (AVs) such as self-driving trucks learn how to drive by navigating around swerving cars, absent-minded pedestrians, and meandering bicyclists. They gain millions of miles of experience practicing common driving scenarios and emergency maneuvers at highway speeds.[1] If a self-driving truck hits a pedestrian, or hits a thousand pedestrians, it's not a big deal. After all, Waabi World is a completely fabricated universe. All the cars, pedestrians, and bicyclists in this simulator are digital representations. Millions of miles of simulated driving around (and sometimes into) these digital artifacts generate synthetic data that trains AVs to drive in the real world. This Article explores the technical and legal benefits of synthetic data and proposes an innovation ecosystem to promote its robust and responsible development.

Synthetic data will define the future of artificial intelligence (AI). In general, AI refers to technologies that automate tasks "normally requiring human intelligence."[2] One of the most prominent AI technologies is machine learning (ML), which encompasses systems that "detect[] useful patterns in large

---

[1] *See* Waabi, *How Waabi World Works*, WAABI, https://waabi.ai/how-waabi-world-works/#:~:text=Waabi%20World%20uses%20AI%20to,play%20out%20like%20a%20movie [https://perma.cc/E7TS-4Z6X] (last visited Nov. 17, 2024).

[2] *Artificial Intelligence*, OXFORD REFERENCE, https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095426960 [https://perma.cc/EH2R-64A4] (last visited Nov. 17, 2024); *see* Jessica M. Meyers, *Artificial Intelligence and Trade Secrets*, 11 LANDSLIDE 17, 18 (2019) ("AI is defined as *a set of technologies that enable machine intelligence to simulate or augment elements of human thinking*.") (emphasis in original).

amounts of data."[3] A prototypical example of ML is an email spam filter, which "learns" from the content of emails and how a user interacts with them to differentiate spam from legitimate messages.[4] Other examples include Netflix's system of recommending videos, "predictive policing" software that anticipates where crime is likely to occur, and AVs that "learn" how to drive.[5] ML models require enormous amounts of so-called "training data" to discern patterns, make decisions, and render predictions.[6] To date, such training data has largely been real-world data based on observations of actual phenomena. However, the future of AI training data is synthetic.

While AI promises enormous benefits, its voracious appetite for data creates several significant technical and legal challenges. First, collecting massive amounts of real-world data is difficult and expensive.[7] Members of the public, for instance, only have so much tolerance for AVs learning to drive on residential streets where they can hit real pedestrians. Second, gathering massive amounts of training data threatens individual privacy.[8] Third, the real-world data used to train ML systems may be biased or unrepresentative, thus producing discrimination in automated decision making.[9] And fourth, AI systems training on (and copying) enormous amounts of text, images, video, and other real-world data may infringe copyrights on a massive scale.[10]

The limitations and risks of real-world data give rise to a solution that seems like (computer) science fiction: synthetic data. Increasingly, data scientists are generating synthetic data, such as simulated driving environments, fabricated medical records, and fake images, to train ML systems. In a recursive fashion, AI systems can generate synthetic data, which

---

3 Harry Surden, *Artificial Intelligence and the Law: An Overview*, 35 GA. ST. U. L. REV. 1305, 1311 (2019); *see* Nicol Turner Lee, Paul Resnick & Genie Barton, *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*, BROOKINGS (May 22, 2019), https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/ [https://perma.cc/N XE3-9X9Y]; David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 671 (2017).

4 *See* Surden, *supra* note 3, at 1312–16.

5 *See* Lehr & Ohm, *supra* note 3, at 669.

6 *See* Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 742, 745 (2021).

7 *See infra* Part I.A.

8 *See infra* Part I.B.

9 *See infra* Part I.C.

10 *See infra* Part I.D.

then trains other AI systems. In short, artificial data is training artificial intelligence.

The great promise of synthetic data is that it can mitigate many of the technical and legal challenges of real-world data. Real-world data is expensive to collect, riddled with errors, and must often be labeled by hand so that ML systems can properly learn from it. However, synthetic data generators can produce endless amounts of cheap, high-quality data with accurate labels automatically attached.[11] Synthetic data may sidestep thorny privacy issues by using information not based on identifiable individuals to train ML systems.[12] Synthetic data can also rectify the biases and lack of representativeness of real-world datasets.[13] Finally, synthetic data may allow developers to avoid copyright infringement from training ML systems on real-world (copyrighted) content.[14]

While this is a positive narrative, synthetic data also has the potential to do great harm. Synthetic data promises to significantly increase the analytic and predictive power of ML models, which parties can utilize for good or ill. Additionally, low-quality, poorly deployed synthetic data can exacerbate, rather than mitigate, the deficiencies of ML systems and even lead to the catastrophic collapse of AI models.[15] Both to realize its benefits and avoid its harms, much is at stake in getting synthetic data right.

AI will fundamentally shape society, and synthetic data will fundamentally shape AI. Already, AI systems are determining who gets jobs,[16] how medical resources are allocated,[17] and the potential recidivism risk of persons in the criminal justice system.[18] The enormous social implications of AI have spurred the European Union to agree to a sweeping AI Act and

---

11 *See infra* Part II.B.1.

12 *See infra* Part II.B.2.

13 *See infra* Part II.B.3.

14 *See infra* Part II.B.4.

15 *See infra* Part II.C.

16 Jeffrey Dastin, *Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women*, Reuters (Oct. 10, 2018, 8:50 PM), https://www.reuters. com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G [https://perma.cc/LC7M-M6XX] (describing Amazon's AI system for screening resumes from job applicants).

17 Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 Sci. 447, 447 (2019) (describing an AI system to determine healthcare costs of patients based on prior healthcare expenditures).

18 Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, ProPublica (May 23, 2016), https://www.propublica.org/article/

led the Biden Administration to issue a Blueprint for an AI Bill of Rights.[19] From a technical standpoint, three trends have driven the growth of AI: increased computing power, more advanced algorithms, and greater amounts of data.[20] With constraints on the availability of real-world data, synthetic data represents a crucial driver of AI, and the emerging market for synthetic data "seems to be having a moment right now."[21] Remarkably, research firm Gartner predicted that by 2024, 60% of the data used to develop AI and analytics projects would be synthetically generated.[22]

Just as synthetic data is a critical input to AI, it, too, has inputs. While synthetic data will shape the future of AI, law and policy can help shape the future of synthetic data. This Article examines several "innovation mechanisms" that can shape how data scientists and firms develop and deploy synthetic data. Innovation mechanisms include open source approaches and various forms of intellectual property protection, including patents, trade secrets, and copyrights. These innovation mechanisms define an ecosystem in which data scientists and firms develop (and sometimes protect) synthetic data. How policymakers structure these innovation mechanisms can

---

machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/5NHU-JB36] (describing an AI system that determines recidivism risk).

19 Adam Satariano, *E.U. Agrees on Landmark Artificial Intelligence Rules*, N.Y. TIMES (Dec. 8, 2023), https://www.nytimes.com/2023/12/08/technology/eu-ai-act-regulation.html [https://perma.cc/XE3S-6F3V]; WHITE HOUSE OFF. OF SCI. AND TECH. POL'Y, BLUEPRINT FOR AN AI BILL OF RIGHTS 24–25 (2022), https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf [https://perma.cc/8JKC-BZCN] [hereinafter OSTP, BLUEPRINT]; *see also* Michael D. Shear, Cecilia Kang & David E. Sanger, *Pressured by Biden, A.I. Companies Agree to Guardrails on New Tools*, N.Y. TIMES (July 21, 2023), https://www.nytimes.com/2023/07/21/us/politics/ai-regulation-biden.html [https://perma.cc/GUY7-GEMC] (describing a White House-brokered agreement with leading tech companies to establish "guardrails" for AI).

20 Meyers, *supra* note 2, at 18; Leinar Ramos & Jitendra Subramanyam, *Maverick* Research: Forget About Your Real Data—Synthetic Data Is the Future of AI*, GARTNER (June 24, 2021) at 5, https://www.gartner.com/en/documents/4002912 [https://perma.cc/9H94-AUUF].

21 Chris Metinko, *Synthetic Data Startups Pick Up More Real Cash*, CRUNCHBASE (Apr. 22, 2022), https://news.crunchbase.com/ai-robotics/synthet ic-data-vc-funding-datagen-gretel-nvidia-amazon/ [https://perma.cc/J3NG-P7BM].

22 Emma Keen, *Gartner Identifies Top Trends Shaping the Future of Data Science and Machine Learning*, GARTNER (Aug. 1, 2023), https://www.gartner.com/en/newsroom/press-releases/2023-08-01-gartner-identifies-top-trends-shaping-future-of-data-science-and-machine-learning [https://perma.cc/3G6Y-H9BX]; *see* Sara Castellanos, *Fake It to Make It: Companies Beef Up AI Models with Synthetic Data*, WALL ST. J. (July 23, 2021, 5:30 AM), https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601 [https://perma.cc/457Y-Z6N6]; Metinko, *supra* note 21.

significantly impact the character of synthetic data. Drawing on the concept of "designing for values" that informs debates over ethical AI,[23] this Article argues that an innovation ecosystem for synthetic data should advance three objectives: provisioning, disclosure, and democratization.

First, innovation mechanisms should encourage the provisioning of high-quality synthetic data. While provisioning information goods represents the primary function of innovation mechanisms, they perform other, less appreciated functions as well. Indeed, in the context of synthetic data, this Article contends that functions related to disclosing and democratizing innovations may be even more important than provisioning. Accordingly, second, this Article argues that innovation mechanisms should promote the disclosure of synthetic data and processes for generating it. AI is subject to a well-known "black box" phenomenon in which it is often difficult to discern how an AI system reaches a particular decision.[24] Such opaqueness is compounded when AI systems train on data whose content and provenance are unknown. Innovation mechanisms can help encourage the disclosure of synthetic data and processes for generating it, thus enabling greater verification of these crucial inputs. Third, innovation mechanisms should promote the democratization of the synthetic data landscape. Such democratization entails both widening access to synthetic data and increasing the number of independent sources generating it. One of the most concerning trends of the AI ecosystem is increasing industry concentration.[25] Large incumbents with enormous troves of data enjoy significant advantages in developing ML systems. However, innovation mechanisms should promote a more democratic landscape in which startups and new entrants can both access and generate synthetic data to train new generations of ML models.

To advance these normative objectives, this Article sketches the contours of an innovation ecosystem to promote the provisioning, disclosure, and democratization of synthetic data. It analyzes innovation mechanisms spanning open source production and various intellectual property regimes, including patents, trade secrets, and copyrights. In so doing, it proposes legal and policy reforms to improve the ability of these regimes

---

23 *See infra* Part III.

24 *See infra* notes 247–60 and accompanying text.

25 *See infra* notes 54–61 and accompanying text.

to promote the provisioning, disclosure, and democratization of synthetic data. It argues for more direct federal support for open source synthetic data, both through research funding and policy directives. It argues for enhanced disclosure regimes in patent law, drawing on recent Supreme Court jurisprudence on enablement and rehabilitation of the best mode requirement. It argues for broad exceptions and limitations to trade secrets and related bodies of law to promote data sharing, democratization, and employee mobility. And it argues for a robust fair use exception to copyright to determine the operation of copyrighted software that generates synthetic data. A diverse ecosystem featuring several innovative mechanisms can ensure the robust, transparent, and widespread development of synthetic data capabilities.

This Article makes several contributions. While AI has attracted enormous attention in legal scholarship, synthetic data—with a few exceptions—has been surprisingly underexplored.[26] To date, legal scholars have primarily examined synthetic data as a mechanism to protect privacy.[27] This Article fills an important gap by examining the broader implications of synthetic data for the future of AI—not just for protecting privacy, but also for correcting biased datasets, avoiding copyright infringement, and enabling a wide array of ML applications that are currently infeasible.

---

[26] Exceptions include César Augusto Fontanillo López & Abdullah Elbi, *On the Legal Nature of Synthetic Data*, NeurIPS (2022), https://openreview.net/pdf?id=M0KMbGL2yr [https://perma.cc/A64R-7XH4]; Michal S. Gal & Orla Lynskey, *Synthetic Data: Legal Implications of the Data-Generation Revolution*, 109 Iowa L. Rev. 1087 (2024); Georgi Ganev, *When Synthetic Data Met Regulation*, ICML 2023 Workshop on Generative AI & L. (2023), https://arxiv.org/abs/2307.00359 [https://perma.cc/G297-XNPE]. Gal and Lynskey's work is notable for its extended analysis of synthetic data and its legal implications. However, the work focuses primarily on the antitrust and privacy implications of synthetic data and conscientiously avoids discussing copyright. *See* Gal & Lynskey, *supra*, at 1092–93, 1156. This Article addresses additional dimensions of synthetic data, including its implications for copyright. Furthermore, this Article focuses substantially on innovation mechanisms to shape the creation of synthetic data, a topic that falls outside the scope of Gal and Lynskey's article.

[27] *See, e.g.*, Steven M. Bellovin, Preetam K. Dutta & Nathan Reitinger, *Privacy and Synthetic Datasets*, 22 Stan. Tech. L. Rev. 1, 2 (2019); Sharon Bassan & Ofer Harel, *The Ethics in Synthetics: Statistics in the Service of Ethics and Law in Health-Related Research in Big Data from Multiple Sources*, 31 J.L. & Health 87, 88 (2018); Liane Colonna, *Privacy, Risk, Anonymization, and Data Sharing in the Internet of Health Things*, 20 U. Pitt. J. Tech. L. & Pol'y 147, 150 (2020); Gal & Lynskey, *supra* note 26, at 1088; Fang Liu, *A Statistical Overview on Data Privacy*, 34 Notre Dame J.L. Ethics & Pub. Pol'y 477, 477 (2020); Alexandra Wood et al., *Differential Privacy: A Primer for a Non-Technical Audience*, 21 Vand. J. Ent. & Tech. L. 209, 209 (2018).

More broadly, this Article expands the set of policy levers available to "regulate" AI. Commentators have proposed numerous legal reforms to regulate AI harms, such as imposing liability for privacy violations, automated discrimination, and copyright infringement. This Article, however, shows that technology itself—such as high-quality synthetic data—can address many of these problems. Importantly, law can play a formative role in shaping such technological solutions. It can do so in a less heavy-handed manner than direct regulation by embedding public policy objectives in "innovation mechanisms" that technology developers voluntarily adopt. In this fashion, law can operate at two levels: it can directly regulate AI systems, and it can indirectly regulate them by shaping the incentives of those who develop critical technical inputs to those systems. Accordingly, this Article represents the first account of how various innovation mechanisms can and should guide the development of synthetic data. Finally, this Article sheds light on the recursive relationship of technology and law. AI will transform society and has already deeply impacted law in a variety of fields. But law helps determine the nature of technology, and laws and policies defining the innovation ecosystem for synthetic data will shape the future of AI.

This Article proceeds in four Parts. Part I examines how the need to train ML systems on massive amounts of real-world data produces several significant technical and legal difficulties. Part II introduces synthetic data, which can mitigate many of the limitations of real-world data. Part III examines three policy objectives that should guide the development of synthetic data: provisioning, disclosure, and democratization. Part IV examines how various open source and intellectual property-based innovation mechanisms—both in their current form and subject to reforms—can advance these objectives.

I
The Limitations and Risks of Using Real-World Data to Train Machine Learning Models

Effectively training ML systems requires "large amounts of high-quality, structured, machine-processable data."[28] Traditionally, developers have used real-world data to train ML systems. Thus, for instance, ML systems train on millions (or

[28] Surden, *supra* note 3, at 1316.

billions) of emails, medical records, and miles driven on actual roads. However, the need to amass huge amounts of high-quality, real-world data leads to several significant technical and legal challenges.[29]

A. The Challenges of Collecting and Labelling Data

First, developers face the sheer difficulty of amassing enough data to train ML systems. Enormous datasets, frequently encompassing millions of observations and beyond, are necessary to reap the predictive benefits of ML.[30] Consider, for instance, the difficulty of training AVs to drive themselves. Enormous volumes of driving data are needed to include every conceivable "edge case"—statistically improbable but possible events—such as a piano falling out of a truck in front of a vehicle.[31] However, training AVs on actual roads faces intrinsic constraints based on the number of physical vehicles available and human personnel to monitor them, in addition to safety concerns from real-world accidents.[32] In this and other contexts, "[c]ollecting quality data from the real world is complicated, expensive and time-consuming."[33]

Much is at stake in obtaining sufficient quantities of training data.[34] Large quantities of data are necessary to ensure

---

29 *See* Ramos & Subramanyam, *supra* note 20, at 3 (observing that real-world data is incomplete, expensive, biased, and restricted via regulations).

30 Lehr & Ohm, *supra* note 3, at 678.

31 Rob Toews, *Synthetic Data Is About to Transform Artificial Intelligence*, FORBES (June 12, 2022, 7:00 PM), https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=33d7bb517523 [https://perma.cc/VW7X-EU5C]; Laurie Clarke, *Is 'Fake Data' the Real Deal When Training Algorithms?*, THE GUARDIAN (June 18, 2022, 9:00 AM), https://www.theguardian.com/technology/2022/jun/18/is-fake-data-the-real-deal-when-training-algorithms [https://perma.cc/DDW5-VAWK]; *see* Ramos & Subramanyam, *supra* note 20, at 5.

32 *Cf.* Clarke, *supra* note 31 (noting the danger of training an ML system to recognize real-life drivers falling asleep at the wheel).

33 Toews, *supra* note 31; Yashar Behzadi, *A Community for Synthetic Data is Here and This is Why We Need It*, KDNUGGETS (Apr. 22, 2022), https://www.kdnuggets.com/f2022/04/community-synthetic-data-need.html [https://perma.cc/KQ4B-RQ9T] (noting the enormous time, labor, and expense needed to obtain and label data for computer vision models).

34 *See* Ramos & Subramanyam, *supra* note 20, at 5 ("[T]he fundamental constraint of AI progress will be data, not model architecture or computing."); Cade Metz et al., *How Tech Giants Cut Corners to Harvest Data for A.I.*, N.Y. TIMES (Apr. 8, 2024), https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html [https://perma.cc/WYN5-AF7E] (discussing how greater amounts of training data enhance the performance of large language models).

that ML systems provide accurate predictions. Notably, with large datasets, even relatively simple models achieve similar performance as more complex architectures.[35] Conversely, insufficiently small datasets may produce misleading, unreliable, or wildly inaccurate predictions. In some areas, such as farming, insufficient data precludes the development of many useful ML applications.[36] Quite simply, the inability to gather enough data significantly limits the development of AI systems.[37]

While the world produces seemingly limitless amounts of data, in some ways data collection is getting more difficult. As discussed further below, aggregating massive amounts of training data, such as by collecting personal information or scraping the web, may expose ML developers to significant liability for violating privacy or copyright laws.[38] In the "desperate hunt" for training data, OpenAI, Google, and Meta have even "cut corners, ignored corporate policies, and debated bending the law."[39] Compounding existing difficulties, firms like OpenAI, Google, Anthropic, and the New York Times have updated their terms of service to prohibit the use of their data to train AI models.[40] Recently, Zoom provoked a backlash with terms of service that seemingly allowed it to use Zoom call data to train AI models without consent, and it quickly reversed course.[41]

---

35 Ramos & Subramanyam, *supra* note 20, at 5.

36 James Steinhoff, *Toward a Political Economy of Synthetic Data: A Data-Intensive Capitalism That is Not a Surveillance Capitalism?*, 26 NEW MEDIA & SOC'Y 3290, 3295 (2024).

37 *See* SERGEY I. NIKOLENKO, SYNTHETIC DATA FOR DEEP LEARNING 12 (Springer vol. 174 2021) ("[M]any problems of modern AI come down to insufficient data."); Metz et al., *supra* note 34 (describing data supply problems with the development of leading AI models and noting that Meta considered buying publishing house Simon & Schuster to obtain training data).

38 *See infra* Part I.B & I.D. In internal meetings, Meta "conferred on gathering copyrighted data from across the internet, even if that meant facing lawsuits." Metz et al., *supra* note 34.

39 Metz et al., *supra* note 34.

40 Alistair Barr, *AI Hypocrisy: OpenAI, Google and Anthropic Won't Let Their Data be Used to Train Other AI Models, but They Use Everyone Else's Content*, BUS. INSIDER (June 2, 2023, 1:29 AM), https://www.businessinsider.in/tech/news/ai-hypocrisy-openai-google-and-anthropic-wont-let-their-data-be-used-to-train-other-ai-models-but-they-use-everyone-elses-content/articleshow/100713086.cms [https://perma.cc/DF4P-YSRT]; Kevin Roose, *The Data That Powers A.I. Is Disappearing Fast*, N.Y. TIMES (July 19, 2024), https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html [https://perma.cc/JG3V-UP3K]; Jess Weatherbed, *The New York Times Prohibits Using Its Content to Train AI Models*, THE VERGE (Aug. 14, 2023, 6:26 AM), https://www.theverge.com/2023/8/14/23831109/the-new-york-times-ai-web-scraping-rules-terms-of-service [https://perma.cc/MFR6-FMWU].

41 Melissa Goldin, *Zoom Says It Isn't Training AI on Calls Without Consent. But Other Data Is Fair Game*, ASSOCIATED PRESS (Aug. 9, 2023, 8:55 AM), https://

While the enforceability of terms of service is a matter of some debate, many entities now stringently guard their data, sometimes even from themselves.

Second, beyond requiring large amounts of data, ML systems require large amounts of *high-quality* data.[42] Much real-world data is incomplete or partially inaccurate.[43] Data scientists try to remove incomplete data, impute missing values, and otherwise "clean" data,[44] but these efforts further increase the expense and difficulty of amassing usable training data. According to one survey, the difficulties of gathering large quantities of high-quality data pose a challenge to 96% of companies seeking to implement ML applications.[45]

In some contexts, accurate labelling is critical for ensuring high-quality training data.[46] For so-called supervised learning, which comprises the most common form of ML training, data needs to be labeled correctly so that, for instance, a computer vision system can learn that an image of a horse depicts a "horse" and not a "cow."[47] ML-based AVs, fraud-detection systems, and medical diagnostic tools thus require accurate labels on millions of images, financial transactions, and medical records.[48] Partially or incorrectly labeled training data is beyond worthless; in fact, it can be highly damaging. Such data can train ML systems to make wrong decisions with a high degree of confidence.

---

apnews.com/article/fact-check-zoom-ai-privacy-terms-of-service-06ff47e-47439c2173390a4ca1389f652 [https://perma.cc/BQL9-E6J2].

42 *See* Mark Allinson, *Data Annotation as the Key to Successful AI Implementation*, Robotics & Automation News (May 3, 2023), https://roboticsandautomationnews.com/2023/05/03/data-annotation-as-the-key-to-successful-ai-implementation/68052/ [https://perma.cc/B9HZ-XBU2].

43 Lehr & Ohm, *supra* note 3, at 681.

44 *Id.* at 681–83; Allinson, *supra* note 42.

45 Dimensional Research, Artificial Intelligence and Machine Learning Projects are Obstructed by Data Issues 13 (2019), https://cdn2.hubspot.net/hubfs/3971219/Survey%20Assets%201905/Dimensional%20Research%20Machine%20Learning%20PPT%20Report%20FINAL.pdf [https://perma.cc/6T9X-MBAC]; *see also* Gal & Lynskey, *supra* note 26, at 1090.

46 *See* Evan Nisselson, *Deep Learning with Synthetic Data Will Democratize the Tech Industry*, TechCrunch (May 11, 2018, 11:11 AM), https://techcrunch.com/2018/05/11/deep-learning-with-synthetic-data-will-democratize-the-tech-industry/ [https://perma.cc/7UU6-W5KM].

47 *See* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Wash. L. Rev. 579, 591 (2018) (noting that supervised learning comprises the "technique overwhelmingly used to train commercial AI systems"). Alternate approaches, such as unsupervised learning and reinforcement learning, do not require labelled data, though they may benefit from it.

48 Allinson, *supra* note 42.

The need to accurately label massive amounts of training data produces two related challenges. First, it heightens the difficulty and expense of using real-world data. While some aspects of data labeling can be automated, in most ML systems to date, human workers label training data by hand. Such hand labeling is time consuming, expensive, and prone to error.[49] One industry participant estimates that sourcing, annotating, and cleaning real-world data can consume 80% of data scientists' time.[50] Second and relatedly, the individuals who hand-label training data often work in deplorable conditions. Commentators warn that "so-called AI systems are fueled by millions of underpaid workers around the world, performing repetitive tasks under precarious labor conditions."[51] Such grueling, rote, and unrecognized labor has been termed "ghost work."[52] One study found that the median wage for hand-labeling datasets on Amazon's Mechanical Turk was $1.77 per hour.[53]

The difficulties of amassing large amounts of high-quality, labeled data contribute to a third limitation of real-world data: significant concentration in ML industries. Large incumbents with ready access to data, such as Apple, Facebook, and Google, enjoy distinct advantages in training ML systems over small entities.[54] Large incumbents may generate in-house data from their own platforms,[55] buy data from external sources,

---

49 Steinhoff, *supra* note 36, at 3295; Metinko, *supra* note 21; Toews, *supra* note 31; Ramos & Subramanyam, *supra* note 20, at 10.

50 Sam Forsdick, *Artificial Advantage: Can Synthetic Data Make AI Less Biased?*, Raconteur (Aug. 1, 2022), https://www.raconteur.net/technology/artificial-advantage-can-synthetic-data-make-ai-less-biased/ [https://perma.cc/4WCA-L7Z6] (quoting Steve Harris, CEO of synthetic data firm Mindtech Global).

51 Adrienne Williams, Milagros Miceli & Timnit Gebru, *The Exploited Labor Behind Artificial Intelligence*, Noema (Oct. 13, 2022), https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/[https://perma.cc/VW6W-AEA8].

52 *See, e.g.*, Mary L. Gray & Siddharth Suri, Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass (2019).

53 Kotaro Hara et al., *A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk*, CHI '18: Proc. of the 2018 CHI Conf. on Human Factors in Computing Sys. 1, 1 (2018), https://arxiv.org/pdf/1712.05796 [https://perma.cc/H7TL-W8AR]. The working conditions of data labelers is, of course, a complex issue to assess. While labeling gigs are arguably exploitative, they also provide income to many under-resourced individuals, particularly in developing countries. At the very least, the conditions and remuneration of such work are alarming byproducts of ML systems' enormous need for labeled data.

54 *See* Levendowski, *supra* note 47, at 597–99.

55 Lina M. Khan, *Lina Khan: We Must Regulate A.I. Here's How.*, N.Y. Times (May 3, 2023), https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html [https://perma.cc/XY6U-5T2K]; Metz et al., *supra* note 34. While privacy laws and corporate policies sometimes prevent companies from

or simply acquire firms that possess such data.[56] Either way, vast stores of data raise barriers to entry for smaller entities.[57] For example, enormous troves of images and videos amassed by tech incumbents have been likened to a "moat that keeps the advances of machine learning out of reach from many."[58] While third-party data vendors can sell data to new entrants, they may charge prohibitively high fees.[59] As Federal Trade Commission (FTC) Chair Lina Khan observes, "The expanding adoption of A.I. risks further locking in the market dominance of large incumbent technology firms."[60] These incumbents leverage their control over data and other inputs to "exclude or discriminate against downstream rivals."[61]

B. Threats to Privacy

The need to train ML algorithms on massive amounts of data also threatens individual privacy.[62] According to Khan, AI tools can be "trained on private emails, chats and sensitive data, ultimately exposing personal details and violating user privacy."[63] Consider, for instance, GPT-3, an earlier version of the model underlying ChatGPT. OpenAI trained GPT-3 on 300 billion "tokens" (words or parts of words),[64] which, according to business information professor Uri Gal, were "systematically scraped from the internet."[65] These tokens encompass "books, articles, websites and posts—including personal information

---

training ML models on user data, firms such as Google have revised their terms of service to allow greater use of user data for this purpose.

56 *See* Levendowski, *supra* note 47, at 606–09 (exploring "build it" and "buy it" approaches to obtaining data by incumbents like Facebook and IBM).

57 Nisselson, *supra* note 46; Levendowski, *supra* note 47, at 609.

58 Nisselson, *supra* note 46.

59 *See* Samuel A. Assefa et al., *Generating Synthetic Data in Finance: Opportunities, Challenges, and Pitfalls*, ICAIF '20: Proc. of the First ACM Int'l Conf. on AI in Fin. 1, 3 (2020), https://dl.acm.org/doi/abs/10.1145/3383455.3422554 [https://perma.cc/5XRF-8EK2] (noting that exchanges and market data vendors sell financial data, but "the cost associated with accessing highly granular data is typically a deterrent to many").

60 Khan, *supra* note 55.

61 *Id.*

62 *See, e.g.*, Steinhoff, *supra* note 36.

63 Khan, *supra* note 55.

64 Tom B. Brown et al., Language Models are Few-Shot Learners 8 tbl 2.1 (2020) (indicating that all versions of GPT-3 were trained on 300 billion tokens), https://arxiv.org/pdf/2005.14165 [https://perma.cc/L39C-FGNV]; Metz et al., *supra* note 34 (explaining the relationship between tokens and words).

65 Uri Gal, *ChatGPT is a Data Privacy Nightmare. If You've Ever Posted Online, You Ought to be Concerned*, The Conversation (Feb. 7, 2023, 8:06 PM), https://

obtained without consent."[66] Gal writes that such massive data collection "is a clear violation of privacy."[67]

While data scientists attempt to anonymize data to protect individual privacy, these efforts are not always successful.[68] This problem is particularly salient for medical data,[69] where individual patient information has been "reidentified" from presumably anonymized data.[70] The limitations of anonymization have led to more sophisticated techniques to protect privacy, such as *k*-anonymity and differential privacy.[71] However, they are subject to limitations and present unsatisfactory tradeoffs between protecting privacy and maintaining data utility.[72]

Developers of ML systems face considerable liability for privacy violations.[73] For instance, collecting and using personal data to train ML systems may violate the European Union's General Data Protection Regulation (GDPR).[74] OpenAI appears to be violating the GDPR by offering no mechanism for users to check whether it stores their personal information or request that it be deleted.[75] In the United States, states have led the drive to regulate data privacy, including the use of personal data to train ML systems.[76] For instance, the

---

theconversation.com/chatgpt-is-a-data-privacy-nightmare-if-youve-ever-posted-online-you-ought-to-be-concerned-199283 [https://perma.cc/HV4W-EXNS].

66 *Id.*

67 *Id.*; *cf.* Helen Nissenbaum, *Privacy as Contextual Integrity*, 79 WASH. L. REV. 119 (2004) (noting that using publicly available data for an unauthorized purpose—such ML training—can breach "contextual integrity" and violate privacy).

68 *See* Bellovin, Dutta & Reitinger, *supra* note 27, at 13–16.

69 Jason Walonoski et al., *Synthea: An Approach, Method, and Software Mechanism for Generating Synthetic Patients and the Synthetic Electronic Health Care Record*, 25 J. AM. MED. INFORMATICS ASSOC. 230, 231 (2018).

70 *Id.*; *see also* Bellovin, Dutta & Reitinger, *supra* note 27, at 14–15.

71 *See* Bellovin, Dutta & Reitinger, *supra* note 27, at 16–19.

72 *See id.* at 18–20.

73 Jennifer Bryant, *Generative AI: A 'New Frontier,'* IAPP (Feb. 28, 2023), https://iapp.org/news/a/generative-ai-a-new-frontier/ [https://perma.cc/JCR5-CP7X].

74 Commission Regulation 2016/679, 2016 O.J. (L 119) 1 [hereinafter GDPR]; Adam Satariano, *G.D.P.R., a New Privacy Law, Makes Europe World's Leading Tech Watchdog*, N.Y. TIMES (May 24, 2018), https://www.nytimes.com/ 2018/05/24/technology/europe-gdpr-privacy.html [https://perma.cc/9988-9DR8]; *see* Steinhoff, *supra* note 36, at 3291; Allan Tucker, Zhenchen Wang, Ylenia Rotalinti & Puja Myles, *Generating High-fidelity Synthetic Patient Data for Assessing Machine Learning Healthcare Software*, 3 NPJ DIGIT. MED., 1, 1 (2020).

75 Gal, *supra* note 65.

76 Peter Karalis, *ANALYSIS: As AI Meets Privacy, States' Answers Raise Questions*, BLOOMBERG L. (Nov. 13, 2022, 9:00 PM), https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-as-ai-meets-privacy-states-answers-

California Privacy Rights Act's prohibition against "repurposing" data would prevent a firm from using data collected for one purpose to then train ML systems unless users consented or such training was consistent with the original rationale for data collection.[77] More recently, the California Privacy Protection Agency has advanced proposed rules to regulate business practices concerning AI and the collection of personal information.[78]

While the federal government lacks a comprehensive data privacy law akin to the GDPR, numerous federal laws potentially constrain the use of personal data to train ML systems. Federal laws such as the Family Educational Rights and Privacy Act (FERPA) and the Health Information Portability and Accountability Act (HIPAA) protect individual data.[79] Furthermore, the FTC has become more stringent in regulating firms' use of personal data to train ML systems. In 2022, the FTC forced Weight Watchers and its subsidiary, Kurbo, to delete a trove of data and destroy any models derived from it because the companies obtained the data in violation of children's privacy laws.[80] Similarly, the FTC ordered photo storage company Everalbum to delete photos and videos obtained in violation of privacy laws and to destroy any resulting models.[81] The FTC imposed a similar sanction on Cambridge Analytica.[82] The prospect of such "algorithmic destruction" poses enormous risk for companies collecting personal data to train ML systems.[83]

---

raise-questions [https://perma.cc/D7FT-LTG7] (discussing privacy laws in California, Virginia, Colorado, and Connecticut).

77 Eli MacKinnon & Jennifer King, *Regulating AI Through Data Privacy*, STAN. U. HUMAN-CENTERED A.I. (Jan. 11, 2022), https://hai.stanford.edu/news/regulating-ai-through-data-privacy [https://perma.cc/EZS4-NUQZ].

78 Khari Johnson, *Large California Companies Will Soon Face New Rules on How They Use AI*, CALMATTERS (Mar. 13, 2024), https://calmatters.org/economy/technology/2024/03/california-ai-rules-business/ [https://perma.cc/79Z2-HU47].

79 Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g; Health Insurance Portability and Accountability Act of 1996, 42 U.S.C. § 1320d; Assefa et al., *supra* note 59, at 1; *see* Bellovin, Dutta & Reitinger, *supra* note 27, at 8 n.26 (listing over a dozen federal statutes protecting privacy).

80 *See* United States v. Kurbo Inc., No. 3:22-cv-00946-TSH (N.D. Cal. Mar. 3, 2022).

81 *See* In re Everalbum, Inc., No. C-4743, at 4–5 (F.T.C. May 6, 2021).

82 *See* In re Cambridge Analytica, LLC, No. 9383, at 4 (F.T.C. Nov. 25, 2019).

83 *See* Rina Diane Caballar, *"Algorithmic Destruction" Policy Defangs Dodgy AI*, IEEE SPECTRUM (Apr. 15, 2022), https://spectrum.ieee.org/ai-concerns-algorithmic-destruction [https://perma.cc/VD8K-Z7ED]; Katharina Koerner, *Privacy and Responsible AI*, IAPP (Jan. 11, 2022), https://iapp.org/news/a/privacy-and-responsible-ai/ [https://perma.cc/4AA8-7QVW].

Liability for violating privacy rules hinders several aspects of developing ML systems. First, of course, it complicates and raises the cost of amassing enormous training datasets of personal information. Second, privacy concerns also prevent the productive sharing of data between firms and sometimes even between units within the same firm.[84] Sharing training data among firms in a given industry can accelerate collective development and refinement of ML systems. However, privacy laws complicate, and in some cases prohibit, such data sharing.[85] In sum, violating individual privacy is a major challenge of training ML systems with real-world data.

C. Bias in Automated Decision Making

Real-world data may also be biased, thus leading to discrimination in automated decision making. At least two kinds of bias are possible: training data may not accurately represent reality, or it may accurately represent a reality that reflects a legacy of discrimination.[86] ML models trained on such data, moreover, can replicate and "exacerbate problems of bias."[87] Given the increasing importance of ML systems in determining everything from who gets hired to how healthcare funds are allocated, bias in automated decision making can be extremely harmful.[88] According to FTC Chair Khan, "Because they may

---

84 *See* Assefa et al., *supra* note 59, at 1–2.

85 Dov Lieber, *The People in This Medical Research Are Fake. The Innovations Are Real.*, WALL ST. J. (Apr. 6, 2021), https://www.wsj.com/articles/the-people-in-this-medical-research-are-fake-the-innovations-are-real-11617717623 [https://perma.cc/2RS3-AATJ] (noting complications from sharing medical data); *see* Chao Yan et al., *A Multifaceted Benchmarking of Synthetic Electronic Health Record Generation Models*, 13 NATURE COMM'NS 1, 1 (2022) (same).

86 EXEC. OFF. OF THE PRESIDENT, PREPARING FOR THE FUTURE OF ARTIFICIAL INTELLIGENCE 30 (2016), https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf [https://perma.cc/9MP5-SX3N]; *cf.* Lee, Resnick & Barton, *supra* note 3 ("Data sets, which may be under-representative of certain groups, may need additional training data to improve accuracy in decision-making and reduce unfair results.").

87 EXEC. OFF. OF THE PRESIDENT, *supra* note 86, at 30; *see* Lee, Resnick & Barton, *supra* note 3; Eric Lander & Alondra Nelson, *Americans Need a Bill of Rights for an AI-Powered World*, WIRED (Oct. 8, 2021, 8:00 AM), https://www.wired.com/story/opinion-bill-of-rights-artificial-intelligence/ [https://perma.cc/V9VL-UGG9]. Algorithmic bias can arise from several sources beyond training data. For example, how an algorithm specifies a problem may be inherently biased. Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 675 (2016). Furthermore, "AI's largely homogenous community of creators, which skews toward white men," also contributes to bias. Levendowski, *supra* note 47, at 583.

88 This phenomenon has been well covered in the literature, and this Article's discussion will be brief. *See, e.g.*, Barocas & Selbst, *supra* note 87; Levendowski, *supra* note 47, at 586–87 (reviewing the relevant literature).

be fed information riddled with errors and bias, [AI] technologies risk automating discrimination—unfairly locking out people from jobs, housing or key services."[89]

Examples abound of ML-based discrimination due to biased training data.[90] Bias was evident in Amazon's AI hiring tool used to screen resumes from job applicants. Amazon trained the tool on resumes (mainly from men) it received over a ten-year period, and it quickly developed an anti-female bias.[91] In another example, researchers found that three facial recognition software systems most accurately recognized individuals with light complexions and males and were less accurate for darker females.[92] Among other factors, the lack of representation of darker females in training data contributed to distorted outcomes.[93] Additionally, an algorithm that identified patients with high healthcare needs based on past medical expenses systematically underestimated the needs of African American patients, who historically have had less access to healthcare.[94] Researchers found that the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, a system used by courts to assess an individual's recidivism risk, consistently predicted higher risks of reoffending for African Americans compared to similarly situated whites.[95] However, its predictions were largely inaccurate.[96] To the extent that training data such as arrest and incarceration rates reflect disparities in police practices and other criminal justice inequities,

---

89 Khan, *supra* note 55. While this discussion focuses on bias in automated decision making, biased training data affects the outputs of ML systems more generally, including content produced by generative AI systems. *See, e.g.*, Leonardo Nicoletti & Dina Bass, *Humans are Biased. Generative AI Is Even Worse*, BLOOMBERG (June 9, 2023), https://www.bloomberg.com/graphics/2023-generative-ai-bias/ [https://perma.cc/8TAE-6VVM] (examining how biased training data helps amplify race- and gender-based stereotypes in images generated by Stable Diffusion).

90 OSTP, BLUEPRINT, *supra* note 19, at 24–25; Obermeyer, Powers, Vogeli & Mullainathan, *supra* note 17, at 447; Levendowski, *supra* note 47, at 580–81 (describing bias in Google's word2vec toolkit trained on a corpus of data from Google News).

91 Dastin, *supra* note 16.

92 Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RES. 1, 12 (2018); *see also* Levendowski, *supra* note 47, at 584–85.

93 Buolamwini & Gebru, *supra* note 92, at 12; Lee, Resnick & Barton, *supra* note 3.

94 Obermeyer, Powers, Vogeli & Mullainathan, *supra* note 17, at 450.

95 Angwin, Larson, Mattu & Kirchner, *supra* note 18; *see* Levendowski, *supra* note 47, at 599–601.

96 Angwin, Larson, Mattu & Kirchner, *supra* note 18.

ML systems like COMPAS will tend to reflect such inequities as well.[97]

Preventing discrimination in automated decision making is one of the core principles of the Biden Administration's Blueprint for an AI Bill of Rights.[98] Among other protections, the Blueprint urges developers to use "representative data" to train AI systems. In sum, one of the stark limitations of real-world training data is that biased or incomplete datasets may produce discrimination in automated decision making.

D. The Potential for Massive Copyright Infringement

Another problem with using real-world data to train ML systems is the potential for massive copyright infringement. This concern has been most acute for generative AI systems like ChatGPT and Dall-E 3.[99] Much of the data that trains the models underlying these systems comes from readily available sources on the internet, such as news articles, blog posts, social media messages, photographs, videos, and software code.[100] Companies like OpenAI and Google have even transcribed over a million hours of YouTube videos to obtain data to train models.[101] Given the extremely low threshold for copyright protection—consisting primarily of originality and fixation in a tangible medium of expression[102]—most of this training "data" is copyrighted expression.[103] To the extent that training

---

97 Lee, Resnick & Martin, *supra* note 3, at 7.

98 OSTP, BLUEPRINT, *supra* note 19, at 23–29; *see* Lander & Nelson, *supra* note 87, at 4.

99 Concerns over infringing copyrights apply to other contexts as well. *See, e.g.*, Walonoski et al., *supra* note 69, at 230–31 (noting that intellectual property restrictions complicate the use of electronic health records in medical AI systems).

100 *See* James Vincent, *The Scary Truth About AI Copyright Is Nobody Knows What Will Happen Next*, THE VERGE (Nov. 15, 2022, 10:00 AM) https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data [https://perma.cc/DL7S-X5YE]. [hereinafter Vincent, *AI Copyright*], So-called "foundation models" like GPT-3/4, Stable Diffusion, and Codex are ML models pretrained on large-scale internet data and serve as the foundation for numerous downstream applications. *See generally* Peter Henderson et al., *Foundation Models and Fair Use* 1 (Stan. L. and Econ. Olin, Working Paper No. 584, 2023).

101 Metz et al., *supra* note 34.

102 17 U.S.C. § 102(a) ("Copyright protection subsists, in accordance with this title, in original works of authorship fixed in any tangible medium of expression . . . .").

103 *See* Lemley & Casey, *supra* note 6, at 745; Henderson et al., *supra* note 100, at 2; Vincent, *AI Copyright*, *supra* note 100; James Vincent, *Getty Images Sues AI Art Generator Stable Diffusion in the US for Copyright Infringement*, THE VERGE (Feb. 6, 2023, 11:56 AM), https://www.theverge.com/2023/2/6/23587393/

generative AI involves copying this content without authorization, such training potentially exposes ML developers to staggering copyright infringement liability.[104]

Indeed, numerous copyright owners have sued developers of ML systems for copyright infringement. In the United Kingdom, stock photo distributor Getty Images sued Stability AI, alleging that it copied 12 million images without authorization to train its Stable Diffusion image generator.[105] In the United States, lawyers are seeking class certification for a copyright infringement suit against Microsoft, its subsidiary GitHub, and its partner OpenAI for their AI-powered coding assistant GitHub Copilot.[106] In another case brought by the same lawyers, three artists are suing the AI art generation companies Stability AI, Midjourney, and DeviantArt for

---

ai-art-copyright-lawsuit-getty-images-stable-diffusion [https://perma.cc/H5MD-RTBB] [hereinafter Vincent, *Getty*].

104 *See* Lemley & Casey, *supra* note 6, at 754 ("There is at least one obstacle standing in the way of ML's seemingly inexorable learning curve. Virtually all the data used to compile training sets is protected by copyright."). Increasing risk for AI developers, copyright infringement is a strict-liability offense that provides for significant statutory damages, and opportunistic copyright owners are likely to register their works and may be tempted to sue if those works are used to train AI systems. *See id.* at 758-60; *see also* Pamela Samuelson, *How to Think About Remedies in the Generative AI Copyright Cases*, 67 COMMS. ACM 27, 29 (2024) ("If the plaintiffs succeed . . . , copyright statutory damage awards would almost certainly be staggeringly large as millions of works may have been used as training data."). Most commentary presumes that training ML models on copyrighted content constitutes prima facie copyright infringement, thus placing significant emphasis on whether such training constitutes fair use. However, this is not a universal view; several commentators suggest that such training does not constitute copyright infringement in the first place. *See, e.g.*, Oren Bracha, The Work of Copyright in the Age of Machine Production 8 (Feb. 16, 2024) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4581738 [https://perma.cc/6C88-JKZH] ("Notwithstanding the physicalist fact of reproduction, training copies involve no reproduction of copyrightable subject matter and therefore are not infringing."); Levendowski, *supra* note 47, at 595 (noting debates over whether copies made to train ML systems constitute "copies" under the Copyright Act for purposes of infringement). This Article analyzes the mainstream view, which is likely to feature prominently in litigation, while acknowledging that it is not universally shared.

105 Vincent, *Getty*, *supra* note 103; Gal, *supra* note 65 (noting that much of the data scraped from the internet to train ChatGPT is likely copyrighted).

106 Complaint Class Action & Demand for Jury Trial, Doe 1 v. GitHub, Inc., No. 3:22-cv-06823-KAW (N.D. Cal. Nov. 3, 2022); *see* James Vincent, *The Lawsuit That Could Rewrite the Rules of AI Copyright*, THE VERGE (Nov. 8, 2022, 11:09 AM), https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data [https://perma.cc/UMQ4-VMXT] [hereinafter Vincent, *Rewrite*].

copyright infringement.[107] The plaintiffs allege that the three firms infringed the copyrights of millions of artists by training their ML systems on five billion images scraped from the web.[108] The New York Times, comedian Sarah Silverman, the Authors Guild, and others have brought high-profile lawsuits against generative AI firms for copyright infringement.[109] Lawyers suggest that we are in a "Napster-era of AI" in which copyright infringement runs rampant, thus creating the possibility of industry-changing liability.[110]

Some have argued that certain uses of copyrighted content to train ML systems constitute fair use.[111] However, the extent to which this safe harbor applies is unclear and depends significantly on context.[112] Under U.S. copyright law, certain unauthorized uses of copyrighted content constitute fair use and are not infringing.[113] Some commentators suggest that using copyrighted content to train ML systems that do not themselves generate content should constitute "fair learning."[114]

---

107 Complaint Class Action & Demand for Jury Trial, Andersen v. Stability AI Ltd., No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023); James Vincent, *AI Art Tools Stable Diffusion and Midjourney Targeted with Copyright Lawsuit*, THE VERGE (Jan. 16, 2023, 6:28 AM), https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart [https://perma.cc/X9VA-XMDW] [hereinafter Vincent, *Art Tools*].

108 Vincent, *Art Tools*, *supra* note 107; Order on Motions to Dismiss & Strike, Andersen, v. Stability AI Ltd., Case No. 23-cv-00201-WHO (N.D. Cal. Oct 30, 2023) (dismissing most claims, allowing amendments to some claims, and permitting a claim for copyright infringement against Stability AI to proceed).

109 *See* Michael M. Grynbaum & Ryan Mac, *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*, N.Y. TIMES (Dec. 27, 2023 https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html [https://perma.cc/4Z5W-JPE7]); Riddhi Setty, *Sarah Silverman, Authors Hit OpenAI, Meta With Copyright Suits*, BLOOMBERG L. (July 10, 2023, 9:11 AM), https://news.bloomberglaw.com/ip-law/sarah-silverman-authors-hit-openai-meta-with-copyright-suits [https://perma.cc/SK8K-VEG5] (describing two class action suits involving Sarah Silverman and other authors); Max Zahn, *Authors' Lawsuit Against OpenAI Could 'Fundamentally Reshape' Artificial Intelligence, According to Experts*, ABC NEWS (Sept. 25, 2023, 3:50 PM), https://abcnews.go.com/Technology/authors-lawsuit-openai-fundamentally-reshape-artificial-intelligence-experts/story?id=103379209#:~:text=The%20case%20could%20fundamentally%20shape,legal%20analysts%20told%20ABC%20News [https://perma.cc/R3KR-UFUD] (describing litigation involving the Authors Guild and several prominent authors).

110 Vincent, *Rewrite*, *supra* note 106.

111 *See, e.g.*, Levendowski, *supra* note 47, at 622–30.

112 Lemley & Casey, *supra* note 6, at 776–79.

113 Courts apply a four-factor test that considers: "(1) the purpose and character of the use . . . ; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used . . . ; and (4) the effect of the use upon the potential market for or value of the copyrighted work." 17 U.S.C. § 107.

114 Lemley & Casey, *supra* note 6; Henderson et al., *supra* note 100, at 2.

In somewhat analogous fashion, courts have held that copying computer code simply to access unprotectable ideas, facts, or functionality constitutes fair use.[115] Put differently, "reading" (and in the process, copying) copyrighted content for nonexpressive purposes should not count as infringement.[116] However, if a model trained on copyrighted content generated outputs similar to that content, then such training would likely not constitute fair use.[117] The provenance of the training data also matters. If academic researchers and nonprofits generate training data and models, they are more likely to qualify for fair use.[118] On the other hand, commercial generation and use of copyrighted content to train ML systems is less likely to constitute fair use. Recent Supreme Court cases articulating both expansive[119] and narrow[120] conceptions of "transformative use"—a key factor that weighs in favor of fair use—provide little direct guidance.

Ultimately, there is no clear consensus on whether copying copyrighted content to train ML systems constitutes fair use.[121] And uncertainty in the face of staggering liability is highly unsettling.[122] The recent moves by Adobe, Microsoft, and IBM to indemnify the customers of their AI products from copyright infringement and other IP claims may allay the concerns of

---

115 *See* Sega Enterprises Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992); James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 662 (2016); Lemley & Casey, *supra* note 6, at 761–62.

116 Grimmelmann, *supra* note 115, at 662; *see* Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM L. REV. 1887, 1903–06 (2024); *see also* Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) (holding that digitally scanning library books to facilitate research, including text data mining and machine learning, constitutes fair use); Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) (holding that copying entire books to reveal snippets as part of the Google Books initiative was transformative and weighed in favor of fair use); *see* Henderson et al., *supra* note 100, at 5–7 (discussing several fair use decisions and their applicability to generative AI).

117 Lemley & Casey, *supra* note 6, at 777–78; Henderson et al., *supra* note 100, at 2.

118 Vincent, *AI Copyright*, *supra* note 100.

119 Google LLC v. Oracle Am., Inc., 593 U.S. 1, 30–31, 40 (2021) (holding that Google's copying of Sun's Java Application Programing Interface (API) "to create new products" was transformative and constituted fair use).

120 Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 550 (2023) (holding that Andy Warhol's illustrations derived from a copyrighted photograph did not constitute transformative use in part because the works served a shared "purpose," namely as magazine illustrations).

121 Lemley & Casey, *supra* note 6, at 746 ("Given the doctrinal uncertainty and the rapid development of ML technology, it is unclear whether machine copying will continue to be treated as fair use."); *see* Vincent, *AI Copyright*, *supra* note 100.

122 Lemley & Casey, *supra* note 6, at 769; Henderson et al., *supra* note 100, at 2; Levendowski, *supra* note 47, at 596–97.

some end users.[123] However, this move also increases potential liability for such developers, and it reflects enduring concerns over the ability of ML models to infringe copyrights.

Furthermore, whatever conclusions are drawn in the current landscape are subject to change. The European Union's AI Act will require entities to disclose any copyrighted material used to train foundation AI models.[124] This position reflects a "compromise between ignoring copyright and banning the use of copyright [sic] content in training AI models."[125] By requiring such disclosure, the proposal signals palpable unease with massive numbers of copyrighted works being used to train AI systems without authorization. Even if the proposed act does not lead to infringement liability, the requirement of having to identify potentially billions of pieces of copyrighted content in training data would impose an enormous burden on ML developers.

In sum, "[d]ata is rare, expensive, and time consuming to label, and access to the data that exists is often difficult, impossible, or ethically unsound."[126] To correct for various technical and legal difficulties of real-world data, data scientists are turning to a seemingly paradoxical solution: synthetic data.

II
SYNTHETIC DATA

A. Synthetic Data: An Overview

Synthetic data is artificially created data, such as fabricated numerical values, text, images, and videos.[127] Through synthetic data, developers aim to "reproduce the statistical properties and patterns of an existing data set by modeling its

---

123 Isaiah Poritz, *IBM Joins Microsoft, Adobe in Protecting AI Customers From Suits*, BLOOMBERG L. (Sept. 28, 2023, 5:00 PM), https://news.bloomberglaw.com/ip-law/ibm-joins-microsoft-adobe-in-protecting-ai-customers-from-suits [https://perma.cc/B5GT-CCVS].

124 Ryan Morrison, *EU Says Generative AI Makers Must Declare Copyrighted Content*, TECH MONITOR (Apr. 28, 2023), https://techmonitor.ai/technology/ai-and-automation/generative-ai-european-union-eu-copyright [https://perma.cc/X889-Y3PZ]; Supantha Mukherjee, Foo Yun Chee & Martin Coulter, *EU Proposes New Copyright Rules for Generative AI*, REUTERS (Apr. 28, 2023, 2:51 AM), https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/ [https://perma.cc/59HN-3QDP].

125 Morrison, *supra* note 124,

126 Steinhoff, *supra* note 36, at 3295.

127 *See* Bellovin et al., *supra* note 27, at 21; Gal & Lynskey, *supra* note 26, at 1090.

probability distribution and sampling it out."[128] In an iterative manner, developers often use AI to generate synthetic data, which then trains other AI models.[129] While synthetic data has been around in some form for decades, the widespread use of synthetic data to train ML models is a relatively recent development.[130] This application is highly promising given that synthetic data can mitigate several of the technical and legal limitations of real-world data.[131] Soon, the majority of data used to train AI systems will be synthetic.[132]

Synthetic data can take many different forms, and different kinds of such data differ with respect to how "synthetic" they are. While all synthetic data is at some level based on real data, the proximity of real and synthetic data is a question of degree.[133] These distinctions largely correlate with different technological methods for synthesizing data.

At one end of the spectrum, "data augmentation" generates synthetic data through modifying existing data.[134] Such processes are based on statistical modeling, and the resulting synthetic data represents extensions, extrapolations, or reconfigurations of real-world data. For instance, a system that creates synthetic overhead imagery may take a real-world satellite photo of an airplane pointing north and generate a synthetic image of that airplane pointing west. As another example, Israeli startup MDClone created a synthetic dataset of COVID-19 patients that mixed and reconfigured elements from actual electronic medical records.[135] A related approach involves "data perturbation," in which developers add noise to

---

128 Fernando Lucini, *The Real Deal About Synthetic Data*, MIT SLOAN MGMT. REV. (Oct. 20, 2021), https://sloanreview.mit.edu/article/the-real-deal-about-synthetic-data/ [https://perma.cc/SY3A-XWQ4].

129 *Id.*

130 Gal & Lynskey, *supra* note 26, at 1091.

131 *See infra* Part II.B.

132 Toews, *supra* note 31; *see also* Madhumita Murgia, *Why Computer-Made Data Is Being Used to Train AI Models*, FIN. TIMES (July 18, 2023), https://www.ft.com/content/053ee253-820e-453a-a1d5-0f24985258de [https://perma.cc/E8HN-76XT] (quoting Sam Altman, CEO of OpenAI, as stating that he was "pretty confident that soon all data will be synthetic data").

133 Even broad-based simulated universes, which generate highly synthetic data, are modeled in some fashion on known physical properties.

134 Steinhoff, *supra* note 36, at 3296; Ramos & Subramanyam, *supra* note 20, at 6.

135 Lieber, *supra* note 85.

an original dataset to synthesize new data.[136] In such cases, synthetic data is relatively proximate to real-world data.

Further along the spectrum, model-based synthetic data generators learn deep patterns in datasets that allow them to create novel outputs.[137] For instance, by scanning one hundred real faces, Datagen can train its ML system to create "millions of new identities."[138] Other synthetic data generators harness large language models (LLMs) like OpenAI's GPT-3.[139] Additionally, diffusion models "learn by corrupting their training data with incrementally added noise and then figuring out how to reverse this noising process to recover the original image."[140] By learning these patterns, diffusion models can then synthesize data by denoising random input.[141] Relatedly, neural radiance fields (NeRF) are a powerful technology for turning two-dimensional images into three-dimensional scenes.[142]

Generative adversarial networks (GANs) are a particularly prominent approach to synthesizing data.[143] GANs are systems that pit two neural networks against each other—for instance, one generating new images and the other trying to determine whether they are synthetic.[144] The generative model operates like a counterfeiter while the discriminative model functions like the police trying to detect counterfeit currency; these models compete until "the counterfeits are indistinguishable from the genuine articles."[145] This technology, which is used to create deepfakes, increases the fidelity of synthetic data to real data.[146]

At the most "synthetic" end of the spectrum, simulators create entirely new virtual worlds, and with them, new universes of synthetic data.[147] Examples include Facebook's AI

---

136 *See* Tucker et al., *supra* note 74, at 1.

137 Tucker et al., *supra* note 74, at 1; Ramos & Subramanyam, *supra* note 20, at 9.

138 Forsdick, *supra* note 50 (quoting Ofir Chakon, CEO of Datagen).

139 Toews, *supra* note 31.

140 *Id.*

141 *Id.*

142 *Id.*

143 *See generally* Bellovin et al., *supra* note 27, at 31–32; Gal & Lynskey, *supra* note 26, at 1098.

144 Toews, *supra* note 31.

145 Ian J. Goodfellow et al., *Generative Adversarial Nets*, 27 ADVANCES IN NEURAL INFO. PROCESSING SYS. 1, 1 (2014).

146 Castellanos, *supra* note 22.

147 Ramos & Subramanyam, *supra* note 20, at 9; Gal & Lynskey, *supra* note 26, at 1100–01.

Habitat and VIVID (Virtual Environment for Visual Deep Learning), which creates full cityscapes with moving vehicles, pedestrians, and dynamic weather.[148] As noted, AV company Waabi has created Waabi World to train self-driving vehicles,[149] and competitor Waymo has done similarly with its "Simulation City."[150] Simulated environments represent the "holy grail" of synthetic data because they produce entirely novel data that is "not extrapolated from an existing dataset."[151]

While synthetic data may replace real data, frequently it augments it.[152] This could be done serially, as when developers train a model on real-world data and then refine it with synthetic data.[153] Reversing the order, some medical researchers test hypotheses using synthetic data, then retest those hypotheses using real data from patients.[154] Data scientists also frequently combine real and synthetic data in the same dataset to train ML systems. For instance, medical researchers have augmented real data with synthetic data to boost the representation of undersampled groups.[155] AV companies combine millions of miles of real-world driving with billions of miles of synthetic driving to train self-driving automobiles.[156] American Express augments real data with synthetic data to train ML systems to identify rare types of credit card fraud.[157]

While it is tempting to think of synthetic data as a static set of "things," it may be more accurate to conceptualize it as a dynamic, customizable service. For instance, Datagen creates generative models of human faces that "spit out a completely new image each time" the system runs.[158] According to one commentator, "[i]n a nutshell, synthetic data technology allows

---

148 *See* Manolis Savva et al., *Habitat: A Platform for Embodied AI Research*, INT'L CONF. ON COMPUT. VISION (2019); Kuan-Ting Lai, Chia-Chih Lin, Chun-Yao Kang, Mei-Enn Liao & Ming-Syan Chen, *VIVID: Virtual Environment for Visual Deep Learning*, 2018 ACM MULTIMEDIA CONF. 1356, 1356 (2018).

149 *See* Waabi, *supra* note 1.

150 Steinhoff, *supra* note 36, at 3297.

151 *Id.*

152 Toews, *supra* note 31.

153 Forsdick, *supra* note 50.

154 Lieber, *supra* note 85.

155 Tucker et al., *supra* note 74, at 2; Laboratory for Information and Decision Systems, *The Real Promise of Synthetic Data*, MIT NEWS (Oct. 16, 2020) https://news.mit.edu/2020/real-promise-synthetic-data-1016 [https://perma.cc/KT5S-UNJR] [hereinafter LIDS].

156 Toews, *supra* note 31.

157 Castellanos, *supra* note 22.

158 Forsdick, *supra* note 50.

practitioners to simply digitally generate the data that they need, on demand, in whatever volume they require, tailored to their precise specifications."[159]

The market for synthetic data is large, and it is poised to explode. Research firm GlobalData identifies over 330 companies engaged in developing and applying synthetic data.[160] In 2021, the market for synthetic data was more than $110 million, and by 2027, it is projected to be $1.15 billion.[161] Numerous startups have arisen that generate synthetic data, and they have received significant funding.[162] As of 2022, leading synthetic data startups include Synthesis AI, Datagen, Anyverse, Truata, and Mostly AI.[163] In addition, large incumbents are developing synthetic data capabilities, either by acquiring startups or developing such capacity in-house. For example, Facebook acquired startup AI.Reverie, and Microsoft and NVIDIA are developing in-house synthetic-data generators.[164] Even non-tech companies such as J.P. Morgan, John Deere, and American Express are producing synthetic data to train ML models.[165]

B. The Benefits of Synthetic Data

Synthetic data has enormous benefits. While accuracy is always a concern,[166] several empirical studies find that models trained on synthetic data perform very similarly to those trained

---

159 Toews, *supra* note 31.

160 GlobalData, *Artificial Intelligence Innovation: Leading Companies in Synthetic Data*, VERDICT (June 2, 2023), https://www.verdict.co.uk/innovators-ai-synthetic-data-technology/#catfish [https://perma.cc/27Q6-CED7].

161 Metinko, *supra* note 21.

162 Elise Devaux, *[New] List of Synthetic Data Vendors—2022*, MEDIUM (Oct. 6, 2022), https://elise-deux.medium.com/new-list-of-synthetic-data-vendors-2022-f06dbe91784#:~:text=Mindtech%3A%20vendor%20of%20a%20synthetic,platform%20for%20deep%20learning%20applications [https://perma.cc/53BQ-N6PA]; Metinko, *supra* note 21; 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/G22N-BP4R].

163 Steinhoff, *supra* note 36, at 3295.

164 *See* Metinko, *supra* note 21 (discussing Facebook's acquisition of synthetic data firm AI.Reverie); Steinhoff, *supra* note 36, at 3295 (discussing Microsoft's open source Synthetic Data Generator); *NVIDIA Announces Omniverse Replicator Synthetic-Data-Generation Engine for Training AIs*, NVIDIA (Nov. 9, 2021), https://nvidianews.nvidia.com/news/nvidia-announces-omniverse-replicator-synthetic-data-generation-engine-for-training-ais [https://perma.cc/H4WK-PRUQ].

165 Steinhoff, *supra* note 36, at 3295.

166 Forsdick, *supra* note 50 (citing Professor Marek Rei, Imperial College London).

on real-world data.[167] Of course, synthetic data is far from a panacea for all that ails AI.[168] Synthetic data can dramatically increase the analytic power of ML systems—for good or ill—and poorly designed synthetic data can cause significant harms.[169] However, high-quality, conscientiously deployed synthetic data can mitigate many of the technical and legal difficulties of using real-world data to train ML systems.

1. *Enabling the Creation of Large Amounts of High-Quality Data*

Perhaps most importantly, synthetic data offers the possibility of virtually limitless, labeled data to train ML systems. First, synthetic data reduces the need for firms to engage in the time-consuming, expensive, and laborious process of collecting real-world data.[170] For example, in 2016, AV leader Waymo logged three million miles of real-world driving and 2.5 billion miles of simulated driving.[171] Synthetic data can provide ample instances of "edge cases" that are important for training ML systems, such as a piano falling out of a truck in front of an AV.[172] Second, while real-world data is often messy and incomplete, synthetic data offers the prospect of complete, error-free datasets.[173] Furthermore, software programs can automatically label synthetic data, thus obviating the need for human labeling.[174] Such automation can even label information "that is difficult or impossible for humans to label, such as velocity,

---

167 Steinhoff, *supra* note 36, at 3296.

168 For instance, synthetic data will not mitigate (and will likely accelerate) replacing human workers with AI systems. *See infra* note 226 and accompanying text.

169 *See infra* Part II.C.

170 Toews, *supra* note 31; *see* Lucini, *supra* note 128 (noting that synthetic data exponentially increases the amount of data available to train ML models).

171 Toews, *supra* note 31.

172 *Cf.* NVIDIA, *supra* note 164 ("Data generated in these virtual worlds can cover a broad range of diverse scenarios, including rare or dangerous conditions that can't regularly or safely be experienced in the real world."); *see* Clarke, *supra* note 31.

173 *See* Gal & Lynskey, *supra* note 26, at 1102–09 (discussing several benefits of synthetic data). Another benefit of synthetic data is that it can be purposefully designed to *include* errors and biases. Such data is useful for stress testing models, which can help winnow out underperforming models and refine promising ones. *See* Ramos & Subramanyam, *supra* note 20, at 12.

174 Forsdick, *supra* note 50; Toews, *supra* note 31; *see* NVIDIA, *supra* note 164 (describing how NVIDIA's synthetic data generator "augments costly, laborious human-labeled real-world data, which can be error prone and incomplete"). Additionally, synthetic data can also reduce the cost of data storage, which can be substantial. Gal & Lynskey, *supra* note 26, at 1103–04.

depth, occluded objects, adverse weather conditions or tracking the movement of objects across sensors."[175]

By resolving these limitations, synthetic data can greatly improve the performance, accuracy, and capabilities of ML systems.[176] The availability of high-quality, granular synthetic data promises to significantly expand applications of ML systems and can accelerate time to market for new services.[177] The ability of synthetic data to cover previously unknown scenarios, such as rare manufacturing defects, expands ML functionality.[178] Already, synthetic data has improved AI-based detection of credit card fraud and provided more thorough training for AI-based customer service chatbots.[179]

However, enhancing the performance of ML models with synthetic data is a double-edged sword.[180] While improved data can increase the benefits of ML systems, it can also increase the ability of such systems to "profile, nudge, exploit and manipulate individuals, with ramifications for the interpersonal, commercial, social, and political spheres."[181] As noted, synthetic data is not a panacea for all that ails AI and ML. Indeed, rather than obviating the need for exogenous regulation, the widespread use of synthetic data to create more powerful ML models may increase the need for regulatory monitoring and intervention.

Third, synthetic data also democratizes the data landscape.[182] As Rob Toews observes, "[o]ne of the main reasons that tech giants like Google, Facebook and Amazon have achieved such market dominance in recent years [in ML] is their unrivaled volumes of customer data."[183] The wide availability of cheap, accurate synthetic data can enable standalone firms, including startups and new entrants, to develop ML systems even when those firms do not have ready access to in-house

---

175 NVIDIA, *supra* note 164.

176 *See* Gal & Lynskey, *supra* note 26, at 1144–45 (noting that synthetic data can improve the completeness and accuracy of datasets).

177 *See* Assefa et al., *supra* note 59, at 3; Lucini, *supra* note 128.

178 Ramos & Subramanyam, *supra* note 20, at 8; *see* Gal & Lynskey, *supra* note 26, at 1104.

179 Castellanos, *supra* note 22.

180 *See* Gal & Lynskey, *supra* note 26, at 1145–46.

181 Gal & Lynskey, *supra* note 26, at 1093; *see* Jiahong Chen, *The Dangers of Accuracy: Exploring the Other Side of the Data Quality Principle*, 36 EUR. DATA PROT. L. REV. 42 (2018).

182 Nisselson, *supra* note 46; Gal & Lynskey, *supra* note 26, at 1113–14.

183 Toews, *supra* note 31.

training data.[184] ML systems trained by insurgents can then compete against, and perhaps even outperform, ML systems from large firms. For instance, startup AiFi is using synthetic visual data to develop a checkout-free retail system similar to Amazon Go.[185] According to Toews, "[t]he net effect of the rise of synthetic data will be to empower a whole new generation of AI upstarts and unleash a wave of AI innovation by lowering the data barriers to building AI-first products."[186]

2. *Mitigating Privacy Concerns*

Synthetic data can also alleviate—though not completely eliminate—privacy concerns over using personal data to train ML systems.[187] According to one commentator, it is "virtually impossible" to reverse engineer synthetic data or the algorithm used to create it to reveal underlying personal data.[188] While the privacy safeguards of synthetic data are helpful across all fields,[189] they are particularly valuable in two highly regulated industries: health and finance.[190]

In the medical field, synthetic data can resolve privacy concerns "that for years have held back the deployment of AI in healthcare."[191] For example, the National Institutes of Health

---

184 *See* Forsdick, *supra* note 50 ("For those companies without access to platforms like Instagram, there is another answer: synthetic data.").

185 Nisselson, *supra* note 46.

186 Toews, *supra* note 31. It is important to note a countervailing risk that synthetic data could enhance the power of industry incumbents if they possess real-world data that is essential to creating synthetic data. Gal & Lynskey, *supra* note 26, at 1114–15. However, these circumstances are becoming less likely, and the weight of authority suggests that synthetic data will democratize the data landscape. It is also possible that widespread use of synthetic data could lead to a decrease in antitrust enforcement and ultimately increase industry concentration. If synthetic data is widely available, antitrust authorities may be more permissive toward mergers and acquisitions of entities possessing large amounts of real-world data and relax current rules mandating data sharing, access, portability, interoperability, and standardization. *See* Gal & Lynskey, *supra* note 26, at 1116, 1118–20.

187 Koerner, *supra* note 83; Forsdick, *supra* note 50; *see* Gal & Lynskey, *supra* note 26, at 1122–26.

188 Lucini, *supra* note 128.

189 *See, e.g.*, Amazon Staff, *How Amazon Protects Customer Privacy While Making Alexa Better*, AMAZON (Jan. 28, 2022), https://www.aboutamazon.com/news/devices/how-amazon-protects-customer-privacy-while-making-alexa-better [https://perma.cc/CZA6-PP5P] (discussing how Amazon uses synthetic data to train its Alexa speech recognition system while protecting individual privacy).

190 Metinko, *supra* note 21; *see* LIDS, *supra* note 155 (noting increasing interest in synthetic data in the banking industry due to privacy concerns).

191 Toews, *supra* note 31.

(NIH) is working with startup Syntegra to create a synthetic dataset of COVID-19 patient records that duplicates the properties of real-world data without containing any links to the original information.[192] Anthem, a major health insurer, is partnering with Google Cloud to generate massive amounts of synthetic medical histories, healthcare claims, and related data to train ML systems on fraud detection and personalized health care.[193] Similarly, Illumina, a leading genetic sequencing company, is partnering with Gretel.ai to create synthetic genomic datasets.[194] By mitigating privacy concerns, synthetic data can speed up ML-based medical innovation.[195]

Addressing privacy concerns also renders synthetic data much more shareable.[196] This is evident not only in healthcare,[197] but also in finance. At one bank, due to privacy and security concerns, financial "data was so highly protected, gaining access to it was an arduous process, even for purely internal use."[198] Currently, privacy regulations limit data sharing between banks. Accordingly, banks largely rely on their own in-house data to train fraud-detection systems. However, if banks could pool their synthetic data, they could obtain a more holistic account of how people interact with banks in general, not just their own institutions.[199] By sidestepping privacy concerns and enabling greater data sharing, synthetic data can accelerate ML development by multiple parties in parallel.[200]

Notably, use of synthetic data does not eliminate all privacy concerns.[201] Notwithstanding assertions to the contrary, there is a risk that as synthetic data becomes less distinguishable from the real data upon which it is based, it becomes easier to reconstruct that real data.[202] Depending on how an ML model

---

192 Lucini, *supra* note 128.

193 Toews, *supra* note 31.

194 *Id.*

195 Lieber, *supra* note 85.

196 Ramos & Subramanyam, *supra* note 20, at 19 ("[S]ynthetic data is also inherently more shareable, avoiding the privacy pitfalls that plague real datasets.").

197 Toews, *supra* note 31.

198 Lucini, *supra* note 128; *see* LIDS, *supra* note 155 ("Companies and institutions, rightfully concerned with their users' privacy, often restrict access to datasets—sometimes within their own teams.").

199 Lucini, *supra* note 128.

200 *See* Assefa et al., *supra* note 59, at 2.

201 An important related question is whether synthetic data falls within the scope of privacy laws, which generally exclude anonymous data. *See* Gal & Lynskey, *supra* note 26, at 1126–37.

202 Toews, *supra* note 31; Lieber, *supra* note 85.

is trained, a synthetic dataset could in theory "leak."[203] One early-stage player, DataCebo, even allows users to calibrate the trade-off between privacy and fidelity to real-world data when generating synthetic data.[204] More broadly, as synthetic data increases the analytic power of ML models, such models will be able to analyze and manipulate individuals—thus implicating the core concerns of privacy law—even if they do not directly use personal information to do so.[205] As a general matter, however, synthetic data promises to improve data privacy, thereby enabling greater data sharing and promoting more robust innovation.

3. *Reducing Bias in Automated Decision Making*

Third, synthetic data can counteract the systemic bias that fuels automated discrimination in ML systems. Recall that real-world datasets may not accurately represent reality or may accurately represent a reality that reflects a history of discrimination.[206] If biased data trains ML models, those models can amplify such biases in their decisions and predictions. However, data scientists can sidestep or supplement real data by using synthetic data designed to ensure diversity and representativeness to train ML models.[207] They may use entirely synthetic datasets or selectively augment real data, filling in gaps and bolstering underrepresented groups with synthetic data.[208] Such injection of "domain knowledge" from real-world experience can enhance the functionality of ML systems.[209] As research firm Gartner observes, "Real datasets are typically incomplete, imbalanced and not fully representative of the business domain. Synthetic data is designed to address these shortcomings."[210]

Importantly, synthetic data is not a "silver bullet" for eliminating bias in automated decision making.[211] Much depends on the conscientious design and monitoring of data synthesis. One

---

203 Bellovin et al., *supra* note 27, at 37–38; *see id.* at 39–40 (discussing adversarial machine learning, in which an external party seeks to force leaks in a process for generating synthetic data); Gal & Lynskey, *supra* note 26, at 1125–26.

204 Toews, *supra* note 31.

205 Gal & Lynskey, supra note 26, at 1140–42.

206 *See supra* Part I.C.

207 Forsdick, *supra* note 50; Clarke, *supra* note 31.

208 Forsdick, *supra* note 50; Gal & Lynskey, *supra* note 26, at 1144–45.

209 Ramos & Subramanyam, *supra* note 20, at 7; Toews, *supra* note 31.

210 Ramos & Subramanyam, *supra* note 20, at 2.

211 Forsdick, *supra* note 50.

paradigm of synthetic data is statistically modeling it to parallel some real-world dataset. However, if that real-world data exhibits biases, then synthetic data modeled on that data will, too.[212] These biases, moreover, may be amplified considerably given the enormous volume of data that synthetic data generators can fabricate.[213] As noted, firms can use synthetic data to generate entirely new datasets that conform to some predetermined conception of fairness. However, without interrogating that definition of fairness, the synthetic dataset may reflect other biases.[214] More generally, commentators stress the importance of keeping humans "in the loop" to ensure fairness in automated decision making, even when using synthetic data.[215]

4. *Avoiding Copyright Infringement*

Fourth, synthetic data promises to sidestep thorny issues of copyright infringement. As mentioned, ML systems that train on copyrighted works without authorization, such as generative AI platforms like ChatGPT and Stable Diffusion, face potentially enormous infringement liability.[216] It is possible that some uses of copyrighted content to train ML systems constitute fair use.[217] However, the current legal uncertainty creates massive risk for developers. Accordingly, if generative AI systems can train on synthetically generated text, images, sounds, and videos, they can in theory avoid copyright issues.[218] This is an important benefit of synthetic data that the legal literature has not yet fully appreciated. However, commercial firms are already touting this benefit of synthetic data.[219] Sam

---

212 Lucini, *supra* note 128; Forsdick, *supra* note 50.

213 Forsdick, *supra* note 50.

214 Lee et al., *supra* note 3 ("Fairness is a human, not mathematical determination, grounded in shared ethical beliefs.").

215 Joe McKendrick, *Fighting Bias in AI Starts with the Data*, ZDNET (Aug. 13, 2022), https://www.zdnet.com/article/fighting-bias-in-ai-starts-with-the-data/ [https://perma.cc/H9EH-Q83H]. *But see* Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Humans in the Loop*, 76 VAND. L. REV. 429 (2023) (cautioning against haphazard human-in-the-loop governance systems and advocating conscientious regulation).

216 *See* discussion *supra* Part I.D.

217 *See supra* notes 111–23 and accompanying text.

218 *See* Forsdick, *supra* note 50 (discussing this phenomenon in the context of image-based ML systems).

219 *See, e.g.*, Kyle Wiggers, *Synthesis AI Raises $17M to Generate Synthetic Data for Computer Vision*, TECHCRUNCH (Apr. 28, 2022, 5:00 AM), https://techcrunch.com/2022/04/28/synthesis-ai-raises-17m-to-generate-synthetic-data-for-computer-vision/ [https://perma.cc/2TCG-L862] (reporting that computer

Altman, CEO of OpenAI, has lauded synthetic data as a way to develop more powerful models without relying on copyrighted training data.[220]

Notably, the claim that training ML systems on synthetic data can avoid copyright infringement requires substantial qualification. First, if synthetic data itself is copyrighted, then training ML systems on that data may still infringe if ML developers have not cleared relevant copyrights. As discussed further below, certain synthetic text, images, and other data may constitute copyrightable expression, although the authorship requirement would bar protection of synthetic data wholly generated by AI systems with minimal human creative input.[221] If entities holding copyrights on synthetic data do not authorize a third party to use such data to train an ML system, such use may infringe.

Second, if synthetic data infringes other parties' copyrights, then training ML systems with that data may constitute copyright infringement. In this sense, synthetic training data may not resolve issues of copyright infringement so much as shift them earlier in the supply chain.[222] As noted, at some point all synthetic data is based on real data.[223] A generative AI system may try to avoid copyright infringement by training on synthetic images. However, if synthesizing those images involves making unauthorized copies or derivative works of real-world, copyrighted images (e.g., taking a copyrighted photo of an airplane and rotating the plane ninety degrees), then the synthetic images themselves may infringe.[224] And if training the ML system involves making copies or derivative works of infringing synthetic images, then those ML systems may infringe as well. Put differently, the degree to which ML systems

---

vision synthetic data startup Synthesis AI markets its synthetic data as avoiding copyright infringement risk).

220 Metz et al., *supra* note 34.

221 *See infra* notes 399–401 and accompanying text.

222 *See* Katherine Lee, A. Feder Cooper & James Grimmelmann, Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain (Aug. 1, 2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551 [https://perma.cc/U6H2-4Z8C] (discussing the concept of the AI supply chain).

223 Lucini, *supra* note 128.

224 Whether the outputs of an AI model operating with little or no human creative input can infringe a third party's derivative-work right is a complex doctrinal question upon which circuits would likely differ. *See* Daniel J. Gervais, *AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines*, 52 SETON HALL L. REV. 1111, 1127 (2022).

can avoid copyright infringement by training on synthetic data may depend on how "synthetic" that data is. If synthetic data is highly proximate to real-world (copyrighted) data, both the synthetic data and the ML system training on it may infringe copyrights.

C. The Importance of Ensuring High-Quality Synthetic Data

The stakes of getting synthetic data right are extremely high. On the positive side, high-quality synthetic data can mitigate many pressing ills of ML, including the high cost of data collection and labeling, privacy violations, bias in automated decision making, and massive copyright infringement. On the negative side, synthetic data poses several potential harms. As noted, synthetic data can radically enhance the analytic power of ML models, which parties can use toward manipulative and harmful ends.[225] More generally, the increased power of ML from synthetic data can accelerate some of the well-known harms of ML, such as job losses due to automation.[226] While these harms presume that synthetic data "enhances" the capabilities of ML systems, this section explores a more fundamental threat: low-quality synthetic data can exacerbate the limitations of real-world data and severely undermine the functionality of ML systems.

On the one hand, low-quality synthetic data that is too similar to reality can exacerbate the limitations of real-world training data. As noted, the closer that synthetic data is to real data, the more likely it is to leak personal information, thus violating privacy laws.[227] Additionally, synthetic data's fidelity to real-world data, which may be biased and unrepresentative, or ground-truth reality, which may reflect legacies of

---

225 *See supra* notes 180–81 and accompanying text.

226 *See* Cade Metz, *What's the Future for A.I.?*, N.Y. TIMES (Apr. 4, 2023), https://www.nytimes.com/2023/03/31/technology/ai-chatbots-benefits-dangers.html [https://perma.cc/R7Y5-QYUL]; Anton Korinek & Joseph E. Stiglitz, *Artificial Intelligence and Its Implications for Income Distribution and Unemployment*, *in* THE ECONOMICS OF ARTIFICIAL INTELLIGENCE: AN AGENDA 349, 349 (Ajay Agrawal, Joshua Gans & Avi Goldfarb eds., 2019); Cynthia Estlund, *What Should We Do After Work? Automation and Employment Law*, 128 YALE L.J. 254, 257 (2018); Daron Acemoglu & Pascual Restrepo, *The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment*, 108 AM. ECON. REV. 1488, 1488 (2018).

227 Isabelle Bousquette, *AI-Generated Data Could Be a Boon for Healthcare—If Only It Seemed More Real*, WALL ST. J. (Aug. 2, 2023, 7:00 AM), https://www.wsj.com/articles/ai-generated-data-could-be-a-boon-for-healthcareif-only-it-seemed-more-real-5bfe52dd [https://perma.cc/SND2-WJTX].

discrimination, can vastly amplify the problem of bias in automated decision making.[228] Given the enormous amounts of data that synthetic data generators can produce, small biases can lead to very large distortions in ML system outputs. Finally, training ML systems on synthetic data that is only minimally different from real-world, copyrighted content may not avoid copyright infringement issues.

On the other hand, low-quality synthetic data that diverges too much from reality can cause significant harms. Training ML systems on inaccurate or misrepresentative synthetic data can undermine product development, fraud detection, resource allocation, and all of the other critical functions that ML systems perform.[229] For instance, IBM's Watson Health gave incorrect cancer treatment advice due to being trained on erroneous synthetic patient records.[230] In healthcare, concerns that synthetic data does not accurately represent the characteristics of target patient populations has chilled adoption of this potentially useful resource.[231]

At a broader level, misrepresentative or biased synthetic data can have catastrophic effects on the future of AI. In this context, "bias" refers not necessarily to the perpetuation of social inequalities, but more generally to the deviation of synthetic data from reality. While this deviation is problematic for many kinds of synthetic data—including synthetic data deliberately designed to train ML models—it is particularly relevant to "unintentional" synthetic training data, such as artificially generated content scraped from the web that was not intended to train ML models but ends up doing so. As noted, in a recursive fashion, AI systems generate synthetic data, which then trains other AI systems, which then generate more synthetic data, ad infinitum. If most LLMs train on data scraped from the web, "then they will inevitably train on data produced by their predecessors."[232] Such recursive training can lead to "model

---

228 *See* Ramos & Subramanyam, *supra* note 20 ("[T]he data generation process can introduce bias into AI models and inadequately represent the underlying real-world phenomena."); Gal & Lynskey, *supra* note 26, at 1110 ("In some situations, adding synthetic data increases the *risk of duplicating bias or errors*.").

229 Lucini, *supra* note 128.

230 Casey Ross & Ike Swetlitz, *IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show*, STAT (July 25, 2018), https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/ [https://perma.cc/3XQL-D68F].

231 Bousquette, supra note 227; *see* Gal & Lynskey, *supra* note 26, at 1109–10.

232 Ilia Shumailov et al., *AI Models Collapse When Trained on Recursively Generated Data*, 631 NATURE 755, 755 (2024); *see* Ilkhan Ozsevim, *Research Finds*

collapse," which computer scientists define as "a degenerative process affecting generations of learned generative models, in which the data they generate end up polluting the training set of the next generation. Being trained on polluted data, they then mis-perceive reality."[233] While the best antidote for model collapse is actual, real-world data, high-quality synthetic data that more closely represents reality may prevent or slow such collapse.[234] Ultimately, low-quality synthetic data can render AI models irretrievably divorced from reality.

III
Policy Objectives for Developing Synthetic Data

The enormous value of synthetic data and the need to ensure its quality raise pressing questions over how to promote its robust and responsible development. Synthetic data is a critical input to AI that will shape the future of this transformative technology. But inputs have inputs, too. Among the inputs to synthetic data are laws and policies defining the innovation ecosystem in which parties generate synthetic data and use it to train ML systems. This Article sets forth a legal and policy framework to shape that ecosystem. Below, it will consider several "innovation mechanisms" that impact the development of synthetic data and systems for generating it.[235] These mechanisms range from open source production to intellectual property regimes, notably patents, trade secrets, and

---

ChatGPT & Bard Headed for 'Model Collapse,' AI Mag. (June 20, 2023), https://aimagazine.com/articles/research-finds-chatgpt-headed-for-model-collapse [https://perma.cc/UB5S-SHN2].

233 Shumailov et al., *supra* note 232, at 755; *see also* Aatish Bhatia, *When A.I.'s Output Is a Threat to A.I. Itself*, N.Y. Times (Aug. 25, 2024), https://www.nytimes.com/interactive/2024/08/26/upshot/ai-synthetic-data.html [https://perma.cc/U8L7-AM3A].

234 *See* Kamya Pandey, *Sam Altman Says That OpenAI Doesn't Fully Understand What is Going on Inside Its AI Models*, MediaNama (June 4, 2024), https://www.medianama.com/2024/06/223-sam-altman-says-that-openai-doesnt-fully-understand-what-is-going-on-inside-its-ai-models/ [https://perma.cc/M5XP-EZK4] (noting OpenAI CEO Sam Altman's statement that high-quality data—whether real or synthetic—can prevent model training corruption).

235 *See infra* Part IV. Of course, synthetic data raises numerous additional policy concerns that fall outside the scope of this Article. For example, the Food and Drug Administration must address whether and how to approve AI-based software as a medical device (AI-SaMD) when such software is trained on synthetic data. *See* Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson & Faisal Mahmood, *Synthetic Data in Machine Learning for Medicine and Healthcare*, 5 Nature Biomed. Eng'g 493, 493 (2021). This Article brackets such questions and focuses on legal and policy mechanisms that can shape the creation of synthetic data.

copyrights. Before analyzing those innovation mechanisms, however, one must have a sense of what they should try to accomplish. Accordingly, this Part explores several policy objectives that innovation mechanisms should promote.

In so doing, it draws on the concept of "designing for values" that informs current debates about ethical AI. The harms of AI—several of which this Article has examined—have spurred a robust debate on "ethical AI" and how to "align" AI with human values.[236] In this context, the use of synthetic data to train ML systems can be understood as a way to align AI with the values of privacy, nondiscrimination, and respect for the creations of others. One approach to ensuring ethical AI is to deploy external laws and regulations to "keep AI systems in check."[237] More fundamentally, however, commentators advocate a "design for values" approach that integrates values directly into the technical design of AI systems.[238] This approach involves "coding" ethical constraints and values into AI systems.[239] One variant of this approach involves "Constitutional A.I.," in which designers provide an AI model with a list of principles (a constitution) to govern its operation.[240]

This Article applies a "design for values" approach to laws and policies governing the innovation ecosystem that will produce synthetic data. Thus far, ethical debates over AI have focused on how people should design AI systems. This Article approaches this issue at a meta level, asking how we should design an innovation ecosystem in which people design elements of AI systems, including synthetic data. It focuses on one set of policy tools—those aimed at promoting innovation—that can shape the character of synthetic data. In so doing, it illustrates the multiple ways that law can "regulate" AI. Certainly, law can directly regulate AI by imposing liability for privacy violations,

---

236 Virginia Dignum, *Responsible Artificial Intelligence: Designing AI for Human Values*, ITU J.: ICT Discoveries 1 (2017) ("Currently, there is an increasing awareness that a responsible approach to AI is needed to ensure the safe, beneficial and fair use of AI technologies.").

237 Dan Hendrycks et al., *Aligning AI with Shared Human Values*, Int'l Conf. on Learning Representations 2021 1, 9 (2021).

238 Dignum, *supra* note 236, at 2.

239 Francesca Rossi & Nicholas Mattei, *Building Ethically Bounded AI*, Thirty-Third AAAI Conf. on A.I. (AAAI-19) 9785, 9786 (2019); Hendrycks et al., *supra* note 237, at 1 (discussing the need for algorithms to be fair, safe, prosocial, and useful).

240 Kevin Roose, *Inside the White-Hot Center of A.I. Doomerism*, N.Y. Times (July 11, 2023), https://www.nytimes.com/2023/07/11/technology/anthropic-ai-claude-chatbot.html [https://perma.cc/9J5G-QV62].

discrimination, and copyright infringement. Additionally, law can indirectly regulate AI by shaping the innovation ecosystem and incentives of those who design its critical technical inputs. This Article argues that legal and policy mechanisms should aim to create a robust and varied innovation ecosystem that incentivizes the creation of high-quality synthetic data, encourages the disclosure of synthetic data and the processes used to create it, and ensures multiple sources of innovation. Accordingly, it argues that an innovation ecosystem for synthetic data should promote the values of provisioning, disclosure, and democratization.

A. Provisioning

First, and most foundationally, innovation mechanisms should facilitate the provisioning of synthetic data and processes for generating it. Synthetic data, like other information assets, is a public good. Such goods are nonrival, which means that one party's consumption of the good does not reduce its availability for others.[241] Furthermore, such goods are nonexcludable, which means that in the absence of some kind of legal protection, it is generally difficult to exclude parties from consuming such goods.[242] As commonly understood, public goods such as synthetic data may be subject to undersupply in a competitive economy. Put differently, while synthetic data may be initially expensive to produce,[243] it is trivially inexpensive to copy. Free riders can copy millions of synthetic medical records, numerical figures, and images with the click of a button, potentially undermining incentives to create synthetic data in the first place. Innovation mechanisms should thus encourage the production of this valuable public good.

However, while provisioning represents the central function of innovation mechanisms, the need to perform this function is somewhat limited in the context of synthetic data. Firms have strong market incentives to develop synthetic data,

---

241 *See* Mark A. Lemley, *Property, Intellectual Property, and Free Riding*, 83 TEX. L. REV. 1031, 1050–51 (2005).

242 *See id.* at 1051. *See generally* Kenneth J. Arrow, *Economic Welfare and the Allocation of Resources for Invention*, *in* THE RATE & DIRECTION OF INVENTIVE ACTIVITY: ECON. AND SOC. FACTORS 609, 614–16 (Nat'l Bureau Comm. for Econ. Rsch., Comm. on Econ. Growth of the Soc. Sci Rsch. Couns. eds., 1962) (observing the difficulties of preventing outside parties from appropriating existing information).

243 Lucini, *supra* note 128 (noting that data synthesis requires "very specific, sophisticated frameworks and metrics that enable it to validate that it created what it set out to create"); *see* Bousquette, *supra* note 231.

and analysts predict a rapid increase in the use of such data.[244] These factors suggest relatively little need for exogenous "innovation mechanisms," such as intellectual property rights, to provide incentives to create. At the margin, however, innovation mechanisms can shore up incentives to develop *high-quality* synthetic data, which may be more expensive to create. This Article contends that the greater value of innovation mechanisms lies not in their classic provisioning function but in the other ways that they shape innovative activity. Accordingly, the next two sections argue that innovation mechanisms should also advance the objectives of technical disclosure and democratization.

B. Disclosure

In addition to encouraging the provisioning of synthetic data, innovation mechanisms should also encourage the disclosure of such data and processes for generating it. As noted, abundant supplies of synthetic data are useless (and potentially extremely harmful) if they are low-quality and unverifiable.[245] The value of independent examination and validation places a premium on the disclosure, sharing, and transparency of synthetic data and synthetic data generators.[246]

Disclosure is particularly important given that AI systems often operate like a "black box" where it is unclear how they arrived at a particular outcome.[247] This lack of transparency may even create due process concerns when government decisions are based on AI.[248] Commentators have advocated for transparency in AI design that satisfies "the need to describe, inspect, and reproduce the mechanisms through which AI systems make decisions and learn to adapt to their environment, and to the governance of the data used or created."[249] Such

---

244 *See* Ramos & Subramanyam, *supra* note 20.

245 *See supra* Part II.C.

246 *Cf.* Gal & Lynskey, *supra* note 26, at 1149 (discussing the legal requirements of explainability and interpretability, which advance norms of transparency and reason giving).

247 Dignum, *supra* note 236, at 5–6; Tucker, Wang, Rotalinti & Myles, *supra* note 74, at 2; *see* Jonathan Zittrain, *The Hidden Costs of Automated Thinking*, New Yorker (July 23, 2019), https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking [https://perma.cc/L7F2-2HQX].

248 *See* Meyers, *supra* note 2, at 21.

249 Dignum, *supra* note 236, at 5.

transparency is key to enhancing the trustworthiness of AI and safeguarding its adoption.[250]

This black box phenomenon also applies to the data used to train AI systems. Given that an AI system is only as good as the data that trains it, there is a pressing need to open up training data for scrutiny.[251] Access to data (real or synthetic) is necessary, for instance, to identify and correct for discriminatory bias.[252] Public access to the data used to train ML systems is especially important when those ML systems make decisions with public policy implications.[253] Data transparency is particularly valuable given that firms sometimes release AI models on an open source basis but do not disclose the data that trained them.[254] As noted, much is at stake in ensuring the transparency and quality of synthetic data. Indeed, one approach to preventing "model collapse" involves widespread coordination of parties developing ML systems to share information on the provenance of training data.[255] Accordingly, to the extent possible, innovation mechanisms should encourage the disclosure of synthetic data.

It is also important to get "under the hood" to examine not only synthetic data, but also the processes used to generate it.[256] According to commentators, "Synthetic data derived from methods without complete documentation cannot be validated, reducing the utility of such methods for the wider scientific community."[257] Greater access to AI models has helped independent parties catch their flaws.[258] In similar fashion, greater

---

250 Meyers, *supra* note 2, at 21.

251 *See* Lemley & Casey, *supra* note 6, at 748.

252 Levendowski, *supra* note 47, at 583.

253 Lemley & Casey, *supra* note 6, at 757; *see* Zittrain, *supra* note 247.

254 Levendowski, *supra* note 47, at 599 ("Several dominant AI players, including Google, IBM, and Microsoft, have released some of their algorithms as open source. Releasing underlying datasets is far less common."). IBM's recent decision to publish the training data for its generative AI systems reflects growing user demand for transparency. Steve Lohr, *IBM Tries to Ease Customers' Qualms About Using Generative A.I.*, N.Y. Times (Sept. 28, 2023), https://www.nytimes.com/2023/09/28/business/ibm-ai-data.html [https://perma.cc/4EUW-3UNK].

255 Shumailov et al., *supra* note 232, at 759.

256 *See* Gal & Lynskey, *supra* note 26, at 1150 (emphasizing accountability in the synthetic data generation process, particularly when explainability or interpretability of synthetic data itself is not feasible).

257 Walonoski et al., *supra* note 69, at 231.

258 Will Douglas Heaven, *The Open-source AI Boom is Built on Big Tech's Handouts. How Long Will It Last?*, MIT Tech. Rev. (May 12, 2023), https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta/ [https://perma.cc/TW6U-4ZNW].

access to processes for generating synthetic data (including AI models and their training data) can help reveal their flaws. For instance, researchers found that a popular synthetic image generator was biased toward producing images of white males because of biases in its training data.[259] This emphasis on disclosure and transparency is consistent with the EU's AI Act and the Biden Administration's voluntary safeguards for AI companies.[260] In sum, to the extent possible, innovation mechanisms should prioritize disclosure of synthetic data and processes to generate it.

C. Democratization

In addition to provisioning and disclosure, innovation mechanisms should also promote "democratization" in the synthetic data landscape. Such democratization has two related components. First, it entails widening access to synthetic data to a broader swath of users. Second and relatedly, democratization also entails increasing the number of independent generators of synthetic data. Pluralizing sources of synthetic data will enhance access to synthetic data itself, and it will also promote innovation, facilitate cross-checking, and counteract data monopolies.

As discussed, among the many concerns raised by AI and ML are anxieties over industry concentration.[261] Large incumbents like Amazon, Facebook, and Google generate (or can purchase) enormous amounts of data to train their ML systems. Such vast stores of data function as a "moat" that raises barriers to entry for new firms seeking to develop ML applications.[262] To be sure, such advantages yield some benefits, as they increase the productivity of leading firms. However, they also inhibit the spread of technical knowledge, slow aggregate productivity growth, and increase wage inequality.[263] Data concentration has even drawn antitrust scrutiny; FTC Chair

---

259 Niharika Jain, Alberto Olmo, Sailik Sengupta, Lydia Manikonda & Subbarao Kambhampati, *Imperfect ImaGANation: Implications of GANs Exacerbating Biases on Facial Data Augmentation and Snapchat Face Lenses*, 304 A.I. 1, 2–4 (2022).

260 Satariano, *supra* note 19; Kevin Roose, *How Do the White House's A.I. Commitments Stack Up?*, N.Y. TIMES (July 22, 2023), https://www.nytimes.com/2023/07/22/technology/ai-regulation-white-house.html [https://perma.cc/QJH5-AS4Q].

261 *See supra* notes 54–61 and accompanying text.

262 Nisselson, *supra* note 46; James Bessen, *The Policy Challenge of Artificial Intelligence*, CPI ANTITRUST CHRONICLE 2, 6 (June 2018).

263 Bessen, *supra* note 262, at 6.

Khan has identified the "vast stores of data" possessed by large incumbents as driving problematic concentration in AI fields.[264]

In light of these concerns, innovation mechanisms should prioritize democratizing access to synthetic data. Synthetic data can counteract this concentrating effect by radically reducing the cost of generating high-quality datasets. In so doing, synthetic data can enable startups and smaller entities to develop their own ML systems to compete against (and perhaps displace) the ML offerings from large data incumbents.[265]

Relatedly, innovation mechanisms should also promote the existence of numerous independent sources of synthetic data. One way to multiply the sources of synthetic data is to empower startups and new entrants that are building ML models to generate their own synthetic data. Another way is to enable the existence of independent, third-party synthetic data generators to serve external clients.[266] Pluralizing the sources of synthetic data offers several benefits. First, competition among generators can reduce the price of and increase access to synthetic data. Second, pluralizing the sources of synthetic data can accelerate innovation in the field, leading to higher-quality synthetic data. Theoretical and empirical accounts suggest that parallel innovation by multiple parties, instead of controlled development by one or a few actors, often leads to the most robust technological advancements.[267] In more practical terms, "AI won't thrive if just a few mega-rich companies get to gatekeep this technology or decide how it is used."[268] Finally, multiple sources of innovation, coupled with disclosure and transparency, can facilitate cross checking and validation of synthetic data and processes for generating it.

---

264 Khan, *supra* note 55. This interest in widening access to data resonates with the "Neo-Brandeisian" movement in antitrust law, which seeks to ensure that small and medium-sized enterprises can viably compete against large incumbents. *See* Jonathan B. Baker, *Finding Common Ground Among Antitrust Reformers*, 84 ANTITRUST L.J. 705, 705–06 (2022).

265 Ramos & Subramanyam, *supra* note 20, at 18 ("Synthetic data democratizes the playing field by allowing smaller organizations to create AI models without a lot of data, effectively solving their *cold-start* problem.").

266 *Id.* at 19.

267 *See, e.g.*, Robert P. Merges & Richard R. Nelson, *On the Complex Economics of Patent Scope*, 90 COLUM. L. REV. 839, 843–44 (1990); *see also* Mark Zuckerberg, Open Source AI is the Path Forward, July 23, 2024, https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/ [https://perma.cc/E9XF-XN5G] ("Lots of people see that open source is advancing at a faster rate than closed models, and they want to build their systems on the architecture that will give them the greatest advantage long term.").

268 Heaven, *supra* note 258.

IV
Innovation Mechanisms for Developing Synthetic Data

Building on the previous normative analysis, this Part assesses how various "innovation mechanisms" can promote the provisioning, disclosure, and democratization of high-quality synthetic data. It first explores open source production before turning to several proprietary mechanisms based on intellectual property rights: patents, trade secrets, and copyrights. Throughout, it analyzes how these innovation mechanisms can promote provisioning, disclosure, and democratization, and it proposes policy reforms to help them advance these objectives.

A. Nonproprietary and Open Source Approaches

1. *Overview*

While "innovation mechanisms" conjures up notions of intellectual property rights, numerous innovation mechanisms other than exclusive rights can promote the generation of synthetic data.[269] Focusing first on public approaches, government agencies could directly fund the development of synthetic data, as they do for other public goods, such as basic scientific research.[270] Indeed, various federal agencies already fund research and development efforts focused on synthetic data.[271] The government could also subsidize the development of synthetic data through tax breaks or by offering prizes.[272] Such nonproprietary innovation mechanisms could encourage the development of this valuable information asset without directly subjecting it to exclusive rights.

Turning from public to private approaches, numerous innovation mechanisms incentivize for-profit entities to develop synthetic data without recourse to intellectual property rights.[273]

---

269 *See generally* Daniel J. Hemel & Lisa Larrimore Ouellette, *Innovation Policy Pluralism*, 128 Yale L.J. 544, 551–58 (2019).

270 *See* Lemley, *supra* note 241, at 1050.

271 *See, e.g.*, *Simulated and Synthetic Data for Infrastructure Modeling (SSDIM)*, U.S. Nat'l Sci. Found. (Mar. 30, 2017), https://new.nsf.gov/funding/opportunities/simulated-synthetic-data-infrastructure-modeling [https://perma.cc/3JTU-9ERB]; *Researchers Receive $1.2 Million NIH Grant to Study Synthetic Data Use in Health Care*, UC Davis Health (Apr. 27, 2022), https://health.ucdavis.edu/health-magazine/issues/fall2022/noteworthy/study-synthetic-data-use.html#:~:text=This%20spring%2C%20UC%20Davis%20researchers,predict%2C%20diagnose%20and%20treat%20diseases [https://perma.cc/W223-K2P3].

272 *See* Hemel & Ouellette, *supra* note 269, at 551–52.

273 In this context, "innovation mechanisms" include economist David Teece's concept of "regimes of appropriability," which allow an innovator to "capture the

First and perhaps most importantly, firms' own desire to develop new ML systems can motivate significant investments in generating synthetic data to train them. This is a variant of so-called user innovation, in which a party's own use for a resource provides the incentive to develop it, even absent formal property rights.[274] Relatedly, a firm may pursue vertical integration by combining the "upstream" generation of synthetic data with the "downstream" training of ML models with that data.[275] Thus, for instance, large incumbents like Facebook and Google may invest in synthetic data to train their consumer-facing, profit-generating ML systems. Turning to (nonintegrated) firms that may sell synthetic data to outside parties, first-mover advantage (perhaps accompanied by brand recognition) can allow them to appropriate returns from investing in synthetic data.[276] Notably, classic empirical research suggests that non-IP innovation mechanisms are more significant than exclusive rights for promoting innovation in most contexts.[277]

Approaches based on open source software have been particularly important for generating synthetic data. Open source software refers to software distributed with its source code and subject to licenses in which the copyright holder grants subsequent users rights to use, modify, and distribute such software.[278]

---

profits generated by an innovation." David J. Teece, *Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy*, 15 RES. POL'Y 285, 287 (1986). This Article uses the more expansive term "innovation mechanisms" to also include approaches where innovators may not seek to internalize profits, such as with open source approaches.

274 *See generally* Katherine J. Strandburg, *Users as Innovators: Implications for Patent Doctrine*, 79 U. COLO. L. REV. 467 (2008) (exploring the phenomenon of user innovation).

275 *See generally* Teece, *supra* note 273, at 300 (providing an example of vertical integration as an appropriability regime); *see also* JONATHAN M. BARNETT, INNOVATORS, FIRMS, AND MARKETS 3 (2021) (noting the advantage of vertically integrated incumbents in appropriating returns from innovation in the absence of strong intellectual property protection).

276 *Cf.* Stuart J.H. Graham, Robert P. Merges, Pam Samuelson & Ted Sichelman, *High Technology Entrepreneurs and the Patent System: Results of the 2008 Berkeley Patent Survey*, 24 BERKELEY TECH. L.J. 1255, 1289 (2009) (finding that first-mover advantage was the most important appropriability mechanism for startups).

277 *See generally* Richard C. Levin et al., *Appropriating the Returns from Industrial Research and Development*, 18 BROOKINGS PAPERS ON ECON. ACTIVITY 783 (1987); Wesley M. Cohen, Richard R. Nelson & John P. Walsh, *Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (Or Not)* (Nat'l Bureau of Econ. Rsch., Working Paper No. 7552, 2000), http://ssrn.com/abstract=214952 [https://perma.cc/H4AT-XGUV].

278 *See generally* CHRISTOPHER M. KELTY, TWO BITS: THE CULTURAL SIGNIFICANCE OF FREE SOFTWARE (2008) (discussing free and open source software).

As has been widely studied, open source software facilitates commons-based "peer production" in which large numbers of unconnected programmers contribute to massive, collective software projects.[279] Prominent examples of open source software products include the Apache HTTP Server and Mozilla Firefox web browser.[280] Although nominally based on copyright, the legal openness of open source software challenges the notion—central to intellectual property law—that exclusive rights are necessary to motivate investments in information goods. Notably, while open source software is considered nonproprietary, many companies have effectively monetized such software, such as by providing service and support for open source software implementations. For example, Red Hat (which was acquired by IBM) provides open source software products and charges fees for support, training, and integration.[281] Recently, Meta embraced a strategy of open source AI models, noting that it would lead to more robust, secure, and customizable AI development.[282]

Numerous government, academic, and nonprofit initiatives have developed open source synthetic data generators.[283] In the healthcare sector, a consortium of nonprofit and academic researchers created Synthea, an "open-source synthetic health simulation . . . that simulates synthetic patients from cradle to grave."[284] Synthea has generated a million synthetic medical records for fictitious patients in a virtual Commonwealth of Massachusetts.[285] The developers of Synthea have made these

---

279 YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM 5, 320–23 (2006) (examining commons-based peer production in several domains, including open source software).

280 Greg R. Vetter, *Commercial Free and Open Source Software: Knowledge Production, Hybrid Appropriability, and Patents*, 77 FORDHAM L. REV. 2087, 2111 (2009).

281 David L. Olson, Bjorn Johansson & Rogerio Atem De Carvalho, *Open Source ERP Business Model Framework*, 50 ROBOTICS & COMPUTER-INTEGRATED MFG. 30, 32 (2018).

282 Zuckerberg, *supra* note 267; Mike Isaac, *Mark Zuckerberg Stumps for 'Open Source' A.I.*, N.Y. TIMES (July 23, 2024), https://www.nytimes.com/2024/07/23/technology/mark-zuckerberg-meta-open-source-ai.html [https://perma.cc/X5TZ-3MGA].

283 *See* Evgeniya Panova, *Synthetic Data Tools: Open Source or Commercial? A Guide to Building vs. Buying*, STATICE.AI (Sept. 23, 2023), https://www.statice.ai/post/synthetic-data-open-source-tools-guide-building-buying [https://perma.cc/BFV5-JTJM] (listing almost two dozen open source synthetic data tools from various sectors).

284 Walonoski et al., *supra* note 69, at 232. Other synthetic medical data generators include the Synthetic Electronic Medical Records Generator (EMERGE) and the medical Generative Adversarial Network (medGAN). *Id.* at 231.

285 *Id.* at 232.

records publicly available for public and private users "free of legal, privacy, security, financial, and intellectual property restrictions."[286] Another prominent example is MIT's Synthetic Data Vault (SDV), a set of open source synthetic data generation tools unveiled in 2020.[287] The SDV represents a "one-stop shop where users can get as much data as they need for their projects, in formats from tables to time series."[288] It represents the largest open source ecosystem for synthetic data.[289]

Additionally, for-profit firms have also pursued open source synthetic data generation. Microsoft has partnered with Harvard University to provide an open source synthetic data generator aimed at enhancing data privacy.[290] In addition to large incumbents, numerous startups offer open source synthetic data, sometimes on the Red Hat model of profiting off of customization, service, and support. For instance, Gretel.ai, which recently partnered with Google Cloud, generates "anonymized, safe-to-share, and privacy-first synthetic data."[291]

2. *Analysis and Prescriptions*

Returning to the normative objectives discussed above, open source synthetic data generation offers the best of many worlds. First, it addresses the provisioning problem of creating valuable information goods that are costly to develop but cheap to copy. Open source approaches marshal the provisioning power of communal peer production to generate synthetic data without subjecting it to exclusive rights. While it is perhaps unsurprising that government, academic, and nonprofit entities have embraced such approaches, many for-profit firms—including both large incumbents and small startups—have also invested considerably in open source synthetic data generation.

---

286 *Id.*

287 LIDS, *supra* note 155.

288 *Id.*

289 Toews, *supra* note 31.

290 Andreas Kopp, *Create Privacy-preserving Synthetic Data for Machine Learning with SmartNoise*, MICROSOFT (Feb. 18, 2021), https://opensource.microsoft.com/blog/2021/02/18/create-privacy-preserving-synthetic-data-for-machine-learning-with-smartnoise/ [https://perma.cc/SW29-8UCE].

291 *Gretel Partners with Google Cloud to Harness the Power of Synthetic Data and Accelerate Adoption of Safer Generative AI in the Enterprise*, BUSINESS WIRE (Mar. 14, 2023, 9:01 AM), https://www.businesswire.com/news/home/20230314005528/en/Gretel-Partners-With-Google-Cloud-to-Harness-the-Power-of-Synthetic-Data-and-Accelerate-Adoption-of-Safer-Generative-AI-in-the-Enterprise [https://perma.cc/2MN6-GXH8]; *see* Panova, *supra* note 283.

Second and perhaps more importantly, open source approaches fully disclose synthetic data and processes for generating it. Such transparency counteracts the black box character of synthetic training data and facilitates its evaluation and validation.[292] Reflecting the notion that "[g]iven enough eyeballs, all bugs are shallow," open source synthetic data generation allows anyone to correct and improve upon existing source code.[293] Addressing the general advantages of open source, Mark Zuckerberg recently noted that "open source should be significantly safer since the systems are more transparent and can be widely recognized."[294] For instance, one of the benefits of Synthea, the open source generator of synthetic medical data, is that it "can be easily inspected, modified, and refined, facilitating transparency and continuous improvement."[295] Indeed, because it is open source, "it is continually being tweaked by researchers to create more-accurate disease models."[296]

It is important to note that while disclosure and wide access are generally seen as virtues of open source, they may have some drawbacks. For instance, some firms are releasing their AI models in a "controlled way according to their potential risk of causing harm or being misused."[297] The same may be true for open source synthetic data and data generators, which could be used for nefarious purposes.

Third, open source approaches also promote democratic access to synthetic data and tools for generating it.[298] As this Article has shown, gathering sufficient real-world data to train an ML model is expensive and time consuming. Open source synthetic data libraries and tools mitigate these constraints by offering free or low-cost access to large amounts of data and pretrained models that would be difficult to build from scratch.[299] Open source also means that researchers and developers "can

---

292 *See supra* notes 251–60 and accompanying text; Alex Engler, *How Open-source Software Shapes AI Policy*, BROOKINGS (Aug. 10, 2021), https://www.brookings.edu/research/how-open-source-software-shapes-ai-policy [https://perma.cc/J8FU-E3Z6].

293 ERIC S. RAYMOND, THE CATHEDRAL AND THE BAZAAR 29 (1999).

294 Zuckerberg, *supra* note 267.

295 Walonoski et al., *supra* note 69, at 231.

296 Lieber, *supra* note 85.

297 Heaven, *supra* note 258.

298 *Cf.* Zuckerberg, *supra* note 267 ("Open source will ensure that more people around the world have access to the benefits and opportunities of AI, that power isn't concentrated in the hands of a small number of companies, and that the technology can be deployed more evenly and safely across society.").

299 Heaven, *supra* note 258.

study, build on, and modify" these models.[300] Such openness democratizes synthetic data in two ways. First, it provides access to synthetic data to under-resourced entities (including startups) to develop and train ML systems. Second, it enables standalone data vendors that can utilize open source models to produce synthetic data for external clients.

Given the substantial benefits of open source synthetic data generation, this Article offers several prescriptions to promote this practice. First and most obviously, the government can expand its funding to specifically support open source synthetic data generation. Second and relatedly, government, academic, and nonprofit entities can catalyze open source synthetic data generation by providing the necessary infrastructure to support it. In the context of the Human Genome Project, NIH promoted rapid disclosure of DNA sequence data by both enacting data-disclosure policies and maintaining a central repository (GenBank) in which researchers could deposit their data.[301] Similarly, funding agencies, universities, and nonprofits can supply necessary infrastructure by hosting open source synthetic data generation projects, as MIT did with the SDV. Third, the government can indirectly encourage the adoption of open source synthetic data by modifying its data sharing requirements for taxpayer-funded research. Both the National Science Foundation (NSF) and NIH require recipients of research grants to make data arising from taxpayer-funded research available to other researchers.[302] These agencies should specify that such data sharing requirements also apply to synthetic data. Furthermore, given the unique nature of synthetic data, these policies should require grant recipients to disclose not just synthetic data but also methods used to generate it.

The government can also promote open source synthetic data through its procurement powers. The government is a major purchaser of technology, and government procurement has accelerated the development of many technological

---

300 *Id.*

301 *See* Peter Lee, *Centralization, Fragmentation, and Replication in the Genomic Data Commons*, *in* GOVERNING MEDICAL KNOWLEDGE COMMONS 46, 49–50 (Katherine J. Strandburg et al. eds., 2017).

302 *See Preparing Your Data Management and Sharing Plan*, U.S. NAT'L SCI. FOUND., https://new.nsf.gov/funding/data-management-plan#nsfs-data-sharing-policy-1c8 [https://perma.cc/6FTU-Z2LW]; (last visited Nov. 17, 2024), *Final NIH Policy for Data Management and Sharing*, NAT'L INST. OF HEALTH (Oct. 29, 2020), https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-013.html [https://perma.cc/6LXZ-C5BL].

industries.[303] The government will likely increase its procurement of synthetic data and ML systems trained on such data, and it can use the power of the purse to compel contractors to generate synthetic data in an open source manner. Notably, the Federal Acquisition Regulations (FAR) provide, with some exceptions, the federal government with "unlimited rights" in data first produced under subject contracts and data delivered under subject contracts.[304] Policymakers should clarify that such regulations also apply to synthetic data (whether open source or proprietary) generated under a federal contract. More generally, government procurement can widen access to data even if it was not originally generated in an open source manner. Federal regulations enable the government "to use, disclose, reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display [data], in any manner and for any purpose, and to have or permit others to do so."[305] These regulations allow the government to handle data in an open source manner even if it is not technically open source.

While open source initiatives are an important innovation mechanism to encourage the creation of synthetic data, the remainder of this Part focuses on proprietary mechanisms based on intellectual property rights. Under the traditional view, the grant of exclusive rights over an information good excludes free riders, thus maintaining incentives to create. In the context of synthetic data, however, the provisioning function of patents, trade secrets, and copyrights may not be as important as their ability to promote disclosure and democratization in synthetic data generation.

B. Patents

1. *Overview*

Patents are a classic innovation mechanism that can encourage the development of synthetic data-related technologies. Patents confer twenty years of exclusive rights over novel, useful, and nonobvious inventions.[306] This section evaluates the

---

303 *See* Peter Lee, *Enhancing the Innovative Capacity of Venture Capital*, 24 YALE J.L. & TECH. 611, 696 (2022) (discussing the role of government procurement in jumpstarting the nuclear power, computer, semiconductor, and aerospace industries).

304 48 C.F.R. §§ 52.227-14(b)(1) (2023).

305 48 C.F.R. §§ 52.227-14(a) (2023).

306 35 U.S.C. §§ 101, 102, 103, 154(a)(2).

patentability of synthetic data, concluding that while synthetic data itself is not patentable, processes for generating it generally are. It then examines how patents can advance the provisioning, disclosure, and democratization of synthetic data, and it suggests doctrinal reforms to improve their ability to do so.

At the outset, it is highly doubtful that synthetic data itself is patentable. Among other obstacles, synthetic data does not comprise patentable subject matter, which the patent statute defines as processes, machines, manufactures, or compositions of matter.[307] Synthetic numbers, text, and other data are clearly not processes, machines, or compositions of matter. Notwithstanding courts' broad interpretations of the term "manufacture,"[308] not all things that are made by people (such as poems) are "manufactures." Relatedly, the Court of Appeals for the Federal Circuit has ruled that all categories of patentable subject matter other than processes "must exist in some physical or tangible form."[309] This would exclude data itself (either real-world or synthetic) from eligibility for patenting as a "manufacture." Relatedly, the court has ruled that "[d]ata in its ethereal, non-physical form is simply information that does not fall under any of the categories of eligible subject matter under 101."[310]

Beyond patentable subject matter, the inventorship requirement presents another obstacle to patenting synthetic data. Under U.S. patent law, whoever "invents" a technology may obtain a patent.[311] Accordingly, U.S. patent applications must list the "true and only" inventors of a claimed technology.[312] As noted, synthetic data is often produced by AI systems. However, the U.S. Patent and Trademark Office (USPTO) has roundly rejected the inventorship status of AI.[313] The Federal Circuit has followed suit, categorically ruling that only natural persons can

---

307 35 U.S.C. § 101.

308 Diamond v. Chakrabarty, 447 U.S. 303, 308–09 (1980).

309 Digitech Image Tech's, LLC v. Elecs. for Imaging, Inc., 758 F.3d 1344, 1348 (Fed. Cir. 2014).

310 *Id.* at 1350; *see also* In re Nuijten, 500 F.3d 1346, 1357 (Fed Cir. 2007) (holding that even the physical embodiment of data in a signal does not comprise patentable subject matter).

311 35 U.S.C. § 101.

312 Hess v. Advanced Cardiovascular Sys., Inc., 106 F.3d 976, 979–80 (Fed. Cir. 1997).

313 *See* Thaler v. Vidal, 43 F.4th 1207, 1209–10 (Fed. Cir. 2022); In re Application of Application No.: 16/524,350, Dec. Comm'r Pat. (July 20, 2020) (rejecting patent applications listing an AI system as the inventor).

qualify as "inventors" under the Patent Act.[314] The USPTO recently launched a "listening tour" to solicit input on the patentability of AI-generated inventions, and it is possible that the legal landscape may change.[315] For present purposes, however, even if synthetic data comprised patentable subject matter, it would not be patentable if it were solely produced by AI.

While synthetic data itself is not patentable, processes for generating synthetic data likely are. This would include AI systems designed (by humans) to generate synthetic data. Thus, for instance, advancements in generative adversarial networks (GANs) that produced higher-quality synthetic images could be patented.[316] As noted, processes are a recognized category of patentable subject matter,[317] and they need not have any physical or tangible element.[318] In recent years, courts have taken a narrower approach to the patent eligibility of processes, particularly those manifesting in software.[319] However, these cases typically involved inventions that merely adapted existing ideas to a software or online environment. Software that enables new functionality remains patentable subject matter.[320]

---

314 *Thaler*, 43 F.4th at 1210.

315 *See Request for Comments on Artificial Intelligence and Inventorship*, 88 FED. REG. 9492 (Feb. 14, 2023).

316 While this section focuses on processes to generate synthetic data, a host of other processes related to synthetic data are eligible for patenting, such as: methods for enriching, anonymizing, or representing data; processes for training an ML model using synthetic data; and methods to test, evaluate, and validate synthetic data. Joshua D. Berk, Lily Zhang & Terri Shieh-Newton, *AI in Biotech and Synthetic Biology: What Can Be Protected? What Should Be Kept Secret?*, NAT'L L.J. (Aug. 11, 2021), https://www.natlawreview.com/article/ai-biotech-and-synthetic-biology-what-can-be-protected-what-should-be-kept-secret [https://perma.cc/2SU4-QV8Q].

317 35 U.S.C. § 101.

318 Digitech Image Tech's, LLC v. Elecs. for Imaging, Inc., 758 F.3d 1344, 1348 (Fed. Cir. 2014).

319 Courts have established a two-part patent eligibility test that asks 1) if a claim is directed to a patent-ineligible concept, and 2) if so, if the claim includes an "inventive concept" that differentiates it from merely covering the patent-ineligible concept. *See* Alice Corp. v. CLS Bank Int'l, 573 U.S. 208, 217 (2014) (citing Mayo Collaborative Servs. v. Prometheus Lab'ys. Inc., 566 U.S. 66, 77–79 (2012)). Courts have applied this framework to reject the patent eligibility of several software-based processes by reasoning that they merely claim abstract ideas. *See id.* at 212 (holding a computer-implemented process of intermediated settlement ineligible for patenting); Intell. Ventures I LLC v. Cap. One Bank (USA), 792 F.3d 1363, 1365 (Fed. Cir. 2015) (affirming the ineligibility of two patents directed to internet-based activities).

320 *See, e.g.*, McRO, Inc. v. Bandai Namco Games A., Inc., 837 F.3d 1299, 1302–03 (Fed. Cir. 2016) (holding that software-based methods for automatically animating characters comprised patentable subject matter); Enfish, LLC v. Microsoft Corp., 822 F.3d 1327, 1336 (Fed. Cir. 2016) (finding that claims directed

Assuming that the other requirements for patentability, such as utility, novelty, nonobviousness, enablement, and written description,[321] were satisfied, processes for generating synthetic data would be patentable.

Indeed, numerous firms have patented processes for synthesizing data. They include large incumbents operating in both technological (e.g., Microsoft[322]) and non-technological (e.g., Capital One Services[323]) industries. Currently, the top five patentees in the synthetic data space are IBM, Microsoft, Baidu, Alphabet, and Meta Platforms.[324] In addition, many startups have patented processes for generating synthetic data. For example, synthetic visual data firm Synthesis AI has four U.S. patents.[325] MDClone also has four U.S. patents as well as numerous published patent applications.[326] Howso, an AI platform that provides synthetic data as well as several other services, has three patents in their name as well as over sixty "patent assets."[327] While firms patent for a variety of reasons, this practice suggests that at least some firms view patents as important for protecting investments in generating synthetic data.

2. *Analysis and Prescriptions*

Returning to the three normative objectives above, patents can do much to promote provisioning, disclosure, and democratization, though certain doctrinal reforms are warranted. First, patents are a classic "provisioning" mechanism

---

to a self-referential database were eligible for patenting and did not merely cover abstract ideas).

321 *See* 3 U.S.C. §§ 101, 102, 103, 112.

322 *See* U.S. Patent No. 11,580,329 B2 (issued Feb. 14, 2023).

323 *See* U.S. Patent No. 10,884,894 B2 (issued Jan. 5, 2021).

324 GlobalData, *Artificial Intelligence Innovation: Leading Companies in Synthetic Data*, VERDICT (June 2, 2023), https://www.verdict.co.uk/innovators-ai-synthetic-data-technology/#catfish [https://perma.cc/WL5H-U9SF].

325 *Patent Public Search Basic (PPUBS Basic)*, U.S. PAT. & TRADEMARK OFF., https://ppubs.uspto.gov/pubwebapp/static/pages/ppubsbasic.html [https://perma.cc/NP3E-UNRF] (search for "(Synthesis).aanm. AND (AI).aanm,").

326 *Patent Public Search Basic (PPUBS Basic)*, U.S. PAT. & TRADEMARK OFF., https://ppubs.uspto.gov/pubwebapp/static/pages/ppubsbasic.html [https://perma.cc/A4YA-FMY4] (search for "((mdclone).aanm. OR (mdclone).as"); MDCLONE, https://www.mdclone.com/services/synthetic-data/ [https://perma.cc/SR6M-9JK9].

327 *Patent Public Search Basic (PPUBS Basic)*, U.S. PAT. & TRADEMARK OFF., https://ppubs.uspto.gov/pubwebapp/static/pages/ppubsbasic.html [https://perma.cc/32PM-6CBZ]; (search for "Howso") About Us, HOWSO, https://www.howso.com/about-us/ [https://perma.cc/6CZW-3BTT].

that can encourage parties to develop new processes for generating synthetic data. By excluding free riders who would copy such processes for free, patents shore up incentives to invent. Furthermore, by allowing inventors to internalize a greater share of the value of their creations, patents may encourage inventors to develop synthetic data generators with higher functionality than open source varieties. For instance, startup Mostly AI argues that its proprietary synthetic data generator produces higher-quality data than open source tools from the SDV.[328] While such self-interested statements must be scrutinized carefully, there are theoretical reasons to posit that proprietary approaches can encourage greater investment in synthetic data generation, thus yielding functionally superior results. Patents represent a valuable complement to open source and other innovation mechanisms in a diversified ecosystem for generating synthetic data.

Of course, the provisioning benefits of patents on processes for generating synthetic data must be weighed against the familiar costs of intellectual property protection. Exclusive rights diminish static efficiency because they subject nonrivalrous information goods (such as processes to generate synthetic data) to artificial scarcity.[329] Such losses are ordinarily justified by patents' presumptive contributions to dynamic efficiency by shoring up incentives to invent.[330] However, a wide literature has explored how patents can harm dynamic efficiency by raising the cost of cumulative innovation. To the extent that innovators must gather together or build off of patented technologies to further innovate, such patents may create "anticommons" or thickets that inhibit technological progress.[331] While there is no empirical evidence of such innovation-inhibiting effects

---

328 Tobias Hann, *SDV vs MOSTLY AI: Which Synthetic Data is Better?*, MOSTLY AI (Aug. 19, 2022), https://mostly.ai/blog/sdv-vs-mostly-ai-synthetic-data-generators-comparison#:~:text=To%20assess%20the%20quality%20of,MOSTLY%20AI's%20Synthetic%20Data%20Platform [https://perma.cc/J9SC-57QB]; Michael Platzer & Thomas Reutterer, *Holdout-Based Empirical Assessment of Mixed-Type Synthetic Data*, 4 FRONTIERS OF BIG DATA 1, 7, 12 (2021) (finding that Mostly AI and one other synthetic generator achieved the highest fidelity scores among several tested generators but acknowledging a potential conflict of interest because one of the coauthors is a cofounder of Mostly AI).

329 *See* Thomas Cheng, *Putting Innovation Incentives Back in the Patent-Antitrust Interface*, 11 NW. J.L. & TECH. & INTELL. PROP. 385, 388–90 (2013).

330 *Id.*

331 *See* Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCI. 698 (1998); Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, *in* 1 INNOVATION POL'Y & ECON. 119 (2001).

in the context of synthetic data, policymakers should closely monitor the uptick in patents on synthetic data generation.

Second, and perhaps more importantly, patents can promote the disclosure of processes for generating synthetic data. Unlike open source approaches, patents are "closed," proprietary innovation mechanisms, which suggests little emphasis on openness and transparency. However, the patent system embodies a societal quid pro quo in which inventors must disclose their inventions to receive exclusive rights.[332] The disclosure requirements of patentability fall under 35 U.S.C. § 112, which requires that inventors enable their inventions, adequately describe them, and disclose any "best mode" they are aware of for practicing them.[333] Enablement is particularly important for disclosure, as it requires that a patent teach a person of ordinary skill in the art how to make and use a claimed invention.[334] For instance, a patent on a process for generating synthetic data should disclose that invention in sufficient detail so that technical artisans can practice it without undue experimentation. Such disclosure would be highly valuable in countering the black box quality of AI generally and synthetic data specifically.

In a valuable development, the Supreme Court recently heightened the enablement requirement. In its 2023 decision in *Amgen v. Sanofi*, the Court rejected Amgen's attempt to patent entire classes of antibodies defined by their function, holding that Amgen's patent failed to enable the full range of claimed inventions.[335] It ruled that a patent must do more than simply provide "research assignments" to adequately enable all embodiments in a claim.[336] Applied to the present context, the Court's ruling should increase disclosure of patented processes for generating synthetic data, particularly processes claimed at a high level of generality. Commentators note that "it may be necessary under *Amgen* to disclose how training of the AI was performed, the training data sets used, and the different weights applied to the data within the data sets to enable a patent implementing an AI system."[337]

---

332 Universal Oil Prods. Co. v. Globe Oil & Refining Co., 322 U.S. 471, 484 (1944); Amgen Inc. v. Sanofi, 598 U.S. 594, 605 (2023).

333 35 U.S.C. § 112.

334 *Id.*

335 *Sanofi*, 598 U.S. at 614.

336 *Id.*

337 David McCombs, Dina Blikshteyn, Eugene Goryunov & Matthew Beck, *Navigating the Murky Waters of Patent Claims Involving AI After* Amgen v. Sanofi,

More broadly, while AI models that generate synthetic data may operate like "black boxes," the inscrutability of such inventions does not preclude enablement. Drawing an analogy to patented biological innovations whose innerworkings are not understood, legal scholar Dan Burk argues that enabling AI models may simply require public deposit of the models, including in some cases the data used to train them.[338] Such public deposit, moreover, would greatly facilitate cross checking and validation of such patented processes.

While strengthening the enablement requirement is useful, this Article argues for rehabilitating the best mode requirement to further increase the disclosure of processes for generating synthetic data.[339] Notably, courts and commentators have criticized the patent disclosure requirements as limited and easily circumvented.[340] To help remedy this state of affairs, this Article argues for shoring up the best mode requirement. As noted, the patent statute formally requires patent applicants to disclose any known best mode, which refers to any "specific instrumentalities or techniques" known to the applicant as the best way to practice an invention.[341] The best mode requirement demands more disclosure than enablement, which only mandates disclosing enough information to practice a basic version of an invention.[342] Due to legislative reforms in 2011, the best mode requirement is largely unenforced.[343] The result is that patentees routinely obtain exclusive rights on their inventions while keeping valuable technical information about

---

DRUG DISCOVERY ONLINE (Aug. 24, 2023), https://www.drugdiscoveryonline.com/doc/navigating-the-murky-waters-of-patent-claims-involving-ai-after-amgen-v-sanofi-0001 [https://perma.cc/XX8X-GBQR].

338 Dan L. Burk, *AI Patents and the Self-Assembling Machine*, 105 MINN. L. REV. HEADNOTES 301, 313–14 (2021).

339 *See generally* Peter Lee, *Best Mode*, ELGAR ENCY. OF INTELL. PROP. (forthcoming 2025).

340 *See* Brenner v. Manson, 383 U.S. 519, 534 (1966) (acknowledging "the highly developed art of drafting patent claims so that they disclose as little useful information as possible—while broadening the scope of the claim as widely as possible"); Sean B. Seymore, *The Teaching Function of Patents*, 85 NOTRE DAME L. REV. 621, 634–36 (2010); Jeanne C. Fromer, *Patent Disclosure*, 94 IOWA L. REV. 539, 552–553 (2009).

341 Spectra-Physics, Inc. v. Coherent Inc., 827 F.2d 1524, 1532 (Fed. Cir. 1987).

342 35 U.S.C. § 112; David S. Levine & Joshua D. Sarnoff, *Compelling Trade Secret Sharing*, 74 HASTINGS L.J. 987, 1013–14 (2023).

343 Brian J. Love & Christopher B. Seaman, *Best Mode Trade Secrets*, 15 YALE J.L. & TECH. 1, 8–9 (2012). The Leahy-Smith America Invents Act significantly weakened the best mode requirement by establishing that noncompliance with the requirement is no longer a ground for cancelling, invalidating, or rendering unenforceable a patent. *See* 35 U.S.C. § 282(b)(3)(A).

the best way to practice them to themselves. Such concealment offends the basic quid pro quo of the patent system.[344] As advocated elsewhere, this Article argues for restoring the best mode requirement as a fully enforceable requirement of patentability.[345] While this reform would pay dividends across all types of inventions, it would be especially useful for increasing disclosure of processes for generating synthetic data.

Third, patents can play a surprising role in democratizing access to synthetic data. This is somewhat ironic given that such proprietary innovation mechanisms seem the exact opposite of open source approaches that are freely available to all. However, patents can promote democratization by enabling the existence of standalone synthetic data generators. A wide literature has argued that patents enable the existence of small, research-based technology firms that rely on exclusive rights to protect technological outputs.[346] A classic—though contested—example is small, research-based biotechnology firms, which patent synthesized compounds and license them to large pharmaceutical firms for commercialization.[347] In the absence of patent protection, such firms would likely not exist. Rather, these "upstream" entities would be vertically integrated into larger firms that also perform "downstream" commercialization.

As applied in the present context, patents on processes to generate synthetic data enable the independent existence of standalone firms that rely on patented methods to synthesize data. In this vein, patents can diversify the ecosystem of actors generating such data. Indeed, synthetic data startup Synthesis AI, which holds several patents, has stated that one of its aims is to democratize access to data, enabling a wide range of entities to

---

344 Love & Seaman, *supra* note 343, at 3 ("Traditionally, trade secrecy and patent rights have been considered mutually exclusive.").

345 *See, e.g.*, Peter Lee, *New and Heightened Public-Private Quid Pro Quos: Leveraging Public Support to Enhance Private Technical Disclosure*, *in* INTELLECTUAL PROPERTY, COVID-19, AND THE NEXT PANDEMIC: DIAGNOSING PROBLEMS, DEVELOPING CURES 39, 48–52 (Madhavi Sunder & Haochen Sun eds., 2024).

346 *See* Ashish Arora & Robert P. Merges, *Specialized Supply Firms, Property Rights and Firm Boundaries*, 13 INDUS. & CORP. CHANGE 451, 455 (2004); Jonathan M. Barnett, *Intellectual Property as a Law of Organization*, 84 S. CAL. L. REV. 785, 838–39 (2011); Bronwyn H. Hall & Rosemarie Ham Ziedonis, *The Patent Paradox Revisited: An Empirical Study of Patenting in the U.S. Semiconductor Industry, 1979–1995*, 32 RAND J. ECON. 101, 119–20 (2001).

347 *See* Peter Lee, *Innovation and the Firm: A New Synthesis*, 70 STAN. L. REV. 1431 (2018) (challenging the notion that patents promote vertical disintegration when patented technologies require significant tacit knowledge to transfer and practice).

develop ML systems.[348] The entry of such startups helps counteract the dominance of large incumbents such as Facebook and Google that have ready access to vast stores of data.

What patents give with one hand, however, they take away with the other. While patents can promote industry diversification when wielded by startups and new entrants, they can exacerbate industry concentration when wielded by large incumbents.[349] Thus, for instance, if Google or Facebook obtained large numbers of patents on processes to generate synthetic data, such patents could block new entrants and accelerate concentration in ML fields.[350] More broadly, a proliferation of patents in the synthetic data space could raise barriers to entry by creating innovation-dampening anticommons and thickets.[351] Given that the welfare effects of a patent depend considerably on the kind of entity wielding it, this Article argues for a differential patent policy that favors patent acquisition by small entities over large ones.[352] Lower patent fees for small entities, coupled with aggressive application of patent misuse and antitrust doctrines to curb exclusionary practices by patentees, could help patents promote startup formation while limiting their ability to stifle new entry.[353]

C. Trade Secrets

1. *Overview*

Trade secrets represent another intellectual property regime that can encourage the development of synthetic data.[354] Trade secrecy arises from state and federal laws that protect technical

---

348 Wiggers, *supra* note 219 (describing Synthesis AI as being focused on, among other objectives, "democratizing access" to data).

349 *See* Peter Lee, *Reconceptualizing the Role of Intellectual Property Rights in Shaping Industry Structure*, 72 VAND. L. REV. 1197, 1201–02 (2019).

350 *Cf.* Brenda M. Simon & Ted Sichelman, *Data-Generating Patents*, 111 NW. L. REV. 377, 379 (2017) (noting that so-called "data-generating patents" allow patentees to enjoy exclusivity in large amounts of data, which may have pernicious social consequences).

351 *See* Heller & Eisenberg, *supra* note 331; Shapiro, *supra* note 331.

352 *See* Peter Lee, *Churn*, 99 WASH. U. L. REV. 1, 53–62 (2021) [hereinafter Lee, *Churn*].

353 *See id.*; U.S. DEPT. OF JUST. & FED. TRADE COMM'N, DRAFT MERGER GUIDELINES 19–20 (2023) (emphasizing that mergers should not extend a dominant position, such as by exacerbating barriers to entry due to control over patents).

354 *See* Kewanee Oil Co. v. Bicron Corp., 416 U.S. 470, 484–85 (1974) (recognizing that trade secrets provide an "incentive to invention"); Mark A. Lemley, *The Surprising Virtues of Treating Trade Secrets as IP Rights*, 61 STAN. L. REV. 311, 329–32 (2008) (applying the incentive to invent justification for exclusive rights to trade secrets) [hereinafter Lemley, *Trade Secrets*].

and business information that is the subject of reasonable efforts to maintain secrecy and that derives economic value from such secrecy.[355] Protectable subject matter for trade secrets is very broad, encompassing "virtually any useful information."[356] This would include both technical and nontechnical information as well as source code and raw data.[357] The threshold for conferring economic value is quite low; a trade secret's economic value need only be "actual or potential" to qualify.[358] Unlike patents, which require lengthy applications and robust technical disclosure,[359] trade secrets are automatically protected without application, registration, or public disclosure.[360] Furthermore, while patents last for twenty years from the date of filing an application, trade secrets last indefinitely—as long as information remains secret and commercially valuable.[361]

While attractive to innovators in many ways, trade secrecy is an inherently "leaky" regime that does not confer strict exclusive rights. A trade secret loses protection as soon as it is disclosed.[362] Furthermore, trade secrecy only protects against misappropriation of such information, such as through breach of a confidential duty or improper means.[363] Independent invention or reverse engineering of a trade secret does not constitute misappropriation and could lead to the termination of trade secret protection.[364]

---

355 *See* Unif. Trade Secrets Act § 1(4) (2018); Defend Trade Secrets Act, 18 U.S.C. § 1839(3); Meyers, *supra* note 2, at 18.

356 JAMES POOLEY, TRADE SECRETS § 1.01, at 1–6 (2014).

357 *Id.* at § 1.01, at 1-1, 1-5–6 (indicating that trade secret protection can cover customer lists, financial projections, pricing data, and marketing plans); Marguerite McConihe & Meena Seralathan, *Benefits of and Best Practices for Protecting Artificial Intelligence and Machine Learning Inventions as Trade Secrets*, NAT'L L. REV. (Feb. 10, 2022), https://www.natlawreview.com/article/benefits-and-best-practices-protecting-artificial-intelligence-and-machine-learning [https://perma.cc/S4TN-BSKY]; Matthew Kohel, *Trade Secrets May Offer the Best Protection for AI Innovation*, LAW360 (Feb. 21, 2023), https://www.law360.com/articles/1577741/trade-secrets-may-offer-the-best-protection-for-ai-innovation [https://perma.cc/37AQ-B463]; *see also* 18 U.S.C. § 1893(3); Experian Info. Sols., Inc. v. Nationwide Mktg. Servs. Inc., 893 F.3d 1176, 1179 (9th Cir. 2018) (recognizing that lists of information can be protected as trade secrets).

358 Kohel, *supra* note 357; *see* 18 U.S.C. § 1893(3)(B).

359 35 U.S.C. § 112.

360 Meyers, *supra* note 2, at 19.

361 *Id.*

362 *Id.*

363 Unif. Trade Secrets Act § 1(2) (2018); *see* Meyers, *supra* note 2, at 19.

364 Unif. Trade Secrets Act § 1 cmt. 2; *see also* Pamela Samuelson & Suzanne Scotchmer, *The Law and Economics of Reverse Engineering*, 111 YALE L.J. 1575, 1582 (2002).

Many aspects of AI may be protected as trade secrets, including the "structure of the AI/ML model, formulas used in the model, *proprietary training data*, a particular method of using the AI/ML model, any output calculated by the AI/ML model that is subsequently converted into an end product for a customer, and similar aspects of the platform."[365] There is no requirement of "human authorship" for trade secret subject matter,[366] so synthetic data wholly generated by AI systems would be eligible for protection. As such, trade secrets are a potentially significant innovation mechanism for protecting synthetic data and processes for generating it.

2. *Analysis and Prescriptions*

Trade secrets can play an important role in promoting the provisioning, disclosure, and democratization of synthetic data. First, trade secrets can incentivize synthetic data generation in ways that complement open source approaches and patent protection. Obviously, unlike open source approaches, trade secrecy allows synthetic data generators to keep proprietary synthetic data and processes secret while appropriating returns from innovation. Compared to patents, the low cost, lack of disclosure, and long-term protection of trade secrets render them particularly attractive. Notably, foundational empirical studies indicate that secrecy (as well as lead time) is generally more important than patents as a mechanism to appropriate returns from investments in innovation.[367]

Second, trade secrets have somewhat surprising effects on promoting the disclosure of synthetic data and processes for generating it. At first glance, it seems obvious that trade secrecy would decrease disclosure and transparency around synthetic data. Trade secrecy demands that claimants keep subject matter secret, and it establishes a cause of action against parties for misappropriation. The desire to maintain trade secret protection may discourage firms from disclosing

365 McConihe & Seralathan, *supra* note 357 (emphasis added); *see also* Meyers, *supra* note 2, at 18.

366 Lauren Castle, *Trade Secrets Summoned to Protect AI Amid Noncompete Uncertainty*, BLOOMBERG L. (July 16, 2024), https://news.bloomberglaw.com/ip-law/trade-secrets-summoned-to-protect-ai-amid-noncompete-uncertainty [https://perma.cc/B83V-PNLC].

367 Bronwyn Hall, Christian Helmers, Mark Rogers & Vania Sena, *The Choice Between Formal and Informal Intellectual Property: A Review*, 42 J. ECON. LIT. 375, 380 (2014).

synthetic data and processes for generating it.[368] This in turn would subvert the strong public policy interest in disclosing such subject matter so that independent parties can validate it.

Upon second glance, however, trade secrecy can, ironically, increase disclosure or at least sharing of "secret" technical information. While various justifications for trade secrets abound, the view that trade secrets advance traditional objectives of intellectual property has gained ascendance.[369] In addition to serving a provisioning function, intellectual property in many ways also promotes disclosure.[370] Though counterintuitive, legal scholar Mark Lemley argues that trade secrets promote the disclosure, or at least sharing, of information that parties would otherwise conceal quite stringently. For example, firms may be more comfortable sharing confidential information with external vendors, contractors, and clients if they can protect such information through trade secrecy.[371] In the absence of this legal protection, firms would likely implement draconian security measures, limit the use of such information to employees, or pursue vertical integration to prevent information leakage. Applied in the current context, protecting synthetic data and processes for generating it as trade secrets may encourage firms to share such secrets with outside parties, which can then independently verify and validate them. In short, in the absence of legal protection for trade secrets, there would be more, not less, secrecy.

Trade secrecy promotes, or at least permits, technical disclosure in other ways as well. Trade secrecy is a leaky regime, and protection ceases upon disclosure of a secret.[372] Employees tend to move from company to company, and given that "company proprietary information is often intertwined with an employee's knowledge and skill," firms likely cannot entirely prevent the leakage of trade secrets.[373] Furthermore, trade secrecy only protects against misappropriation, which does not include independent invention or reverse engineering.[374] Thus, if one firm protected synthetic data and processes for making it

---

368 Meyers, *supra* note 2, at 21.

369 *See* Lemley, *Trade Secrets*, *supra* note 354, at 319–41; Deepa Varadarajan, *Trade Secret Fair Use*, 83 FORDHAM L. REV. 1404, 1413–20 (2014).

370 Lemley, *Trade Secrets*, *supra* note 354, at 332–33.

371 *Id.* at 334–36.

372 Meyers, *supra* note 2, at 19.

373 *Id.*

374 *See supra* notes 363–64 and accompanying text.

as trade secrets, a competing firm could reverse engineer such creations and face no liability.

This Article argues for bolstering limitations on trade secrets to facilitate increased disclosure of synthetic data and processes for generating it. It joins others in advocating for more robust safe harbors to disclose trade secrets when doing so advances important policy interests.[375] This would include instances where a party seeks to improve upon an invention protected as a trade secret[376] and where disclosing a trade secret is important for advancing "public health, safety, and welfare."[377] A wealth of literature has questioned the appropriateness of strict trade secret protection for subject matter of high public policy importance, such as concerning healthcare prices, environmental protection, voting machines, breathalyzer devices, and search engine algorithms.[378] As applied in the present context, disclosure of trade secret-protected synthetic data and processes for creating it may implicate important public health, safety, and welfare interests when ML systems trained on such data decide healthcare expenditures, incarceration, and the allocation of government benefits. Whether through statutory carve-outs or a common-law doctrine of "trade secret fair use,"[379] this Article argues for a robust safe harbor for revealing secret synthetic data and processes for producing it—particularly for "whistleblower"-type disclosures—when important policy interests are at stake.[380]

Third, trade secrets can also advance democratization of the synthetic data landscape. Trade secret protection can perform a similar function as patents in enabling the existence of standalone synthetic data generators that supply data to

---

375 *See, e.g.*, Varadarajan, *supra* note 369, at 1404 (noting that public disclosure of secrets regarding fracking or prices for healthcare devices "ought to be encouraged, or at least, not discouraged"); Peter S. Menell, *Tailoring a Public Policy Exception to Trade Secret Protection*, 105 CALIF. L. REV. 1, 46–48 (2017) (proposing a safe harbor whereby parties subject to NDAs would be entitled to report alleged misconduct to government authorities through confidential communications).

376 *See* Varadarajan, *supra* note 369, at 1440; Derek E. Bambauer & Oliver Day, *The Hacker's Aegis*, 60 EMORY L.J. 1051, 1077 (2011).

377 Varadarajan, *supra* note 369, at 1441; *cf.* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018) (arguing against a trade secret privilege in criminal cases).

378 Varadarajan, *supra* note 369, at 1441–44 (discussing several contributions to the literature).

379 *See id.* at 1446–52.

380 *Cf.* Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 135–37 (2019) (advocating for safe harbors for the limited disclosure of trade secrets in situations involving algorithmic bias).

external clients. In some ways, the democratizing effects of trade secrets are even more pronounced because, unlike patents, they do not require long, expensive examination or public disclosure of information.[381] As such, trade secrets are particularly appealing to startups and small- and medium-sized entities. Indeed, empirical evidence has found that trade secrecy is more important than patent protection for software startups.[382] Given that trade secrecy insists upon "reasonable efforts" to maintain secrecy rather than absolute secrecy, a firm could protect its synthetic data as a trade secret, license it to customers under confidentiality agreements, and maintain protection.[383]

A related mechanism for democratizing access to synthetic data involves allowing employees in synthetic data fields to move to other firms or even start competing firms. Noncompetition agreements, often justified as protecting trade secrets, limit employee mobility and hamper technological diffusion.[384] In so doing, they increase industry concentration, decrease new business formation, and harm innovation.[385] In the present context, noncompetition agreements could prevent employees from leaving technology firms to start their own synthetic data generation ventures, thus reducing competition and parallel innovation. Encouraging in this regard, several states and the federal government have moved to ban or strictly limit noncompetition agreements. California has long banned noncompete agreements in most circumstances.[386] Recently, the FTC proposed a broader rule effectively banning noncompete

---

381 *See* David S. Levine & Ted Sichelman, *Why Do Startups Use Trade Secrets?*, 94 NOTRE DAME L. REV. 751, 761–63, 799 fig 3. (2019) (indicating that cost was the leading factor preventing software startups from patenting inventions that were patentable); Varadarajan, *supra* note 369, at 1405; McConihe & Seralathan, *supra* note 357.

382 Levine & Sichelman, *supra* note 381, at 796.

383 McConihe & Seralathan, *supra* note 357.

384 Bessen, *supra* note 262, at 6. *See generally* Christopher B. Seaman, *Noncompetes and Other Post-Employment Restraints on Competition: Empirical Evidence from Trade Secret Litigation*, 72 HASTINGS L.J. 1183, 1190–91 (2021).

385 *See* Non-Compete Clause Rule, 88 FED. REG. 3482, 3490 (Jan. 19, 2023) (to be codified at 16 C.F.R. pt. 910) [hereinafter Fed. Trade Comm'n, *Non-Compete Clause Rule*] ("There is also evidence non-compete clauses increase industrial concentration more broadly."); *id.* at 3491 ("The weight of the evidence indicates non-compete clauses likely have a negative impact on new business formation."); *id.* at 3492 ("The weight of the evidence indicates non-compete clauses decrease innovation.").

386 *See* Cal. Bus. & Prof. Code § 16600 (rendering noncompete agreements unenforceable except in the context of the sale of the goodwill of a business).

agreements nationwide.[387] The FTC's proposed rule also reaches other provisions, such as nondisclosure agreements, that sweep so broadly so as to function like noncompete agreements.[388] Independent of the FTC's proposed rule, several courts have invalidated nondisclosure agreements that functionally operate like noncompete agreements.[389] Taken together, removing such barriers to employee mobility can increase the number of firms (especially startups) in the synthetic data field, thus promoting the democratization of this technology.

D. Copyrights

1. *Overview*

A final intellectual property regime that can influence the generation of synthetic data is copyright law. Copyrights confer exclusive rights over original expression fixed in a tangible medium of expression.[390] Obtaining a copyright involves no application or examination, and the term of protection for most works is the life of the author plus seventy years.[391] Unlike patents, copyrights do not confer broad rights to exclude over protected subject matter. Rather, copyrights only confer a specific set of statutorily enumerated rights[392] and are further limited by the fair use doctrine.[393] This section explores three different aspects of synthetic data that may qualify for copyright protection: software for generating synthetic data, synthetic data itself, and the selection and arrangement of synthetic data in datasets.

First, software for generating synthetic data would be eligible for limited copyright protection. In general, copyrights do

---

387 *See* Press Release, Federal Trade Commission, FTC Proposes Rule to Ban Noncompete Clauses, Which Hurt Workers and Harm Competition (Jan. 5, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/01/ftc-proposes-rule-ban-noncompete-clauses-which-hurt-workers-harm-competition [https://perma.cc/S8PL-ZGEZ]; Fed. Trade Comm'n, *Non-Compete Clause Rule*, *supra* note 385. As of this writing, a federal judge has temporarily halted enforcement of this ban. *See* Castle, *supra* note 366.

388 Fed. Trade Comm'n, *Non-Compete Clause Rule*, *supra* note 385, at 3482.

389 *See* Camilla A. Hrdy & Christopher B. Seaman, *Beyond Trade Secrecy: Confidentiality Agreements That Act Like Noncompetes*, 133 YALE. L.J. 669, 677 (2024).

390 17 U.S.C. § 102.

391 17 U.S.C. § 102; 17 U.S.C. § 302.

392 *See* 17 U.S.C. §§ 106, 106A.

393 17 U.S.C. § 107.

not protect any "process, system, [or] method of operation."[394] Accordingly, disembodied processes or algorithms for generating synthetic data are not copyrightable.[395] While functional subject matter is generally not protectable, courts have recognized that various aspects of software may comprise copyrightable expression.[396] Thus, for instance, source code and the structure, sequence, and organization of software to generate synthetic data may be copyrighted. However, given the functional nature of software, the scope of copyright would be relatively narrow, largely limited to protecting against virtually identical copying.[397] Furthermore, the Supreme Court has ruled that the fair use doctrine should be construed broadly to permit unauthorized use of at least some kinds of software for certain purposes.[398]

Second, beyond the underlying software itself, synthetic data itself may be copyrightable, but it faces formidable obstacles from the authorship requirement for protectability. Under U.S. law, copyright only extends to "original works of *authorship*."[399] In the context of generative AI, the Copyright Office has stated that works arising solely from a machine with minimal creative input from a human fail the authorship requirement.[400] The U.S. District Court for the District of

---

394 17 U.S.C. § 102(b).

395 *Cf.* Baker v. Selden, 101 U.S. 99, 102 (1880) (holding that protecting a process of double-entry bookkeeping was "the province of letters-patent, not of copyright").

396 *See, e.g.*, Apple Comput., Inc. v. Franklin Comput. Corp., 714 F.2d 1240 (3d Cir. 1983) (holding that source code and object code are copyrightable); Comput. Assoc. Int'l Inc. v. Altai, Inc., 982 F.2d 693 (2d Cir. 1992) (holding that the structure, sequence, and organization of software are copyrightable); Oracle Am., Inc. v. Google LLC, 750 F.3d 1339 (Fed. Cir. 2014) (holding that application programming interfaces are copyrightable).

397 *See* Apple Comput., Inc. v. Microsoft Corp., 35 F.3d 1435, 1439 (9th Cir. 1994) ("When the range of protectable and unauthorized expression is narrow, the appropriate standard for illicit copying is virtual identity.").

398 Google LLC v. Oracle Am., Inc., 593 U.S. 1 (2021).

399 17 U.S.C. § 102 (emphasis added); *see* Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53, 59–60 (1884).

400 Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 FED. REG. 16190, 16192 (Mar. 16, 2023) (to be codified at 88 C.F.R. pt. 202) [hereinafter U.S. Copyright Office, *Artificial Intelligence*]. Earlier, the Copyright Office issued similar guidance:

> Similarly, the Office will not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author. The crucial question is "whether the 'work' is basically one of human authorship, with the computer [or other device] merely being an assisting instrument, or whether the traditional elements of authorship in

Columbia has ruled similarly.[401] These authorities suggest that synthetic text, images, and other "data" created wholly by AI, with minimal human input, would fail the authorship requirement and not be copyrightable. However, U.S. copyright law leaves some avenues by which synthetic data generated by AI may satisfy authorship. To the extent that humans provide creative inputs to AI systems that operate in a fairly understandable or predictable manner, they are more likely to qualify as the authors of resulting outputs.[402] Additionally, modifications to AI-generated outputs (including, presumably, synthetic data) may be copyrightable.[403] For instance, human-modified synthetic data, perhaps as the result of "cleaning," may pass the authorship threshold.[404] Third, a work containing AI-generated content may be copyrightable based on the selection and arrangement of elements in the work, even if the AI-generated content itself is not protectable.[405]

---

the work (literary, artistic, or musical expression or elements of selection, arrangement, etc.) were actually conceived and executed not by man but by a machine."

U.S. Copyright Office, Compendium of U.S. Copyright Office Practices § 313.2 (3d ed. 2021) (quoting U.S. COPYRIGHT OFFICE, REPORT TO THE LIBRARIAN OF CONGRESS BY THE REGISTER OF COPYRIGHTS 5 (1966)); *see* U.S. Copyright Office, Re: Zarya of the Dawn (Registration # VAu0011480196) 4–10 (Feb. 21, 2023) [hereinafter U.S. Copyright Office, *Zarya*] (indicating that images generated from a generative AI system in a comic book are not copyrightable); Michael D. Murray, *Generative and AI Authored Artworks and Copyright Law*, 45 HASTINGS COMM. & ENT. L.J. 27, 32–33, 35 (2023). This is an old debate. *See, e.g.*, Pamela Samuelson, *Allocating Ownership Rights in Computer-Generated Works*, 47 U. PITT L. REV. 1185, 1199 (1986) (arguing that computers should not be treated as authors for copyright purposes).

401 Thaler v. Perlmutter, No. 22-1564, 2023 WL 5333236, at *7–8 (D.D.C. Aug. 18, 2023).

402 *Cf.* U.S. Copyright Office, *Artificial Intelligence*, *supra* note 400, at 16192; U.S. Copyright Office, *Zarya*, *supra* note 400, at 9 ("The fact that Midjourney's specific output cannot be predicted by users makes Midjourney different for copyright purposes than other tools used by artists.").

403 U.S. Copyright Office, *Artificial Intelligence*, *supra* note 400, at 16192–93.

404 Such modifications must be more than "minor and imperceptible" to pass muster. U.S. Copyright Office, *Zarya*, *supra* note 400, at 11.

405 U.S. Copyright Office, *Artificial Intelligence*, *supra* note 400, at 16192; *see* U.S. Copyright Office, *Zarya*, *supra* note 400. Notably, jurisdictions outside of the United States have been more permissive regarding the authorship status of AI. *See, e.g.*, Copyright, Designs and Patents Act 1988, ch. 48, § 9 (UK), https://www.legislation.gov.uk/ukpga/1988/48/section/9/enacted [https://perma.cc/H955-HPXP] ("In the case of literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements for the creation of the work are undertaken."); Copyright Act 1994, § 5(2)(a) (N.Z.), https://www.legislation.govt.nz/act/public/1994/0143/latest/DLM345899.html [https://perma.cc/323C-FMF8].

If authorship can be established, synthetic data is likely copyrightable, though further complications remain. Foundational copyright doctrine holds that facts, including data, are not copyrightable.[406] The defining characteristic of a copyrightable work is originality,[407] which means that the work is independently created and exhibits a modicum of creativity.[408] Real-world facts fail both prongs of this requirement because people discover rather than create facts, and facts (if true) do not display any creativity.[409] However, courts have held that so-called "fictional facts" satisfy originality and constitute creative expression.[410] Thus, for instance, fictional facts from the *Harry Potter* books (such as the "fact" that Harry attended Hogwarts) constitute original, protectable expression.[411] Relatedly, courts have held that facts infused with "professional judgement and expertise" may constitute protectable expression.[412] Thus, for instance, projections of used car prices represent original expression qualifying for copyright protection.[413] As applied here, assuming authorship is satisfied, fictional medical records, fabricated streetscapes, and other kinds of synthetic data may qualify for copyright protection as original expression.[414]

---

406 Feist Pubs., Inc. v. Rural Tel. Serv., 499 U.S. 340, 344 (1991); *see* 17 U.S.C. § 102(b) (providing a nonexhaustive list of exclusions from copyright protection, including "discover[ies]"); DAVID NIMMER, NIMMER ON COPYRIGHT § 2.03[E] (equating facts with "discoveries").

407 *Feist*, 499 U.S. at 345 ("The *sine qua non* of copyright is originality.").

408 *Id.*

409 *See Feist*, 488 U.S. at 347, 356; Experian Info. Sols., Inc. v. Nationwide Mktg. Servs. Inc., 893 F.3d 1176, 1181 (9th Cir. 2018).

410 *See, e.g.*, Ariel M. Fox, *Aggregation Analysis in Copyright Infringement Claims: The Fate of Fictional Facts*, 115 COLUM. L. REV. 661 (2015).

411 Warner Bros. Ent. Inc. v. RDR Books, 575 F. Supp. 2d 513, 535 (S.D.N.Y. 2008) (characterizing "fictional facts" as "entirely the product of the original author's imagination and creation"); Paramount Pictures Corp. v. Carol Publ'g Grp., 11 F. Supp. 2d 329, 333 (S.D.N.Y. 1998) (stating that "[t]he characters, plot and dramatic episodes" of *Star Trek* are the story's "original elements" and are protected by copyright); Castle Rock Ent. Inc. v. Carol Publ'g Grp., Inc., 150 F.3d 132, 139 (2d Cir. 1998) (holding that invented facts from the *Seinfeld* universe represent creative expression); *but see* Jeanne C. Fromer, *An Information Theory of Copyright Law*, 64 EMORY L.J. 71, 100 (2014) (suggesting that fictional facts in compendia should be freely available, just like facts about the real world).

412 CCC Info. Serv. v. MacLean Hunter Mkt. Reps., 44 F.3d 61, 67 (2d Cir. 1994); *accord* CDN Inc v. Kapes, 197 F.3d 1256, 1260 (9th Cir. 1999) (holding that estimates of wholesale prices of collectible coins satisfied the originality requirement because they were "wholly the product of [the copyright owner's] creativity").

413 *CCC*, 44 F.3d at 67.

414 An additional potential objection to the copyrightability of synthetic data is that it is functional (because it is created to train ML systems) and thus not protectable. This reasoning, however, is in tension with the established practice

Third, regardless of whether synthetic data itself is copyrightable, the selection and arrangement of synthetic data in a dataset may be.[415] While facts are not original, the selection and arrangement of facts in a compilation may satisfy originality.[416] Of course, a selection of information that is entirely "typical," "obvious," or "basic" is not creative.[417] Similarly, arranging information in an "age-old," traditional, or "commonplace" manner also lacks creativity.[418] However, other, more creative ways of selecting and arranging data may be protectable.[419] Even a "logical" selection and arrangement of facts may satisfy originality.[420] Most relevant to massive synthetic databases, the Ninth Circuit held in *Experian Information Solutions, Inc. v. Nationwide Marketing Services, Inc.*, that the selection and arrangement of hundreds of millions of fields in a database were copyrightable.[421] Thus, even if courts construe synthetic data as uncopyrightable facts, the generators of synthetic data may claim copyright in the selection and arrangement of those facts. For instance, the selection and arrangement of synthetic medical records in a dataset, which reflect human judgment about which diseases to track and how to organize them, may be copyrightable.

---

of granting copyrights on many works that are admittedly functional, such as projected used car prices, instructional manuals, and software.

415 *See* Feist Pubs., Inc. v. Rural Tel. Serv., 499 U.S. 340, 344 (1991); Experian Info. Sols., Inc. v. Nationwide Mktg. Servs. Inc., 893 F.3d 1176, 1181 (9th Cir. 2018); 17 U.S.C. § 101 (noting that compilations can be "selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship"); § 103(a) (indicating that compilations may comprise copyrightable subject matter).

416 *Feist*, 499 U.S. at 348, 358; *see also CCC*, 44 F.3d at 65–66.

417 *Feist*, 499 U.S. at 362.

418 *Id.* at 363; *accord* Matthew Bender & Co. v. West Publ'g Co., 158 F.3d 674, 677 (2d Cir. 1998) (holding that the selection and arrangement of judicial opinions were "obvious, typical, and lack[ed] even minimal creativity").

419 *See Experian*, 893 F.3d at 1185 ("[T]he creativity that suffices to establish copyright protection in factual compilations is minimal."); *CCC*, 44 F.3d at 67 (holding that a firm's selection and arrangement of used car value projections divided by region and 5,000-mile increments satisfied originality).

420 *See CCC*, 44 F.3d at 67.

421 *Experian*, 893 F.3d at 1185; *accord* Mason v. Montgomery Data, Inc., 967 F.2d 135, 141 (5th Cir. 1992) (ruling that the selection of real estate data in a compilation was copyrightable when the compiler makes "choices . . . independently . . . to select information from numerous and sometimes conflicting sources").

However, even if the selection and arrangement of data are copyrightable, such protection is thin.[422] In *Experian*, the Ninth Circuit ruled that although Experian's selection and arrangement of data in 250 million name-address pairings were copyrighted, Natimark was not liable for infringement because it only copied 200 million name-address pairings.[423] This and other cases address databases consisting of uncopyrightable information, and it is possible that courts would confer greater protection to the selection and arrangement of databases comprised of copyrightable expression. Nonetheless, the selection and arrangement of synthetic data in a database is likely to receive rather limited protection.

2. *Analysis and Prescriptions*

Copyright can help promote the provisioning, disclosure, and democratization of synthetic data, though its contributions are likely to be modest. First, copyright can enhance incentives to invest in creating synthetic data and software for generating it. To the extent that parties can satisfy the authorship and originality requirements for synthetic data and the selection and arrangement of data in synthetic databases, they could press copyright claims against alleged infringers. In this fashion, copyright can exclude free riders and shore up incentives to create. The closest analog may be to data vendors that have asserted copyright claims against parties that have copied huge datasets without authorization.[424] However, if AI systems wholly determine both synthetic data and its selection and arrangement, the authorship requirement would not be satisfied, and these "creations" would not be copyrightable. Firms would stand on firmer ground to assert copyright infringement claims on software for generating synthetic data. As noted, however, copyright protection for software is rather limited and subject to liberal fair use. As such, most entities seeking to assert exclusive rights over software may be better suited to pursue patent or trade secret protection.

---

422 *Feist*, 499 U.S. at 349, 358; *Experian*, 893 F.3d at 1185; *accord* Kregos v. Associated Press, 937 F.2d 700, 702 (2d Cir. 1991) (holding that forms displaying baseball statistics were copyrightable, but that the plaintiff could only prevail against parties who used forms copying this particular selection of information).

423 893 F.3d at 1187–88.

424 *Experian*, 893 F.3d at 1185.

Second, copyright can also enhance the disclosure of synthetic data and processes for generating it.[425] Again, assuming that authorship can be satisfied (which is a big assumption), copyright protection for synthetic data may allow developers of such data to publicly disclose it, reassured that unauthorized copying of such data would comprise copyright infringement. This may provide another option for synthetic data developers who prefer not to release their data as open source or to subject it to trade secrecy. Similarly, copyright protection may also encourage developers to publicize the software underlying synthetic data generation. This would allow external parties to examine and validate the processes used to generate such data.

To aid efforts to analyze software that generates synthetic data, this Article argues that the unauthorized copying of software to determine how it works should weigh strongly in favor of fair use. This proposal would draw upon existing judicial exemptions from the exclusivity ordinarily afforded by copyrights. For example, courts have ruled that unauthorized copying of software to facilitate interoperability constitutes fair use.[426] One principle animating such rulings is that not allowing fair use would provide a de facto copyright over the functional aspects of software, which are not protectable.[427] Similarly, this Article calls on courts to clarify that the unauthorized copying of software (including software that generates synthetic data) to determine how it functions should weigh strongly in favor fair use. This consideration would inform both fair use factor 1 (the purpose of the defendant's use) and factor 4 (market impact), which the Supreme Court has recently held should consider the "public benefits" of unauthorized copying.[428] There is considerable public benefit to allowing independent verification (which may necessitate copying) of software that generates synthetic data.[429]

---

425 *Cf.* Lemley, *Trade Secrets*, *supra* note 354, at 333 (discussing how copyright promotes disclosure in several ways). *But see* James Gibson, *Once and Future Copyright*, 81 NOTRE DAME L. REV. 167, 178 (2005) (critiquing copyright protection of software without requiring disclosure of the underlying source code).

426 *See, e.g.*, Sega Enters., Ltd. v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1992) (holding that Accolade's reverse engineering of Sega's copyrighted code to create interoperable games constituted fair use); Google LLC v. Oracle Am., Inc., 593 U.S. 1, 40 (2021) (holding that Google's copying of the Sun Java API to facilitate interoperability constituted fair use).

427 *Sega Enters.*, 977 F.2d at 1526.

428 *Google*, 593 U.S. at 36.

429 Notably, Congress has also limited copyright to allow for research and security testing on copyrighted material. The Digital Millennium Copyright Act

Finally, copyright can play a supporting role in democratizing the synthetic data landscape. Though not as prominent as for patents, there is some recognition that copyright enables the independent existence of creative firms that sell or license copyrighted outputs to outside parties.[430] Indeed, empirical evidence suggests that copyrights are more important than both patents and trade secrecy for software startups.[431] To the extent that synthetic data is copyrightable, exclusive rights may encourage investment in generating it, particularly by standalone data providers that sell it to external clients. Copyright protection on software to generate synthetic data may provide a similar entrepreneurial incentive, though it must be qualified by the substantial limitations on software copyright.

E. A Diverse Innovation Ecosystem for Synthetic Data and the Recursive Nature of Technology and Law

This Part has sketched the contours of an innovation ecosystem to promote the robust, responsible development of synthetic data. It is important to emphasize that these prescriptions are aimed at improving the quality and accessibility of synthetic data through encouraging innovation, transparency, and parallel development. As such, these prescriptions can counteract the harms of low-quality synthetic data. As mentioned, however, greater access to high-quality synthetic data can cause its own harms by enabling parties to wield high-powered ML systems for nefarious purposes.[432] Accordingly, these prescriptions are intended to augment, rather than replace, traditional regulations of ML systems in general and synthetic data in particular.

It is also worth emphasizing that this Article does not advocate reconfiguring general rules of intellectual property law simply to promote advancements in synthetic data. However, each of these legal fields possesses context-sensitive doctrines

---

(DMCA) establishes liability for circumventing a technological measure—like encryption—that controls access to a copyrighted work. 17 U.S.C. § 1201. However, the DMCA establishes several exceptions permitting circumvention of such measures to aid law enforcement, facilitate interoperability, and promote research. 17 U.S.C. § 1201(e)–(g). *See generally* Pamela Samuelson, *Towards More Sensible Anti-Circumvention Regulations*, *in* FINANCIAL CRYPTOGRAPHY 33 (Yair Frankel ed., 2001).

430 *Cf.* Peter Lee, *Autonomy, Copyright, and Structures of Creative Production*, 83 OHIO ST. L.J. 283 (2022).

431 Levine & Sichelman, *supra* note 381, at 795.

432 *See supra* notes 180–81 and accompanying text.

that courts can contour to specific technologies.[433] This Article maintains, moreover, that many of the doctrinal prescriptions here—such as encouraging technical disclosure, allowing unauthorized use of intellectual property to see how it works, and promoting new entity formation—would stimulate innovation in a wide array of technological fields.

For analytic purposes, this Part has examined individual innovation mechanisms separately. However, these innovation mechanisms can overlap, and firms may use several of them to protect different aspects of synthetic data and processes for generating it.[434] For instance, a firm may patent its general process for generating synthetic data, copyright its software for doing so, and protect its synthetic data as a trade secret. Entities may even combine open source and proprietary approaches. For instance, some startups are drawing on open source synthetic data generators while protecting aspects of implementation and customization as trade secrets. While combining multiple innovation mechanisms may be helpful to entities developing synthetic data, policymakers should be vigilant in ensuring that such layering does not undermine carefully crafted limitations and exceptions to exclusive rights.

This wide variety of innovation mechanisms can promote helpful diversity in the ecosystem of entities researching, refining, and producing synthetic data. Government, academic, and nonprofit entities may be drawn to open source approaches, though we have seen that for-profit entities have pursued them as well. Trade secrets and copyrights may appeal to startups and under-resourced entities due to their low cost of acquisition, while patents may be more attractive to larger, more established players. Consistent with several of the themes of this Article, such diversity of innovation mechanisms and actors provides fertile ground for competition, cross-checking, and ultimately, communal improvement of synthetic data technologies. Just as ML systems improve with more (high-quality) data, a vigorous, diverse ecosystem of numerous kinds

---

433 *See, e.g.*, Dan L. Burk & Mark A. Lemley, *Policy Levers in Patent Law*, 89 VA. L. REV. 1575 (2003). Application of copyright's fair use doctrine is also famously context-specific.

434 *See, e.g.*, Levine & Sichelman, *supra* note 381, at 798, 805–06 (indicating that startups often protect creations with both patents and trade secrets, which operate as complements); McConihe & Seralathan, *supra* note 357 (recommending that firms use both patents and trade secrets to protect elements of AI and ML models). *See generally* Mark P. McKenna, *An Alternate Approach to Channeling?*, 51 WM. & MARY L. REV. 873 (2009).

of entities working in parallel on synthetic data promises the most robust innovation and quality control.

Notably, this Article illustrates that innovation mechanisms perform a variety of functions beyond their classic role of provisioning information goods. Certainly, innovation mechanisms help overcome public goods problems and shore up incentives to create information assets like synthetic data. A less appreciated attribute of innovation mechanisms, however, is that they also encourage the public disclosure or at least sharing of new technical creations. In the context of synthetic data, this disclosure function is particularly important given the policy interest in countering the black box nature of AI and ML systems. Finally, innovation mechanisms can also aid in democratizing technological landscapes. Properly calibrated, innovation mechanisms can allow startups and new entrants to enter markets and compete against incumbents, thus increasing the sources of innovation in a field. Again, in the context of synthetic data, such democratization is particularly important given the need to cross-check, validate, and enable competing sources of synthetic data.

At a broader level, these observations highlight the recursive nature of technology and law. Recursiveness is a theme that runs throughout this Article. At a technical level, AI models often generate synthetic data, which then trains other AI models. The irony is not lost that the prescriptions offered here will lead to more artificial data being used to train artificial intelligence. Machines, in a sense, teaching machines. This Article also highlights the recursive relationship between technology and law. AI and ML pose pressing challenges to law in the form of privacy violations, discrimination, and copyright infringement, and will continue to shape legal regimes going forward. However, in a reciprocal fashion, law is constitutive of technology, and laws and policies determining the innovation ecosystem for synthetic data will shape the future of AI. Modifications to open source policies and intellectual property doctrines governing patents, trade secrets, and copyrights can promote the provisioning, disclosure, and democratization of synthetic data and help unleash the full potential of AI.

Conclusion

This is an Article about inputs. Many of the technical and legal problems of AI and ML derive from the limitations of a critical input, namely real-world data. Amassing huge amounts of high-quality, real-world data is difficult, and

doing so can undermine privacy, introduce bias in automated decision-making, and infringe copyrights on a massive scale. Accordingly, this Article has explored the emergence of a seemingly paradoxical technical input that can mitigate (though not completely resolve) these concerns: synthetic data. Synthetic data is a heterogeneous category encompassing data of differing degrees of artificiality arising from different technological approaches. Yet it is clear that synthetic data will play a dominant role in training the ML systems of tomorrow. Quite simply, the future of AI is synthetic.

But inputs have inputs, too. In light of the enormous importance and value of synthetic data, this Article has explored the contours of an innovation ecosystem to promote synthetic data's robust and responsible development. It has focused on three public policy objectives that should guide the development of synthetic data: provisioning, disclosure, and democratization. This Article has then examined a wide array of innovation mechanisms spanning open source methods and proprietary approaches based on patents, trade secrets, and copyrights. Throughout, it has suggested policy reforms to maintain incentives to create high-quality synthetic data, provide wide access to the technical details of synthetic data and its generation, and pluralize the sources of synthetic data production. In so doing, this Article sheds light on the recursive nature of technology and law. Just as AI will shape legal regimes going forward, law and policy can help determine critical inputs to AI, thus shaping the future of this transformative technology.