The Global Commons of Data

Jennifer Shkabatur*

22 STAN. TECH. L. REV. 354 (2019)

ABSTRACT

Data platform companies (such as Facebook, Google, or Twitter) amass and process immense amounts of data that is generated by their users. These companies primarily use the data to advance their commercial interests, but there is a growing public dismay regarding the adverse and discriminatory impacts of their algorithms on society at large. The regulation of data platform companies and their algorithms has been hotly debated in the literature, but current approaches often neglect the value of data collection, defy the logic of algorithmic decision-making, and exceed the platform companies' operational capacities.

This Article suggests a different approach—an open, collaborative, and incentives-based stance toward data platforms that takes full advantage of the tremendous societal value of user-generated data. It contends that this data shall be recognized as a "global commons," and access to it shall be made available to a wide range of independent stakeholders—research institutions, journalists, public authorities, and international organizations. These external actors would be able to utilize the data to address a variety of public

* Assistant Professor, Lauder School of Government, Diplomacy & Strategy, Interdisciplinary Center (IDC), Herzliya, Israel; and Senior Consultant on Information & Communication Technologies to the World Bank. Harvard Law School, S.J.D. (2012); LLM (2007). I am grateful for comments, insights, and suggestions provided by the fellows of the Lauterpacht Centre for International Law at the University of Cambridge, United Kingdom, and the participants of the International Law & Technology workshop at the Buchmann Faculty of Law in Tel Aviv University, and the Globalization of Law Workshop at the Interdisciplinary Center (IDC), Herzliya, Israel. I am particularly indebted to Eyal Benvenisti and Shay Lavie for their inputs into this Article. All mistakes are mine, and the statements in this Article do not reflect the position of the World Bank.

challenges, as well as observe from within the operation and impacts of the platforms' algorithms.

After making the theoretical case for the "global commons of data," the Article explores the practical implementation of this model. First, it argues that a data commons regime should operate through a spectrum of data sharing and usage modalities that would protect the commercial interests of data platforms and the privacy of data users. Second, it discusses regulatory measures and incentives that can solicit the collaboration of platform companies with the commons model. Lastly, it explores the challenges embedded in this approach.

TABLE OF CONTENTS

I. INTRODUCTION 356
II. THE PUBLIC IMPORTANCE OF PRIVATE DATA PLATFORMS 363
A. Algorithmic Impacts on Democratic Processes and Elections 365
B. Discriminatory Impacts of Algorithms 368
III. REGULATING DATA PLATFORMS: CURRENT APPROACHES 371
A. Algorithmic Transparency and Explanation 371
B. Due Process 375
C. The Inherent Capacity Limitations of Data Platforms 379
IV. THE CASE FOR A GLOBAL COMMONS OF DATA 380
A. Making the Case 380
B. The Spectrum of the Global Data Commons 385
1. Sharing Internal Data Analysis 385
2. Releasing Targeted Data 387
3. Data Pools 390
4. Granting Access to Public Actors 393
5. Open access 395
C. Regulatory Measures & Incentives 398
1. Sticks: Invoking the Public Utilities Doctrine 399
2. Carrots: Invoking Financial and Social Incentives 402
D. Challenges 404
1. Addressing Privacy Concerns 405
2. Ensuring Users' Consent 407
E. Competitive Concerns 409
V. CONCLUSION 410

I. INTRODUCTION

In July 2018, Facebook released to external and independent researchers a trove of data that its users shared on the social network in 2017-2018,[1] kicking off a new initiative to enable credible research about the role of social media in elections.[2] The company vowed that researchers would not require its approval to publish their findings.[3] This initiative came in response to the vehement public outrage regarding the use (and abuse) of social networks during the presidential elections campaigns of 2016[4]—the massive data collection undertaken by Cambridge Analytica to target users with narrowly-tailored political ads,[5] and the alleged activities undertaken by Russian hackers to generate fake profiles and widely disseminate information that fits the Kremlin agenda.

This Article contends that while Facebook may have been publicly impelled to open its data, its initiative shall constitute a general policy, and

1. Facebook released to external and independent researchers a large dataset that contains web page addresses (URLs) that have been shared on Facebook starting January 1, 2017 and ending about a month before the present day. URLs are included if they are shared by at least 20 unique accounts, and at least once publicly. Facebook estimates that the full dataset will contain around 2 million unique URLs that have been shared in 300 million posts, per week. *See* Russel Brandom, *Facebook Opens Up 'Overwhelming Data Set' for Election Research*, THE VERGE (July 11, 2018), https://perma.cc/78FX-M69G; Solomon Messing, Bogdan State, Chaya Nayak, Gary King & Nate Persily, *Facebook URL Shares*, HARVARD DATAVERSE (July 11, 2018), https://perma.cc/NU8L-E9AU.

2. Elliot Schrage & David Ginsberg, *Facebook Launches New Initiative to Help Scholars Assess Social Media's Impact on Elections,* FACEBOOK NEWSROOM (Apr. 9, 2018), https://perma.cc/D5BX-C5L7. Facebook has formed a commission of renowned academic experts, who then issued an open call for proposals, inviting researchers to obtain access to the data. *See Our Facebook Partnership*, SOCIAL SCIENCE ONE, https://perma.cc/LD2T-MWKQ (archived July 26, 2019); Solomon Messing, Chaya Nayak, Gary King & Nathaniel Persily, *Facebook URL Shares: Codebook*, SOCIAL SCIENCE ONE (July 11, 2018), https://perma.cc/R2KK-CG93.

3. Schrage & Ginsberg, *supra* note 2.

4. *See, e.g.*, Sarah Frier, *Zuckerberg's Crisis Response Fails to Quiet Critics*, BLOOMBERG (March 22, 2018), https://perma.cc/P5ZS-99XE; Jason Murdock, *#DeleteFacebook Is Trending, Is This the End of the Social Network?*, NEWSWEEK (Mar. 20, 2018), https://perma.cc/Q8KS-UA6G.

5. Matthew Rosenberg, Nicholas Confessore & Carole Cadwalladr, *How Trump Consultants Exploited the Facebook Data of Millions*, N.Y. TIMES (Mar. 17, 2018), https://perma.cc/TR28-SKB2.

our approach to data generated on social networks and other online platforms shall be thoroughly reconsidered. Companies such as Google, Facebook, Apple, and eBay have amassed more data about people and their behavior, health, markets and networks than many governments and organizations around the globe. This data could enlighten us about ourselves, and instruct us on various matters, such as how to improve our health, make better informed political decisions, or design more accessible and efficient markets. The data could also suggest areas for institutional attention and regulation, and unveil how the algorithms of data companies operate and whether they result in discriminatory or otherwise problematic decisions. A resource that fulfills such a critical function cannot be managed by private commercial entities for profit purposes only.

Rather, this Article argues that data that is accumulated on private data platforms shall be recognized as a "global commons." A commons regime signifies that access to user-generated[6] data possessed by platform companies would not only be available to these companies, but to a broader range of stakeholders. The latter would take advantage of these access rights to discern from the data insights that are valuable for decision-making processes and also monitor from within the operation and impacts of the platform's algorithms. This Article seeks to make the case for the recognition of a "global data commons," explain why other approaches to the regulation of platforms companies are likely to be ineffective, and suggest how a data commons regime could be implemented in practice.

The Article starts by illuminating the formidable public function of data platform companies—private entities that enable various types of online information exchanges among their users.[7] These include social networks (e.g., Facebook, Twitter, Instagram), search engines (e.g., Google, Bing or Yahoo), online marketplaces (e.g., Amazon or eBay), recommendation systems (e.g., Yelp), payment systems (e.g., Google Wallet, PayPal, Visa, MasterCard), virtual labor or service exchanges (e.g., Uber or AirBnb), and

6. User-generated data, also known as user-generated content, is defined as content uploaded and sometimes created by Internet users, rather than produced by the website itself. For a comprehensive definition, see Steven Hetcher, *User-Generated Content and the Future of Copyright: Part One—Investiture of Ownership*, 10 VAND. J. ENT. & TECH. L. 863, 870 (2008)

7. Becky Carter, *Infomediaries and Accountability*, GOVERNANCE & SOC. DEV. RESOURCE CTR. (2016), https://perma.cc/7DWE-VEUM (explaining how information platforms "synthesise, translate, simplify and direct information on behalf of others").

others. These companies play a constitutive role in the twenty-first century's economy and in the daily lives of billions of people.[8] For instance, 45% of Americans get their news on Facebook,[9] which generally consumes an average of fifty minutes of its users' time every single day.[10] Google, Microsoft, and Yahoo together control 98% of the U.S. search-engine market.[11] Amazon accounts for 43% of U.S. online retail sales.[12] Facebook and Google control 73% of all digital advertising in the U.S.[13]

Despite the wide array of services that they offer, data platform companies utilize a similar operation mode: they encourage users to generate and share data on their platforms, and then employ complex "big data"[14] algorithms that aggregate, process, and analyze this user-generated

8. Oren Bracha & Frank Pasquale, *Federal Search Commission? Access, Fairness and Accountability in the Law of Search*, 93 CORNELL L. REV. 1149, 1163 (2008) (explaining argument in literature that "network gatekeepers, who exercise control over the Internet's technological bottlenecks, constitute the new speech intermediaries"); James Grimmelmann, *The Google Dilemma*, 53 N.Y.L. SCH. L. REV. 939, 940 (2009) ("Whoever controls the search engines, perhaps, controls the Internet itself."); Neil Weinstock Netanel, *New Media in Old Bottles? Barron's Contextual First Amendment and Copyright in the Digital Age*, 76 GEO. WASH. L. REV. 952, 953 (2008) (stating that the "bulk of scholarly and activist attention" has moved to ensuring "access to the conduits of digital communication"); Christopher S. Yoo, *Free Speech and the Myth of the Internet as an Unintermediated Experience*, 78 GEO. WASH. L. REV. 697, 697 (2010) ("In recent years, concerns about the role of Internet intermediaries have continued to grow."); Editorial, *The Google Algorithm*, N.Y. TIMES (July 14, 2010) (calling Google "the gatekeeper of the Internet"), https://perma.cc/57AG-KNNS.

9. Elisa Shearer & Jeffrey Gottfried, *News Use Across Social Media Platforms*, PEW RES. CTR. (Sept. 7, 2017), https://perma.cc/JM49-EUFA.

10. James B. Stewart, *Facebook Has 50 Minutes of Your Time Each Day. It Wants More*, N.Y. TIMES (May 5, 2016), https://perma.cc/9T3W-JHH9.

11. *Market Share of Search Engines in the United States from December 2008 to 2018,* STATISTA (Feb. 2019), https://perma.cc/PSY5-2DQJ.

12. Business Insider Intelligence, *Amazon Accounts for 43% of US Online Retail Sales*, BUSINESS INSIDER (Feb. 3, 2017), https://perma.cc/4XJQ-FPPL.

13. Davey Alba, *Google and Facebook Still Reign over Digital Advertisng,* WIRED (July 29, 2017) https://perma.cc/H74R-SAVC; Jillian D'Onfro, *Google and Facebook Extend Their Lead in Online Ads,* CNBC (Dec. 20, 2017), https://perma.cc/K6T4-E2VS; Reuters, *Why Google and Facebook Prove the Digital Ad Market Is a Duopoly* (July 28, 2017), https://perma.cc/B7BS-5DKK.

14. "Big data" can be defined as high-volume, high-velocity (the rate at which data is generated), or high-variety (the type of data collected) information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. Big data is being generated by everything around us at all times. Every digital process and social media exchange produces it, and a wide variety of systems, sensors and mobile devices transmit it. *See* Jonathan Stuart Ward & Adam Barker, Undefined by Data: A Survey of Big Data Definitions 1-2 (Sept. 20,

data. The outputs of these algorithms inform decisions that affect many aspects of our society. They can determine which school a child can attend,[15] whether a person will be offered a bank credit,[16] what products are advertised to consumers in specific locations,[17] and whether a job applicant will be granted an interview.[18] Government officials also use them to predict issues such as where crimes will take place,[19] who is likely to commit a crime, and whether someone should be allowed out of jail on bail.[20]

The algorithmic prowess of data platform companies also has a staggering political impact. During the 2016 presidential election campaigns in the United States, for instance, it was widely discussed in the media that the algorithms of Google or Facebook could prioritize some types of political contents over other, and thus affect the results of the election.[21] Both companies have forcefully denied these allegations, but there is little controversy that their algorithms are generally capable of such distortions. Social experiments that have been carried out by researchers with access to the databases of these companies confirmed the potential of their

2013) (unpublished manuscript), https://perma.cc/5VPT-5SRB (describing several definitions of "big data"); *see also* Danah Boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFO. COMM. & SOC'Y 662, 663 (2012).

15. Benjamin Herold, *"Open Algorithms" Bill Would Jolt New York City Schools, Public Agencies*, EDUC. WEEK (Nov. 8, 2017), https://perma.cc/N6RL-PLML.

16. Bruce Schneier, *The Risks—and Benefits—of Letting Algorithms Judge Us*, CNN (Jan. 11, 2016), https://perma.cc/2T47-SW5E.

17. Ryan Singel, *Analysis: Google's Ad Targeting Turns Algorithms on You*, WIRED (Nov. 11, 2009), https://perma.cc/C937-AB7G.

18. *Now Algorithms Are Deciding Whom to Hire, Based on Voice*, NPR (Mar. 23, 2015) https://perma.cc/L9DH-MA5Q.

19. Justin Jouvenal, *Police Are Using Software to Predict Crime*, WASH. POST (Nov. 17, 2016), https://perma.cc/J5ME-Y46Z.

20. Tricia L. Nadolny, *How Computers Are Predicting Crime—And Potentially Impacting Your Future*, THE INQUIRER (Sept. 21, 2017), https://perma.cc/64AK-HVEN.

21. *See, e.g.*, Seth Fiegerman, *Facebook Is Well Aware That It Can Influence Elections*, CNN TECH (Nov. 17, 2016), https://perma.cc/8BWK-K8RL; Trevor Timm, *You May Hate Donald Trump. But Do You Want Facebook to Rig the Election Against Him?,* THE GUARDIAN (Apr. 19, 2016), https://perma.cc/J2K6-CGLU. In the case of Facebook, for instance, concerns were raised that the social network prioritizes liberal over conservative news items in its "trending news" section—a major source of news for millions of Americans. Michael Nunez, *Former Facebook Workers: We Routinely Suppressed Conservative News*, GIZMODO (Sept. 5, 2016), https://perma.cc/PHS3-6TFY. Google has also been accused that its "autofill" function is positively biased towards liberal political candidates. Allana Akhtar, *Google Defends Its Search Engine Against Charges It Favors Clinton*, USA TODAY (June 11, 2016), https://perma.cc/L6DQ-WRWN.

algorithms to frame public opinion and communicate to their users information in ways that no other entities could.[22]

The immense role that online data platforms and their algorithms play in the lives of billions of their users is met with growing anxiety. Policymakers and scholars alike deliberate how to curb the power of these companies by subjecting them to transparency and explanation obligations and attempting to impose on them due process requirements. This Article shows, however, that these approaches impose on data platforms unrealistic and futile requirements. Algorithmic decision-making is inscrutable—"rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension."[23] The power of these rules is not to be intelligible and rational from a human perspective, but rather reveal and predict accurate patterns and correlations that exceed human imagination.[24] The core business model of data platforms is to make accurate predictions regarding their users' preferences, even if they do not understand the rationale of these predictions and even if they reveal unpleasant behavioral patterns. Demanding companies to explain or justify in plain language non-causal, non-intuitive, and inscrutable algorithmic decisions thus defies the logic of algorithmic decision-making, contradicts their business models, and often exceeds the companies' operational capacities. Furthermore, imposing on data platform companies intrusive obligations would generate fierce resistance from the companies and is likely to result in low compliance and high enforcement costs. Some of these costs would naturally have to be absorbed by platform users.

Instead of pursuing an adversary and, in all likelihood, ineffective crusade against data platform companies, this Article suggests considering

22. *See* discussion in Part II; *see e.g.,* Jonathan Zittrain, *Engineering an Election: Digital Gerrymandering Poses a Threat to Democracy*, 127 HARV. L. REV. F. 335 (2014) (discussing how data platforms can affect elections); Robert M. Bond et al., *A 61-Million-Person Experiment in Social Influence and Political Mobilization*, 489 NATURE 7415 (2012) (showing how social ads on Facebook affect voting turn out).

23. Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018); *see also,* Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 922 (2017) ("Even when a model is interpretable, its *meaning* may not be clear. Two variables may be strongly correlated in the data, but the existence of a statistical relationship does not tell us if the variables are causally related, or are influenced by some common unobservable factor, or are completely unrelated."); Kiel Brennan-Marquez, *Plausible Cause: Explanatory Standards in the Age of Powerful Machines,* 70 VAND. L. REV. 1249, 1267-68 (2017).

24. Selbst & Barocas, *supra* note 23, at 6.

a collaborative stance. Rather than imposing on these companies requirements that are above and beyond their capacity and interests, they can be taken as allies. The adverse public impacts of private data platforms are the result of these companies' algorithms, but they constitute a much larger societal challenge. It should not be for Facebook or Google or Twitter alone to figure out the proper role of social networking algorithms in times of elections, or the ways to rectify discriminatory patterns that algorithms accurately reify. Demanding these companies to resolve such challenges single handedly means narrowing down the range of available solutions, increasing transaction costs, and engaging in a constant regulatory battle with noncompliant private entities.

Rather, the challenges associated with adverse public impacts of private data platforms shall be examined and addressed by a wide range of stakeholders—governments, organizations, researchers, journalists, etc. As noted by Eric Raymond with regards to the benefits of open source software[25]—"given enough eyeballs, all bugs are shallow."[26] External and collaborative scrutiny of data that is accumulated on data platforms and of some of their algorithms may reveal the causes for specific algorithmic outputs and help address data manipulations, disinformation, biases, and other adverse results.

Access to data held by platform companies can help public and private actors better understand demographic trends, public sentiment, and the geographic distribution of various phenomena. To note just a few examples, the triangulation of official health records and user-generated data on social platforms may reveal common predictors for heart failures or other diseases and thus contribute to the effectiveness of health policies. An analysis of messages shared on social networks related to suicide or mass-violence may help public authorities to better design preventive approaches. Monitoring geolocation data shared by mobile carriers can reveal how residents of specific localities use public and private transport, and enable local authorities to alleviate traffic congestions. Data that is held by platform companies may also unveil public sentiments regarding reforms or policies, issues related to the quality of education or health services, or geolocation information on where people hard-hit by a disease

25. Open source software is software with source code that anyone can inspect, modify, and enhance. *See What Is Open Source*, OPENSOURCE, https://perma.cc/2X8H-AU4V (archived June 25, 2019).

26. Eric S. Raymond, *The Cathedral and the Bazaar*, 3 FIRST MONDAY, Mar. 2, 1998.

or by a hurricane are located. While data platform companies at times sell their users' data to third parties that may triangulate and analyze the data, these endeavors are typically commercial in nature and do not seek to address to public challenges. External scrutiny of this data by a diverse set of actors may also reveal otherwise hidden algorithmic decision rules and outputs.

The recognition of a "global commons of data" does not imply that data platform companies, which play a critical role in aggregating and processing their users' data, would lose their commercial benefits and decision-making prerogatives. It neither implies that access to all such data on the web becomes free and open to all, thus violating users' privacy rights and inflicting potential security damage. Rather, the data commons regime should offer a spectrum of data access modalities—ranging from the most restrictive access rules to the most permissive ones.

The Article outlines five modalities of data access and usage, and provides real-life examples of their implementation: (i) *sharing internal data analysis*, as part of which a data platform company does not share any data with external stakeholders, but rather conducts its analysis in-house and then publicly releases the findings; (ii) *releasing targeted data*, as part of which the data platform company shares with trusted partners subsets of its data for a specific public-regarding purpose (Facebook's initiative on social media and elections would fall under this category); (iii) *participating in data pools*, as part of which several data platform companies and other stakeholders would grant each other access rights to the data to enable collaborative investigation; (iv) *granting access to public actors*, as part of which state statistical agencies or other government stakeholders would get access to data held by platform companies; and (v) *granting open access* to significant subsets of data to the global community, typically under specific terms of use.

How to implement these data sharing modalities in practice? The Article lays out two types of institutional arrangements: a "sticks" approach—invoking the public utilities doctrine, which would oblige data platform companies to provide fair access and use to the user-generated data they hold; and a "carrots" approach—encouraging platform companies to take on the "data commons" idea through financial incentives and soft "naming and shaming" initiatives. The Article suggests several examples on how such incentives and programs could work.

Lastly, the Article addresses three major critiques that can be levied against the global data commons approach. One major pitfall is privacy. The Article discusses technical measures that can be undertaken to protect users' privacy through data de-identification, and the need to employ a relatively restrictive data sharing modality in case that the data in question may impinge users' privacy rights. Another challenge is the question of users' consent: whether and how users' willingness to contribute their data to the global commons can facilitate the implementation of this idea. The commercial interests and incentives of private data companies constitute a third challenge. Being able to effectively aggregate and analyze data is the core business model of platform companies. They invest significant costs in cleansing, structuring, and processing user-generated data. Granting access to this data to external stakeholders may raise significant opposition on the part of data companies. If effective in the long term, such approach may also disincentivize companies from engaging in data collection endeavors. The Article discusses whether a spectrum of data access modalities, coupled with efficient implementation arrangements, can mitigate these concerns. This Article does not purport that these challenges, as well as other hurdles associated with the global data commons proposal, could be fully and satisfactorily resolved. Rather, it seeks to start a conversation on the role of user-generated data in society and the challenges that may be associated with it.

The structure of the Article is as follows. Part II discusses the growing public importance of private data platform companies. Part III outlines existing approaches that aim to curb the adverse implications of algorithmic decision-making and examines the limitations of these approaches. Part IV makes the case for a "global commons of data," delineates the spectrum of data access and usage rules, and discusses the implementation arrangements and challenges embedded in this approach.

II. THE PUBLIC IMPORTANCE OF PRIVATE DATA PLATFORMS

The information that data platforms provide to their users is defined, to a large extent, by the platforms' algorithms.[27] These algorithms do not only

27. *See* Julie E. Cohen, *Law for the Platform Economy*, 51 U.C. DAVIS L. REV. 133, 148 (2017) ("Massively intermediated, platform-based media infrastructures have reshaped the ways that narratives about reality, value, and reputation are crafted, circulated, and

collect and process data,[28] but "create, tap, or steer information flows in ways that suit their goals and in ways that modify, enable, or disable others' agency, across and between a range of older and newer media settings."[29] They do not only link users together, but also suspend them or guide them toward one piece of information and not another. They do not simply circulate images or text, but algorithmically promote some over others.[30]

These algorithms are designed to collect and analyze vast amounts of data,[31] in order to draw inferences about unknown facts from statistical

---

contested."). Algorithms are understood in this context as generalized procedures for turning disorganized data-inputs into manageable outputs through series of logical rules that provide instructions on how to handle data with specific attributes. *See* Mikkel Flyverbom, Anders Klinkby Madsen & Andreas Rasche, *Big Data as Governmentality: Digital Traces, Algorithms, and the Reconfiguration of Data in International Development* (Human Mgmt. Network Research Paper Series No. 42/15, 2015), https://perma.cc/322R-GP46.

28. The procedure of "data processing" can be defined as follows: "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction." *See* Regulation (EU) 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119/1) [hereinafter GDPR].

29. ANDREW CHADWICK, THE HYBRID MEDIA SYSTEM: POLITICS AND POWER 157 (2013). On information flow control by infomediaries, see also Leah A Lievrouw, *New Media, Mediation, and Communication Study*, 12 INFO., COMM. & SOC'Y 303 (2009); Aaron Shaw & Benjamin Mako Hill, *Laboratories of Oligarchy? How the Iron Law Extends to Peer Production*, 64 J. COMM. 215 (2014).

30. Tarleton Gillespie, *Platforms Intervene*, SOC. MEDIA + SOC'Y, Apr.-June 2015.

31. Big data analysis typically relies on three major sources: (i) data that is generated in online activities, including online transactions, web-data trails, e-mail exchanges, videos, photos, search queries, health records, and social networking activities; (ii) personal information from a variety of offline sources: public records (e.g., criminal records, deeds, corporate filings), retailer's sales records, credit agencies, etc.; and (iii) with the advent of the Internet of Things, enormous amounts of information can be collected from the ever-growing number of devices and appliances that have the capacity to record and transmit information about the world. This encompasses information generated by cell phones, surveillance cameras, global positioning satellites, utility-related sensors, communication networks, and phone-booths, among other sources. *See* EXEC. OFFICE OF THE PRESIDENT, PRESIDENT'S COUNCIL OF ADVISORS ON SCI. & TECH., BIG DATA AND PRIVACY: A TECHNOLOGICAL PERSPECTIVE 22-24 (2014), https://perma.cc/LTN5-V37W; VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 99-100 (2013); Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393, 404-5 (2014) ("To obtain their information, data brokers search through government records,

occurrence and correlation, and thus enable predictions about future patterns of behavior and preferences.[32] When enough detail about the past is gathered and processes, the algorithm calculates how different qualities have been correlated with each other in the past. Unforeseen links and correlations surface, and are then used to make projections about events and actions that are likely to happen in the future.[33] Data platform companies employ these algorithms to predict which pieces of information their users would be interested in, and how this information should be presented to them.[34] These algorithms are intensely used in a variety of fields, including the prediction of shopping patterns, students' grades,[35] employees' behavior,[36] heart disease rates,[37] and more.

The political and social implications of such algorithmic prowess are immense. Two major domains where algorithmic decision-making has already had tangible socio-political impacts are (1) democratic processes and elections; and (2) discrimination in the public sphere.

A. Algorithmic Impacts on Democratic Processes and Elections

Facebook by far leads every other social media site as a source of news, with 45% of Americans getting news on the platform.[38] This news consumption largely depends on Facebook's algorithms, which practically determine to which news items different user would be exposed, at what

purchase histories, social media posts, and hundreds of other available sources."); Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 240, 247-50 (2013).

32. MAYER-SCHÖNBERGER & CUKIER, *supra* note 31, at 11-12; Boyd & Crawford, *supra* note 14.

33. Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55, 57 (2013), https://perma.cc/UW9R-YW8G.

34. Karine Nahon, *Where There Is Social Media There Is Politics*, *in* THE ROUTLEDGE COMPANION TO SOCIAL MEDIA AND POLITICS (Axel Bruns et al. eds., 2016); *see also*, John P. Wihbey, *The Challenges of Democratizing News and Information: Examining Data on Social Media, Viral Patterns and Digital Influence* (Shorenstein Ctr. on Media, Politics & Pub. Policy Discussion Paper Series #D-85, 2014), https://perma.cc/F3Z7-JDQN.

35. Jon Marcus, *Here's the New Way Colleges Are Predicting Student Grades*, TIME (Dec. 10, 2014), https://perma.cc/ZGL9-QU8Q.

36. Jack Clark, *Big Data Knows When You're Going to Quit Your Job Before You Do*, BLOOMBERG (Dec. 30, 2014), https://perma.cc/F3C3-E9KB.

37. Elahe Izadi, *Tweets Can Better Predict Heart Disease Rates than Income, Smoking and Diabetes, Study Finds*, WASH. POST (Jan. 21, 2015), https://perma.cc/LZ2C-9WT4.

38. Shearer & Gottfried, *supra* note 9.

time, and in what format. The considerations that underlie these algorithmic decisions are largely obscure. But it is known, for instance, that Facebook presents users with only a small fraction of the information flows created by their friends.[39] Facebook chooses which posts are pushed to the top of a user's "news feed," which posts are located further down, and which are absent from the "feed" all together. Further, the platform is known to prioritize contents with which the user is more likely to agree,[40] thus contributing to the generation of "information bubbles," which prevent exposure to alternative view-points and deepen one's assertion that hers is the only right position.[41]

Studies have indeed shown that Facebook's algorithms could strategically encourage or discourage users to vote in hotly contested states, thus tilting the results of an election. A controversial social experiment that was conducted in 2010 examined the voting patterns of 61 million American Facebook users, who accessed the platform on the day of the 2010 congressional elections.[42] The researchers concluded that a "social" Facebook message that encouraged users to vote caused an additional 340,000 votes to be cast amidst the 82 million Americans who voted that day. While these figures may appear modest, George W. Bush won the 2000 U.S. presidential election by taking over Florida, where he beat Al Gore by 537 votes.[43] In 2016, Hillary Clinton lost key states such as Pennsylvania by 44,292 votes, Michigan by 10,704 votes, and Wisconsin by 22,748 votes.[44]

---

39. Josh Constine, *Why Is Facebook Page Reach Decreasing? More Competition and Limited Attention*, TECHCRUNCH (Apr. 4, 2014), https://perma.cc/S5LP-F5RX.

40. ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK (reprt. 2012) (2011).

41. *Id*. *See generally* CASS R. SUNSTEIN, REPUBLIC.COM 2.0 (2007).

42. Bond et al., *supra* note 22. Nearly 60 millions of users were shown a graphic within their news feeds with a link to locate their polling place, a button to click "I voted," and the profile pictures of their friends who had indicated that they had already voted. Other 611,000 users only received an "informational message" with polling station details, but were not shown the "social" graphic that included references to friends. The researchers then compared the groups' online behaviors, and matched 6.3 million users with publicly available voting records from precincts across the country. The results revealed that those who got the informational message voted at the same rate as those who saw no message at all. But those who saw the social message were 2% more likely to click the "I voted" button than those who received the informational message, and 0.4% more likely to head to the polls than the other group.

43. Robinson Meyer, *How Facebook Could Tilt the 2016 Election*, THE ATLANTIC (Apr. 18, 2016), https://perma.cc/RP58-LVCL.

44. *Presidential Elections Results: Donald J. Trump Wins*, N.Y TIMES (Aug. 9, 2017),

Most recently, the vast political value of Facebook profiles and connections received worldwide recognition in the wake of the Cambridge Analytica data abuse scandal. Allegedly, Cambridge Analytica—a private data mining and voter-profiling company based in the United Kingdom—harvested through various means personally identifiable data of 87 million Facebook users. The company then used the data to generate psychological profiles of US voters and targeted them with personalized political advertisements in support of Donald Trump and other politicians.[45] There is no direct evidence regarding the contribution of Cambridge Analytica's approach to Trump's electoral victory. There is no doubt, however, that Facebook and other data platforms have turned into crucial political arenas.[46]

Google's algorithmic capacity to frame information and shape behavior is no less formidable. Visibility is key for online presence, and the position of an information item in Google's search results thus plays a definitive role in determining whether this item will ever be noticed.[47] In August 2018, for instance, President Trump accused Google of "an effort to intentionally suppress conservative news outlets supportive of his administration,"[48] by ranking them lower than news from "authoritative" and mainstream news sources on Google News.[49]

In recent years, there has been a growing interest in studying the potential biases of search engine rankings.[50] For instance, to test the

---

https://perma.cc/2MRB-3WZR.

45. *The Cambridge Analytica Files: The Story So Far*, THE GUARDIAN (Mar. 26, 2018), https://perma.cc/Q5W6-G6PX.

46. *See, e.g.*, Zittrain, *supra* note 22.

47. A recent analysis of about 300 million clicks on Google found that 91.5% of those clicks were on the first page of search results, with 32.5% on the first result and 17.6% on the second. The study also reported that the bottom item on the first page of results drew 140% more clicks than the first item on the second page. *See* CHITIKA, THE VALUE OF GOOGLE RESULT POSITIONING (2013), https://perma.cc/5ZQE-PYAQ.

48. Adam Satariano, Daisuke Wakabayashi & Cecilia Kang, *Trump Accuses Google of Burying Conservative News in Search Results*, N.Y. TIMES (Aug. 28, 2018), https://perma.cc/9NZM-M38R.

49. "In a statement, Google said that its search service was 'not used to set a political agenda and we don't bias our results toward any political ideology.'" *Id.*

50. ALEX HALAVAIS, SEARCH ENGINE SOCIETY 85 (2009) ("In the process of ranking results, search engines effectively create winners and losers on the web as a whole."); *see also* SIVA VAIDHYANATHAN, THE GOOGLIZATION OF EVERYTHING (2010); Abbe Mowshowitz & Akira Kawaguchi, *Measuring Search Engine Bias*, 41 INFO. PROCESSING & MGMT. 1193 (2005); Herman Tavani, *Search Engines and Ethics*, *in* THE STANFORD ENCYCLOPEDIA OF

political significance of Google rankings, Epstein and Robertson conducted a randomized controlled trial, as part of which they asked participants unaware of the political candidates in an election to search for the candidates and form an opinion based on the results. By biasing the search results in a controlled manner (placing links focused on one candidate above another), they showed that the mere order of search rankings could affect by 20 percent or more the voting preferences in an election.[51] Furthermore, such rankings can be masked so that voters are unaware of the manipulation.

B. Discriminatory Impacts of Algorithms

The ultimate objective of search engine rankings, ads, or the contents of one's news feed is to correctly predict the specific individual preferences of each particular user. Accuracy—rather than political correctness, affirmative action, or attempts to educate the user to be a better person—is key in this respect. Thus, the goal of an accurate algorithm is to distinguish variations among different groups of users and provide these groups with distinct results. For instance, advertisers are able to target people who live in low-income neighborhoods with high-interest loans.[52] An analysis of The Princeton Review's Prices for online SAT tutoring shows that customers in areas with a high density of Asian residents are often charged more.[53] Shall these distinctions be treated as discrimination or normal market practices?

Algorithmic discrimination occurs when certain groups or individuals unfairly receive unfavorable treatment as a result of algorithmic decision-making.[54] Bias can enter into algorithmic systems regardless of the intent of

---

PHILOSOPHY (Edward N. Zalta ed., 2014); S. Fortunato, A. Flammini, F. Menczer & A. Vespignani, *Topical Interests and the Mitigation of Search Engine Bias*, 103 PROC. NAT'L ACAD. SCI. (PNAS) 12684 (2006).

51. Robert Epstein & Ronald E. Robertson, *The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections*, 112 PROC. NAT'L ACAD. SCI. (PNAS) E4512 (2015).

52. Claire Cain Miller, *When Algorithms Discriminate*, N.Y. TIMES (July 9, 2015), https://perma.cc/U3EF-B8D7.

53. Julia Angwin & Jeff Larson, *The Tiger Mom Tax: Asians Are Nearly Twice as Likely to Get a Higher Price from Princeton Review*, PROPUBLICA (Sept. 1, 2015), https://perma.cc/4Z6V-R8VL.

54. Bryce W. Goodman, *Economic Models of (Algorithmic) Discrimination*, PROC. 29TH CONF. ON NEURAL INFO. PROCESSING SYS. (NIPS), Dec. 5-10, 2016, https://perma.cc/TZC4-59M6.

the provider, and create discriminatory feedback loops.[55] Big data algorithms then reify existing patterns of discrimination—if they are found in the dataset, then by design an accurate classifier will reproduce them. Discrimination occurs because the data being mined is itself a result of past intentional discrimination, and there is frequently no obvious method to adjust historical data to rid it of this taint. In this way, discriminatory decisions can be an outcome of a "neutrally" designed algorithm.

For instance, Latanya Sweeney identified "significant discrimination" in the online ads that appeared following Google searches for black-identifying names versus white-identifying names. Searches for black names were much more likely to return advertisements for arrest records.[56] Sweeney argues that this is a function of Google's AdSense algorithm, which takes user feedback into account to determine which terms are more likely to attract user interest. In a similar manner, researchers discovered that Google's AdSense showed an ad for high-income jobs to men much more often than it showed the ad to women.[57] Google Images search for "C.E.O." produced 11 percent women, even though 27 percent of United States chief executives are women.[58] In fact, Google learned society's racism and discriminatory patterns and fed it back to users.

..*

The algorithmic prowess of data platforms—from shaping one's news diet and encouraging specific voting patterns to generating discriminatory social results—has raised significant anxiety among policymakers and

55. A growing legal debate has emerged in the past years to identify and cope with this challenge. *See, e.g.*, FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? UNDERSTANDING THE ISSUES (2016), https://perma.cc/C2EU-P5P7; VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR (2018); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016) (focusing on the case of employment discrimination); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109 (2018); Zaynep Tufekci, *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*, 13 COLO. TECH. L. J. 203 (2015).

56. Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 COMM. ACM, May 2013, at 44.

57. Amit Datta, Michael Carl Tschantz & Anupam Datta, *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination*, 2015 PROC. ON PRIVACY ENHANCING TECH. Apr. 2015, 92-112, https://perma.cc/H5T2-396G.

58. Matthew Kay, Cynthia Matuszek & Sean A. Munson, *Unequal Representation and Gender Stereotypes in Image Search Results for Occupations*, ACM CHI CONF. ON HUM. FACTORS COMPUTING SYS., Apr. 18, 2015, at 3819, https://perma.cc/TG4A-H4YD.

scholars in recent years.[59] These concerns are aggravated by the immense information asymmetry between data platform companies and their users, and the obscure nature of the algorithms' operation. As noted by Richards and King, "[w]hile big data pervasively collects all manner of private information, the operations of big data itself are almost entirely shrouded in legal and commercial secrecy."[60] Decisions taken by these algorithms on how search findings should be ranked or how information shared by users should be prioritized are "dynamic, all but invisible, and individually tailored."[61] As algorithms become increasingly autonomous and invisible, it becomes even harder for the public to detect and scrutinize their impartiality status and the considerations (or lack thereof) that affect their operation mode.

The question of how to mitigate the adverse social and political implications of algorithmic decision-making has thus become central in legal and policy debates.[62] Part III of this Article discusses some of the key approaches that have been brought forward by policymakers and scholars to regulate data platform companies and make their algorithms more accountable, and shows that these approaches do not achieve their stated objectives.

---

59. *See generally* Lee Rainie & Janna Anderson, *Code-Dependent: Pros and Cons of the Algorithm Age*, PEW RES. CTR. (Feb. 9, 2017), https://perma.cc/5WWY-H528 (summarizing examples of the risks of using algorithms broadly).

60. Neil M. Richards & Jonathan H. King, Three Paradoxes of Big Data, 66 STAN. L. REV. ONLINE 41, 42 (2014), https://perma.cc/5MVV-8YLG.

61. Tufekci argues that this makes algorithmic decision-making different from traditional editorial decisions taken by newspaper editors or TV broadcasters, who also have wide margins of discretion, but their decisions and potential biases are visible to their readers or viewers. *See* Tufekci, *supra* note 55.

62. *See, e.g.*, ROB KITCHIN, THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES AND THEIR CONSEQUENCES (2014); CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION (2016); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Michael Ananny, *Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness*, 41 SCI. TECH. & HUM. VALUES 93 (2015); David Beer, *The Social Power of Algorithms*, 20 J. INFO. COMM. & SOC'Y 1 (2016); Taina Bucher, *'Want to Be on the Top?' Algorithmic Power and the Threat of Invisibility on Facebook*, 14 NEW MEDIA & SOC'Y 1164 (2014); Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1 (2016); Danielle Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Kate Crawford, *Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics*, 41 SCI. TECH. & HUM. VALUES 77 (2016); Nicholas Diakopoulos, *Algorithmic Accountability: Journalistic Investigation of Computational Power Structures*, 3 DIGITAL JOURNALISM 398 (2015); Rob Kitchin, *Thinking Critically About and Researching Algorithms*, 20 INFO. COMM. & SOC'Y 14 (2016).

III. REGULATING DATA PLATFORMS: CURRENT APPROACHES

The technological complexity of big data algorithms has far exceeded legal response capacity. Several directions to curb the immense power of platform companies and mitigate the adverse impacts of their algorithmic decision-making procedures have started to emerge. These focus in particular on measures to prevent manipulation and misinformation, discriminatory patterns and biases that result from algorithmic decision-making. These approaches can be divided into two main categories: (a) algorithmic transparency and explanation; and (b) due process. The subsequent sections discuss each of these approaches and examine their effectiveness.

A. Algorithmic Transparency and Explanation

A common and seemingly intuitive approach to neutralize the negative consequences of obscure algorithms is to shed light on the operation of these algorithms. As part of this rationale, platform companies would be expected to release the source code of their algorithms; explain how websites, people, and events are rated and ranked; or describe how and why one's "news feed" or search results are structured and populated. Several scholars have suggested that such an approach could enable users to properly manage expectations regarding the contents of their online information diet and better understand potential biases or inconsistencies.[63]

The General Data Protection Regulation (GDPR),[64] which was issued by the European Parliament in 2016 and came into force in May 2018, adopted a variation of the algorithmic transparency approach by granting EU residents a novel legal protection tool—a "right to explanation."[65] According

63. PASQUALE, *supra* note 62, at 3-11; Citron & Pasquale, *supra* note 62, at 13-16; Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014).

64. GDPR, *supra* note 28.

65. *Id.* at art. 13. It is commonly assumed that the GDPR may influence global standards related to data platforms. *See, e.g.*, Michael D. Birnhack, *The EU Data Protection Directive: An Engine of a Global Regime,* 24 COMPUTER L. & SECURITY REP. 508 (2008); Sean Michael Kerner, *HPE Explains What European GDPR Privacy Regulations Mean to U.S. Firms*, EWEEK (May 1, 2017), https://perma.cc/LM8H-BGYW ("[T]he GDPR applies to anyone that is doing business in the EU, so anyone selling into it or has [sic] employees

to the Regulation, companies that collect or process personal data of EU residents have a legal duty to provide their users with a "meaningful explanation" on how their algorithms reach decisions. Such explanation may either refer to the algorithmic model or system that produced a certain decision (i.e., the logic, significance, envisaged consequences, and general functionality of an automated decision-making system),[66] or explain a specific decision or query (i.e., specific decision-rules and rationale).[67] Several platform companies already follow this path and explain some of the considerations that go into the operation of their algorithms.[68]

As Ananny and Crawford have observed, transparency interventions are "commonly driven by a certain chain of logic: observation produces insights which create the knowledge required to govern and hold systems accountable."[69] Fung et al. refer to this process as a "transparency action cycle," which generates behavioral changes among both platform companies and their users.[70] For example, regulation that requires restaurants to place on their windows their hygiene rankings (as

there.").

66. This would include setup information (the objectives behind the modelling process, the parameters used for the model, etc.), training metadata (summary statistics and qualitative descriptions of the input data used to train the model, and the model predictions), performance metrics (information on the model's predictive skills), process information (how the model was tested and how undesirable elements were screened out), etc. *See* Lilian Edwards & Michael Veale, *Slave to the Algorithm? Why a 'Right to Explanation' Is Probably Not the Remedy You Are Looking For,* 1 DUKE L. & TECH. REV. 19, 55-56 (2017)

67. *Id*. *See also* Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76 (2017).

68. *See, e.g.*, *How News Feed Works*, FACEBOOK, https://perma.cc/3HLH-2BHW (archived June 25, 2019); *How Search Algorithms Work*, GOOGLE SEARCH, https://perma.cc/8PBK-YEWP (archived June 25, 2019).

69. Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973, 974 (2018).

70. ARCHON FUNG, MARY GRAHAM & DAVID WEIL, FULL DISCLOSURE: THE PERILS AND PROMISE OF TRANSPARENCY (2007). The action cycle consists of four elements: *first*, a transparency policy compels an information provider (e.g., private company, public authority, etc.) to provide salient and accessible information to at least one group of information users; *second*, this information causes users to change their behavior vis-à-vis the information provider; *third*, information users' actions affect information providers in a meaningful manner, and the latter cannot ignore such actions; and *fourth*, information providers change their behavior in response to information users' actions. *See also* Steven Kosack & Archon Fung, *Does Transparency Improve Governance?* 17 ANN. REV. POL. SCI. 65 (2014).

determined by municipal sanitary inspections) provides customers with valuable and easily comprehensible information, which allows them to opt for clean restaurants instead of dirty ones. This behavior cannot be ignored by restaurant owners. In response to it, those with lower rankings are compelled to invest more efforts in improving their hygiene practices to attract customers. In theory, a similar logic could be applied to the algorithms employed by data platform companies—platform companies would explain to their users how their algorithms work, and users would rely on these explanations to decide whether they want to consume the services of the companies. To avoid the loss of customers, platform companies would then, theoretically, adapt their algorithms to users' needs and priorities. The practical implementation of this approach is, however, precarious.

Algorithmic transparency could certainly shed some light on otherwise obscure data processing practices. But algorithmic transparency policies are prone to a range of implementation difficulties, related to the consistency of information collection standards, cognitive limitations of information suppliers and users, and context-related sensitivities.[71] For one, the assumption that algorithmic decision-making can be "explained" in plain language is problematic.[72] Algorithmic outputs are not based on causal relations between variables that can be meaningfully rationalized. Rather, they operate based on correlations among variables that seemingly have no connection to each other but nonetheless have some predictive value. As explained by Edwards and Veale:

> LinkedIn, for example, claim to have over 100,000 variables held on every user that feed into [algorithmic] modelling. Many of these will not be clear variables like 'age,' but more abstract ways you interact with the webpage, such as how long you take to click, the time you spend reading, or even text you write but later delete without posting. These variables may well hold predictive signals about individual characteristics or behaviors, but we lack compelling

71. Daniel E. Ho, *Fudging the Nudge: Information Disclosure and Restaurant Grading*, 122 YALE L. J. 574 (2012); Jennifer Shkabatur, *Transparency With(out) Accountability: Open Government in the United States*, 31 YALE L. & POL'Y REV. 79 (2012).

72. Mike Ananny & Kate Crawford, *supra* note 69, at 974-77; Selbst & Barocas, *supra* note 23, at1085.

> ways to clearly display these explanations for meaningful human interpretation.[73]

A requirement to explain the workings of an algorithm is thus prone to a host of technical difficulties and is, in essence, unrealistic.[74] Explanations would either be generalist and vague (and as such fail to provide practical value) or delve into technical details that lack context or causality.[75]

Algorithmic transparency and explanation obligations could also inflict commercial damage on companies that invest ample resources in developing complex data mining algorithms in order to stand out from their competitors.[76] Google's ranking algorithm is one of its most coveted commercial assets, and an obligation to disclose parts of it can constitute a violation of its proprietary methods and trade secrets. Further, making publicly available the factors crucial for certain scoring techniques might provide opportunities for those scored to act strategically and "game the system," thus undermining the credibility of the algorithms, and eventually rendering them inaccurate.[77] For instance, the process for deciding which tax returns to audit or whom to grant financial benefits may need to be partly opaque to prevent tax cheats or credit to insolvent customers.

Furthermore, when an explanation of how a rule operates requires disclosing private or sensitive data (e.g., in adjudicating a commercial offer of credit, a lender reviews detailed financial information about the applicant), disclosure of the data may be undesirable or even legally barred.[78] Cognizant of these limitations, the "right to explanation" in the GDPR contains multiple exemptions that significantly narrow down cases in which platform companies may be mandated to provide an explanation.[79]

---

73. Edwards & Veale, *supra* note 66, at 59-60.

74. *Id.* at 59 ("[I]n some systems there is no theory correlating input variables to things humans understand as causal or even as 'things.' In [machine-learning] systems . . . the features that are being fed in might lack any convenient or clear human interpretation in the first place").

75. *Id.*

76. Citron & Pasquale, *supra* note 62, at 17.

77. *Id.* at 20, 26.

78. Joshua Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 638-39 (2017).

79. Exemptions include "carve-outs for intellectual property (IP) protection and trade secrets; restriction of application to decisions that are 'solely' made by automated systems; restriction to decisions that produce 'legal' or similarly 'significant' effects." Edwards & Veale, *supra* note 66, at 21. Doubts have also been raised regarding the binding legal nature of the "right to explanation" due to its inclusion in the Recitals of the

This naturally limits the scope and effectiveness of the newly created entitlement even before the practicality of the right to explanation is considered.

The fallacy of the algorithmic transparency approach was recently confirmed by Brauneis and Goodman, who used freedom of information legislation and other state laws to request explanations regarding the operations of algorithmic decision-making in local governments. The authors filed 42 open records requests in 23 states, seeking essential information about six predictive algorithm programs.[80] The results of these requests were largely unsatisfactory.[81] The authors identified two impediments that hinder the algorithmic transparency proposition: first, governments lack appropriate records and information regarding the operation of their algorithms, and are thus unable to release more information than they have;[82] and second, even where governments do have key explanatory records, they may refuse to disclose them in deference to the claims of private vendors that this information is confidential and would violate their trade secrets.[83]

B. Due Process

The introduction of "due process" is another mitigation measure that scholars brought forward to mitigate adverse impacts of algorithmic decision-making.[84] Several directions have been proposed.

One category of approaches puts the emphasis on *ex-ante* procedural regularity and fairness of algorithmic decision-making. For instance, Kroll et al. contend that technical methods[85] can verify that

- The same policy or rule was used to render each decision.

Directive, and not in the main text. *See* Wachter, Mittelstadt & Floridi, *supra* note 67.

80. Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J. L. & TECH. 103, 103-4 (2018).

81. *Id.*

82. *Id.* at 152-53.

83. *Id.* at 153-59.

84. *See generally*, Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008) (on how due process might be extended in the Big Data era).

85. These may include software verification, requirements for systems to create cryptographic commitments as digital evidence of their actions, or analysis of random choices and outputs of a software.

- The decision policy was fully specified (and this choice of policy was recorded reliably) before the particular decision subjects were known, reducing the ability to design the process to disadvantage a particular individual.
- Each decision is reproducible from the specified decision policy and the inputs for that decision.
- If a decision requires any randomly chosen inputs, those inputs are beyond the control of any interested party. [86]

The practical implementation of this approach may raise difficulties. Ensuring procedural regularity and fairness may safeguard from overt mistakes or miscalculations. It may also rectify specific individual misjustice (for instance, a wrongful calculation of one's credit score or insurance rates). But it is unlikely to address—or even unveil—cases in which the algorithm simply reflects widespread societal biases or existing discrimination. For instance, Google search results that associate "black" names with incarceration do not, in all likelihood, result from an intentional or wrongful algorithmic design that seeks to discriminate against African-American users. Rather, they reflect an existing discriminatory pattern in society.

One potential way to overcome this challenge is to inspect biases in algorithmic decision-making through random audits. This approach could identify discriminatory results even if they result from fair algorithmic decision rules. For instance, Christian Sandvig et al. suggest to undertake regular audits that would identify discriminatory or biased algorithmic outputs, [87] and then employ data mining tools to cleanse the algorithmic operation.[88] The problem with this approach, however, is that it relies on

86. Kroll et al., *supra* note 78, at 657.

87. For instance, audits can be carried out to observe whether algorithmic decision-making results in discriminatory or biased outcomes. *See* Christian Sandvig, Kevin Hamilton, Karrie Karahalios & Cedric Langbort, *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, DATA & DISCRIMINATION, May 22, 2014, at 8-18, https://perma.cc/88S4-H5ZN.

88. This can be done, for instance, by cleaning data "in such a way that legitimate classification rules can still be extracted but discriminating rules based on sensitive attributes cannot." *See* Sara Hajian, Josep Domingo-Ferrer & Antoni Martinez-Balleste, *Discrimination Prevention in Data Mining for Intrusion and Crime Detection*, 2011 IEEE SYMP. COMPUTATIONAL INTELLIGENCE CYBER SECURITY (CICS) (April 2011), at 47, https://perma.cc/9H8U-THKA. *See generally* Rupanjali Dive and Anagha Khedkar, *An*

audits of random dataset samples, but cannot account for potentially different findings that can result from a different set of samples.

A distinct approach has recently been brought forward by Selbst and Barocas, who emphasize the need to inspect the projected impact of an algorithm. They suggest to require data platform companies to prepare and publicly release algorithmic "impact statements,"[89] which would be akin to environmental impact statements, required by the National Environmental Protection Act.[90] This approach seems promising as it does not intend to inspect inscrutable algorithms, but rather targets the heart of the problem—the impacts of these algorithms. These may be harmful and discriminatory for reasons that are beyond human comprehension, even if the algorithmic design is perfectly fair and neutral. However, the implementation of this proposal may raise technical challenges, as it would require significant compliance and oversight costs to ensure that companies have the capacity to accurately assess the projected impacts of their algorithms and properly report this as part of their statements.

While all these approaches take an *ex-ante* stance to algorithmic decision-making, the European Union's GDPR suggests imposing on platform companies *ex-post* due process obligations. For instance, Article 22 of the GDPR gives a data subject a "right not to be subject to a decision based solely on automated processing, including profiling"[91] that "significantly affects" the data subject.[92] Exceptions to this provision must be based on authorization by a "Union or Member State law" that also "lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests".[93] By way of remedy, Article 22 provides affected users with a right to "obtain human intervention on the part of the [data platform],

Approach for Discrimination Prevention in Data Mining, 3 INT'L J. APPLICATION OR INNOVATION ENGINEERING & MGMT., June 2014, at 52.

89. Selbst & Barocas, *supra* note 23, at 1134 (the authors define such a statement as "a document designed to explain the process of [algorithmic] decision-making and the anticipated effects of that decision in such a way as to open the process up to the public.").

90. 42 U.S.C. § 4332(C) (2012).

91. Profiling is defined in the Directive as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person." GDPR, *supra* note 28, art. 4(4).

92. The question of what constitutes a "significant effect" is open for interpretation, and it remains to be seen how Member States approach it.

93. GDPR, *supra* note 28, art. 22(2)(b).

express their point of view, and contest the decision."[94] This essentially trades away the right to a human decision-maker for maintaining a right for a "human hearer," and secures some form of due process. It remains to be seen how such due process would take place in practice in the different Member States, and what would be the implications of contesting an algorithmic decision—would it secure a compensation or a right to require the correction of the algorithmic decision?

Jack Balkin, for instance, advocates for the recognition of "algorithmic nuisance,"[95] which would apply in cases where platform companies use algorithms to "make judgments that construct people's identities, traits, and associations that affect people's opportunities and vulnerabilities."[96] Recognizing the social costs that users may incur due to algorithmic decision-making, this approach suggests to impose on platform companies fiduciary obligations and thus pay for the negative externalities of their data collection and analysis.[97]

The idea of algorithmic nuisance builds on an analogy to the common law concepts of public and private nuisance. However, the application of the nuisance doctrine to discrimination in automated search engines or news feed results is only likely to offer recourse in the most extreme cases, where tangible damage to the plaintiff can be proved. But such cases are likely to be rare. Given the inscrutability of big data algorithms, causality between an algorithmic decision system and the nuisance caused to the plaintiff would be difficult to establish.

Balkin's proposal takes a step forward compared to the GDPR's due process approach, since it advocates for compensation to affected users, and not only entitles them to "contest" a decision. But both approaches suffer from a similar limitation. They focus on individual—rather than societal—algorithmic impacts. Even if some form of compensation could be provided to individuals that are unjustly harmed by an algorithm, this would only resolve an individual problem of a person who was sufficiently knowledgeable and willing to complain. This approach would not address wider societal impacts that may result from adverse algorithmic decisions.

---

94. *Id.* at art. 22.

95. Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. DAVIS L. REV 1149, 1163 (2018).

96. *Id*.

97. *Id*.

C. The Inherent Capacity Limitations of Data Platforms

The operation model of data platforms is to accurately identify and predict patterns of users' activities, interests, and preferences. This means that effective big data algorithms may simply reify existing patterns of discrimination in society—if they are found in the dataset, then by design an accurate algorithm will reproduce them. This may well result in discriminatory scoring, ranking, or profiling decisions; politically biased "news feeds;" or misleading ads and recommendations. However, predicting behavioral patterns is different from explaining them, understanding their projected impacts on distinct user groups, or incurring liability for their potential (and unpredicted) implications.

Current approaches to diminish the adverse impacts of algorithmic decision-making—by enhancing transparency, introducing explanations, or observing due process requirements—delineate distinct action paths, but share a major commonality. They demand private companies identify, explain, and ultimately prevent non-causal, non-intuitive, and inscrutable algorithmic decisions, which may in fact accurately reflect existing social biases. Companies are thus expected to invest tremendous resources and efforts in endeavors that are well beyond their business models, capacities, or aspirations.

It may feel intuitively appealing and just to impose such public-regarding obligations on dominant, monopoly-like data platforms such as Facebook, Twitter, or Google. But the larger implications of such requirements are not necessarily positive. For one, these approaches neglect the question of whether data companies possess the internal capacity to comply with the suggested obligations, and what it would take them to develop such capacity. Instead of internalizing high compliance costs, dominant data platforms are likely to impose them on their consumers—in the form of usage fees or more aggressive targeted ads. For smaller companies or startups that build their business models based on the aggregation and analysis of user-generated data, intrusive and costly regulation may turn into a deterrent that rules out their possibility of entering the market.

The feasibility and adverse market implications of transparency, explanation, or due process obligations is not the only problem with the suggested approaches. Scholars suggest that regulation or even mere regulatory scrutiny of big data algorithms could raise legitimate First Amendment concerns. Bambauer, for instance, argues that since the First

Amendment protects the right to create knowledge, data is speech.[98] If accepted, such an understanding makes any interference in data collection or processing constitutionally problematic. Indeed, while the Federal Trade Commission has begun to explore questions of platform governance, it has not gone beyond encouraging companies to "do more to determine whether their own data analytics result in unfair, unethical, or discriminatory effects on consumers."[99]

In sum, given the complexity and obscurity of big data algorithms, the efficacy of regulatory attempts naturally depends on the capacity and willingness of data platforms to comply with them. Unsurprisingly, data platform companies do not welcome such regulatory interventions and often meet them with vehement opposition and counter-lobbying efforts.[100] Rather than investing mutual efforts in resolving common challenges, data platform companies and policy-makers find themselves immersed in bitter battles in Congress or in courts.[101] Such battles would become even worse if concrete and intrusive regulation is at stake.

IV. THE CASE FOR A GLOBAL COMMONS OF DATA

A. Making the Case

Data that each of us generates and leaves behind on social networks, search engines or with mobile carriers includes messages and location traces left by users of social networks, cell phone signals that indicate prior and current physical location, online search queries, and other digital traces of online activities. This data has a tremendous value for modern society, and it fuels significant sectors of economic, social, and political activity.

---

98. Jane R. Bambauer, *Is Data Speech?*, 66 STAN. L. REV. 57, 86-110 (2014). *See also* Bracha & Pasquale, *supra* note 8, at 1188-201 (discussing the impact of the First Amendment on the regulation of search engines' results).

99. FTC Commissioner Julie Brill, Scalable Approaches to Transparency and Accountability in Decisionmaking Algorithms, Remarks at the NYU Conference on Algorithms and Accountability (Feb. 28, 2015), https://perma.cc/2JHT-9L5A.

100. Olivia Solon & Sabrina Siddiqui, *Forget Wall Street—Silicon Valley is the New Political Power in Washington*, THE GUARDIAN (Sept. 3, 2017), https://perma.cc/593Q-HC5W.

101. *See, e.g.*, Michael Gaynor, *Telecom Lobbyists Have Stalled 70 State-Level Bills That Would Protect Consumer Privacy,* MOTHERBOARD (Aug. 6, 2018), https://perma.cc/677H-WN6S; Hamza Shaban, *Google for the First Time Outspent Every Other Company to Influence Washington in 2017*, WASH. POST (Jan. 23, 2018), https://perma.cc/J9LC-FKP9.

While some databases are purely local, containing, for example, information about the inhabitants of a specific municipality, most databases are likely to consist of data collected from numerous local and foreign sources. As noted by Benvenisti, "[t]he data has accumulated over years by the input of billions of users, domestic and foreign alike. Each click, like each drop of rain filling up a reservoir, adds to immense reserves of human knowledge. Just like a giant global lake or a vast international river of knowledge, private and public data banks constitute a new manifestation of the common heritage of humankind."[102]

Currently, calls to require platform companies to share data with users are typically focused on personal data—the ability of an individual user to retrieve the personal data that was collected about her by the data company. In the European Union, the GDPR grants data platform users the rights to access the data that is collected about them, obtain explanation regarding the purpose of the data processing and the third-party recipients to whom the data is disclosed (art. 12); rectify any data inaccuracies (art. 13); demand the erasure of personal data (art. 14), and restrict the processing of personal data under certain conditions.[103] Although the United States lacks an omnibus notion of data protection laws, similar rights emerged in relation to credit scoring in the Fair Credit Reporting Act of 1970.[104]

The concept of providing users access to their own data is grounded in both efficiency and fairness rationales and seems to be straightforward. As Tene and Polonetsky argue, the ability to access and reuse personal data can fuel a whole economy of third-party applications that combine personal data from different sources and services into a single user experience that enables anyone to easily analyze one's data and draw useful conclusions (e.g., change one's diet or improve one's sleeping patterns).[105] Sharing such data in real time and in machine-readable formats could be even more promising in terms of the data usability and usefulness for individuals.

---

102. Eyal Benvenisti, *Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance,"* 29 EUR. J. INT'L L. 9, 81 (2018). On the common heritage idea, see generally Surabhi Ranganathan, *Global Commons*, 27 EUR. J. INT'L L. 693, 704-11 (2016).

103. GDPR, *supra* note 28.

104. *See* Citron & Pasquale, *supra* note 62, at 16.

105. *See generally* Tene & Polonetsky, *supra* note 31, at 263-68.

A similar rationale could support the idea of sharing user-generated data with a broader set of stakeholders. Scholars stress the importance of public access to data to our collective pursuit of knowledge, paying particular attention to the value of recombining data from distinct sources in ways that make the sum worth more than its individual ingredients.[106] Access to data can thus be key for public decision-making: it may supplement scant public statistics and inform policy interventions that would resolve public problems or expedite emergency response. As indicated by Yakowitz, "nearly every recent public policy debate has benefited from mass dissemination of anonymized data."[107]

The benefits of open access to data have been acknowledged by dozens of governments around the world and fueled the establishment of the Open Government Partnership, which supports governments in the implementation of open government programs. In an Executive Order issued in 2013, President Obama acknowledged that "making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans' lives and contributes significantly to job creation."[108] He therefore ordered that "the default state of new and modernized Government information resources shall be open and machine readable."[109] In 2016, the leaders of G8 countries committed to the "open by default" principle, which requires that all government data will be published openly by default and that such data would only be kept closed if there are special considerations that prevent its disclosure.[110] The EU has also embarked on an effort to create a "Digital

106. Christine Borgman, *The Conundrum of Sharing Research Data*, 63 J. AM. SOC'Y INFO. SCI. & TECH. 1059, 1070 (2012) ("Indeed, the greatest advantages of data sharing may be in the combination of data from multiple sources, compared or 'mashed up' in innovative ways."); Michael Mattioli, *The Data-Pooling Problem,* 32 BERKELEY TECH. L. J. 171, 194-204 (2017); Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J. L. & TECH. 1, 8-10 (2011).

107. Yakowitz, *supra* note 106, at 9-10 (referring to data released by the Federal Financial Institutions Examination Council that enables to detect housing discrimination and contribute to public deliberations on home mortgage policies; data collected by Medicare and Medicaid, which significantly contributed to President Obama's health care reform; census data that has been critical to detect racial segregation; public crime data that has revealed unequal allocation of police resources; and more).

108. Exec. Order No. 13,642, 3 C.F.R. 13642 § 1 (2013), https://perma.cc/9YC7-M5PG.

109. *Id.*

110. G8 Open Data Charter and Technical Annex, U.K. CABINET OFF. (June 18, 2013),

Single Market," that is designed "to fully unleash the data economy benefits."[111]

While open data is becoming more common in government, academic and institutional settings, this kind of data availability has not yet been taken up by private companies. The collection and analysis of user-generated data are nowadays chiefly undertaken for commercial purposes. Besides a few isolated and self-proclaimed "data philanthropy" initiatives and other corporate data-sharing collaborations, data platform companies have historically shown resistance to releasing their data to the public. Third parties commonly obtain access to such data through private agreements with the data platforms.[112]

Time has come to reconsider this approach. The societal value of the data held by platform companies is enormous, and the benefits of a collaborative access and usage rights to this data are immense. User-generated data shall thus be part of a "global data commons," [113] which would be responsibly managed in a manner that contributes both to the business models of platform companies and to larger societal objectives. External stakeholders would then be able to analyze this data to identify and prevent algorithmic biases or discriminatory outputs, as well as address a variety of other public problems and challenges.

The global data commons regime should not be restricted to the largest internet monopolies, such as Google, Facebook, Twitter, or Amazon. Rather, any company that operates an online platform that collects, aggregates, and processes its users data shall be encouraged to take part in the data commons regime. The breadth of the commons regime could be inspired by

---

https://perma.cc/PBH7-DNLQ.

111. *Free Flow of Non-Personal Data*, EUR. COMMISSION, https://perma.cc/4XRX-AM8M (archived June 25, 2019).

112. *See* discussion *infra* Part IV.B.

113. "Commons" refers to institutionalized arrangements of community management or governance of shared resources. The concept of commons governance was developed by the late Nobel laureate, Elinor Ostrom, in the context of natural resources and later extended to a wide variety of shared resources and domains. One of such extensions has been the scholarly recognition of a "knowledge commons"—"the institutionalized community governance of the sharing and, in some cases, creation, of information, science, knowledge, data, and other types of intellectual and cultural resources." Brett M. Frischmann, Michael J. Madison & Katherine J. Strandburg, *Governing Knowledge Commons*, *in* GOVERNING KNOWLEDGE COMMONS (Brett M. Frischmann, Michael J. Madison & Katherine J. Strandburg eds., 2014), at 3.

the European GDPR, which applies to any "natural or legal person, public authority, agency or other body which processes personal data."[114]

So far, the idea of a commons-based governance of big data has been primarily discussed in the context of scientific collaborations. Such a commons has been defined as "disparate and diffuse collections of data made broadly available to researchers with only minimal barriers to entry."[115] Responding to calls for increased international scientific collaboration, several expert bodies have developed high-level principles for transborder data sharing,[116] which would apply to anonymized data collected by various public institutions (e.g., tax returns, medical records, standardized tests, and more). Lowering or eliminating barriers of access to such data for research purposes is clearly necessary for the global pursuit of knowledge and development of scientific ideas. But time may have arrived to proceed further. Given the breadth and depth of data aggregated on information platforms and the potential societal benefits of user-generated data that have been accumulated on data platforms, limiting a "global data commons" to scientific purposes only seems to be overly narrow and restricting.

Importantly, the idea of a global data commons abandons the adversary stance of current regulatory approaches. Instead of treating data platform companies as hostile entities that should be approached with suspicion, it regards them as collaborators. After all, adverse and discriminatory impacts of algorithmic decision-making are not in the best interests of data platforms companies.[117] Rather, an open-minded collaboration between data platforms and external stakeholders could be in the mutual interests of all the involved parties. If designed properly, it can contribute to the public good without jeopardizing legitimate market interests of data platform companies, and without impinging the privacy rights of platform users.

---

114. GDPR, *supra* note 28, art. 4(8) (definition of "processor"). For a definition of "data processing," see the text *supra* note 28.

115. Yakowitz, *supra* note 106 at 2-3.

116. International Council for Science, *World Data System Strategic Plans 2019–2023* (2018), https://perma.cc/C5VA-MFMF; *Policy RECommendations for Open Access to Research Data in Europe (RECODE)*, OPENAIRE, https://perma.cc/2PPY-ZNSG (archived June 25, 2019).

117. Facebook, for instance, lost 6% of its shares value in the wake of the Cambridge Analytica scandal. *See, e.g.*, Lucinda Shen, *Why Facebook Suddenly Shed $35 Billion in Value*, FORTUNE (Mar. 19, 2018), https://perma.cc/QNH3-WKHB.

B. The Spectrum of the Global Data Commons

Treating user-generated data as a commons regime is far from suggesting a full "open access" system. Under a commons regime, platform companies, governments, private associations or firms, researchers, and private individuals would all hold well-defined but distinct access and usage rights over specific portions of user-generated data.[118] To derive public value out of this data, while protecting the legitimate commercial interests of data platforms and the privacy of data users, the commons regime should offer a spectrum of access and usage modalities. These modalities would have to tilt a balance between the public value gained from access to data, the platform companies' legitimate commercial interests, and the need to protect users' privacy.

Such a system can be structured as follows. Platform companies would retain their current management rights over user-generated data that is shared on their platforms. Private individuals may have access and usage rights over their own data, similarly to the current legal regulation in Europe. Non-governmental organizations, watchdogs, researchers, journalists, teachers, and any other professionals may obtain "authorized access and usage" rights that reflect the different data sharing modalities.

The Article suggests considering five sharing modalities, ranging from the most restrictive access rules to the most permissive ones. Decisions on which sharing modality is chosen for each data sharing initiative would need to be taken on a case by case basis, in a collaboration among data platform companies, their users, and policymakers.[119] The five proposed modalities are outlined below, along with examples on how these modalities have already been put into practice by some data platform companies. These examples refer to positive, yet sporadic practices. The Article contends that these practices shall be turned into a general policy.

1. Sharing Internal Data Analysis

At the most restrictive end of the data commons spectrum, data platforms may analyze their own data and publicly share insights generated from their internal research, but refrain from sharing the data itself. By

118. *See generally* UNDERSTANDING KNOWLEDGE AS A COMMONS 125 (Charlotte Hess & Elinor Ostrom eds., 2007).

119. See *infra* Part IV.C. on implementation arrangements.

choosing this modality, the company ensures that its data is kept private and secure and avoids external scrutiny. In this pathway, data platforms may either release insights that emerged as a byproduct of their regular business practices, or purposefully bring external researchers into their networks and grant them strict data access permissions. These partnerships may be task-specific or involve lasting and regular access privileges.

Several data platform companies have engaged in this practice in the past years. For instance, the Mastercard Center for Inclusive Growth's *Donation Insights* reports have leveraged Mastercard's anonymized and aggregated transaction data to learn more about the trends in philanthropic donations, such as when donations increase or what types of organizations benefit from donations.[120] These insights can help philanthropic organizations better understand various trends in public giving.

While Mastercard undertakes its research internally, Facebook has opted for engaging external researchers. As part of its CommAI Visiting Researcher Program, Facebook invites researchers from universities and labs to visit its facilities, collaborate on machine learning research, and produce publications and open-source code.[121] This research is intended to advance the state of the art in the field of machine learning, as well as improve Facebook's internal algorithms and data analytics processes.[122] A similar Google Visiting Faculty Program allows researchers to visit Google, collaborate with Google's researchers and tap into the company's resources, and publish new findings based on Google's data.[123]

This modality of data sharing is restrictive and enables the data platform company to retain full control both over the data and the results of the data analysis. But even in its current format, this level of data access can provide external researchers with opportunities to inspect some of the data accumulated by platform companies and come up with potentially valuable insights. For instance, if prior research identified discriminatory

120. *Donation Insights: Closing the Information Gap on Charitable Giving*, MASTERCARD CTR. FOR INCLUSIVE GROWTH, https://perma.cc/JYF8-NPK4 (archived June 25, 2019). *See also* Brice McKeever, Solomon Greene, Peter Tatian & Deondre Jones, *Data Philanthropy: Unlocking the Power of Private Data for Public Good*, URBAN INST. (July 2018).

121. *Facebook Visiting Researchers and Post-doc Program*, FACEBOOK RES., https://perma.cc/F7L4-ZW9F (archived June 25, 2019).

122. *See generally Facebook AI Research*, FACEBOOK RES., https://perma.cc/GB68-M46N (archived June 29, 2019)

123. *See, e.g.*, *Collaborations with the Research and Academic Communities*, GOOGLE AI, https://perma.cc/M9EX-6USP (archived June 29, 2019); *Visiting Research Program*, GOOGLE AI, https://perma.cc/3E32-NNS4 (archived June 29, 2019).

algorithmic outputs of Google's AdSense that offer men and women different ads or images, Google can either conduct its own research to understand whether and how these algorithmic outputs could be neutralized, or grant trusted external researchers strict data access and usage rights and invite them to scrutinize the algorithm's code.

This limited access approach can serve as a first step towards more flexible data sharing modalities, or enable data access in particularly sensitive cases, in which data platform companies are particularly reluctant to relax control.

2. Releasing Targeted Data

The most common sharing modality pursued by platform companies is the release of targeted data that can help address a concrete social problem or cope with an emergency. This pathway typically consists of two options: sharing data with trusted partners to address emergencies or other public challenges, and carrying out competitions that invite qualified applicants to develop new apps or discover innovative data uses.[124]

The trusted partnerships option has proven particularly valuable in the context of public emergencies because platform companies are often uniquely positioned to rapidly capture and process signals sent by their users. As an example, Facebook's Disaster Maps initiative provides partner organizations with three types of aggregated and de-identified maps immediately following a natural disaster[125]: (i) Location density maps that offer aggregate information regarding people's location shifts immediately before, during, and after a natural disaster, thus helping rescue organizations better understand disaster-affected areas; (ii) Movement maps between neighborhoods or cities following a disaster, which shed light

124. *Yelp Consumer Protection Initiative*, GOVLAB, https://perma.cc/JEG9-QBBY (archived June 29, 2019). Interestingly, in some cases data platforms can be on the recipient end of user-generated data. For instance, Yelp collaborates with the investigative journalism organization ProPublica to incorporate data that the latter collects on health care statistics and consumer opinion survey data. This data informs the Yelp business pages of more than 25,000 medical treatment facilities. *Id.*

125. Molly Jackman, *Using Data to Help Communities Recover and Rebuild*, FACEBOOK NEWSROOM (June 7, 2017), https://perma.cc/ZVL2-6THU ("Facebook can help response organizations paint a more complete picture of where affected people are located so they can determine where resources—like food, water and medical supplies—are needed and where people are out of harm's way.").

on common evacuation routes or predict traffic congestions; and (iii) Safety Check maps, which rely on self-reporting by Facebook users who can use the platform's Safety Check feature to mark themselves as safe following the disaster. Each of the maps seeks to strike a balance between the necessity to provide sufficiently granular information to rescue organizations and the need to protect individual privacy and thus avoid revealing any personally identifiable data points.[126]

Initiatives to share location and mobility information have also been adopted by several telecommunications companies, which agreed to share with partner organizations and researchers aggregated call data records. These records include a caller's identity, the time of the call, the phone tower that handled it, and the number called. For instance, Safaricom, the leading mobile service provider in Kenya, shared with researchers de-identified data from 15 million cell phone users in Kenya.[127] They used the data to visualize the dynamics of human carriers of malaria and distinguish between regions that are respectively sources and sinks of this disease. The research revealed, for instance, that many cases of malaria in Nairobi did not actually start in that city but were carried there from elsewhere. The data also helped identify locations that had the highest probability of spreading the disease, thus enabling health workers to prioritize their efforts. Call data records shared by mobile service providers with trusted researchers were used in a similar manner to characterize human mobility during floods[128] or earthquakes.[129] In fact, the assumption that targeted

126. Paige Maas et al., *Facebook Disaster Maps: Methodology,* FACEBOOK RES. (June 7, 2017), https://perma.cc/74KK-36GF. Another related initiative that Facebook has been engaged in is to share its population density data with Red Cross's Missing Maps—a project that aims to improve mapping data for at-risk communities. Facebook donates data related to computer visioning and satellite imagery, helping Missing Maps fill in gaps in their work. *See Missing Maps*, GOVLAB, https://perma.cc/T2HE-5998 (archived June 29, 2019).

127. The mobile provider enjoys a market share of 92% of mobile subscriptions in the country. *See* Amy Wesolowski et al., *Quantifying the Impact of Human Mobility on Malaria*, 338 SCIENCE 267 (2012); Amy Wesolowski et al., *The Use of Census Migration Data to Approximate Human Movement Patterns Across Temporal Scales*, 8 PLOS ONE e52971 (2013).

128. Alfredo J. Morales et al., *Studying Human Behavior Through the Lens of Mobile Phones During Floods*, INT'L CONF. ON ANALYSIS OF MOBILE DATA (NETMOB), Apr. 8-10, 2015, https://perma.cc/D44U-4YJU.

129. Benyounes Moumni, Vanessa Frías-Martínez & Enrique Frías-Martínez, *Characterizing Social Response to Urban Earthquakes using Cell-Phone Network Data: The 2012 Oaxaca Earthquake*, PROC. 2013 ACM CONF. ON PERVASIVE & UBIQUITOUS COMPUTING,

releases of cell phone location data can help address emergencies has become so widely ingrained that a media outcry has ensued in the wake of the Ebola epidemic in 2014 against government agencies and rescue organizations for *not* using call data records.[130]

Targeted data releases to trusted partners can also be undertaken in non-emergency contexts. Mastercard, for instance, has granted access to some of its data to Harvard researchers to explore issues such as the impact of tourism on emerging economies[131] and the role of knowledge exchange between countries.[132] Uber shares aggregated and anonymized mobility data with city planners and officials to help inform urban policies and make transport in cities more efficient.[133]

Data platform companies do not necessarily need to predetermine how their data would be used or what problems its use could address. Rather, they can generate incentives for partners to develop innovative uses and applications of specific subsets of their data. Yelp holds an annual competition, as part of which it releases a dataset and invites students "to conduct research or analysis on our data and share their discoveries with us."[134] Yelp attests that this competition, now in its thirteenth iteration, has generated hundreds of academic papers that rely on its datasets.[135]

Data platform companies that hold such competitions typically undertake a range of measures to de-identify the data and protect the privacy of their users. For example, Orange Telecom, an African communications company, hosted a competition in 2013 that allowed researchers to brainstorm ideas on how to use its data to solve problems related to transportation, health, and agriculture.[136] Notably, the data was

1199 (Sept. 8-12, 2013), https://perma.cc/LTJ2-Q8M4.

130. David Talbot, *Cell-Phone Data Might Help Predict Ebola's Spread*, MIT TECH. REV. (Aug. 22, 2014), https://perma.cc/67J9-CMLG; *Waiting on Hold*, THE ECONOMIST (Oct. 25, 2014), https://perma.cc/6F3L-62ZT.

131. *The Economic Impacts of Foreign Tourism,* MASTERCARD CTR. FOR INCLUSIVE GROWTH (June 17, 2016), https://perma.cc/J9LP-S6Q2.

132. Growth Lab, *Uncovering New Insights for How Business 'Knowhow' Impacts Economic Growth,* HARV. U. CTR. FOR INT'L DEV. (Jan. 20, 2016), https://perma.cc/A8CR-FY67.

133. *Uber Movement*, GOVLAB, https://perma.cc/KFG9-TPSC (archived June 29, 2019).

134. *Yelp Dataset Challenge*, YELP, https://perma.cc/4P42-FCUQ (archived June 29, 2019).

135. *Id.*

136. McKeever et al., *supra* note 120. Researchers received an initial description of

de-identified by Orange's local subsidiary, and antennae locations were blurred "to protect Orange's commercial interests."[137] This enabled researchers to map and track mobility and communication network activity without knowing the names of the users in question.

These different data sharing approaches share a common feature—the data platform retains full access and control of the raw user-generated data and to its algorithms. Researchers, rescue organizations, or competition participants only have access to data subsets that the data platform decides to release and can only use the data in line with the data platform's conditions and restrictions.[138] Facebook's newest initiative is similar, inviting researchers to explore the role of social media in a democracy by investigating specific subsets of its data.[139] The challenge and the corresponding scope of the data release are determined by the data platform, but external stakeholders may get a significant level of access. These stakeholders scrutinize the data in order to address pressing social challenges. A similar approach could be employed to investigate algorithmic decision-making and adverse algorithmic outputs.

3. Data Pools

Under the previous data sharing modalities, platform companies retained full control over the data they released to trusted partners. The data pools model relaxes this assumption. A data pool is a horizontal partnership between two or more companies or organizations that agree to share and analyze each other's data, and help fill knowledge gaps while minimizing duplicative efforts.

The idea of pooling resources to facilitate collaborative access and usage has been tested in a closely related area—access to patents. Michael Mattioli shows that large patent holders often form patent pools— an agreement by

---

the data, then sent in abstracts on the basis of which they received four datasets. The datasets showed traffic between antennae, movement trajectories for a large group of phone users referenced by antenna location, movement trajectories according to country administrative regions, and communication network details for a group of users.

137. Vincent D. Blondelet et al., Data for Development: the D4D Challenge on Mobile Phone Data (2013) (unpublished manuscript), https://perma.cc/W86M-C29Y.

138. Linnet Taylor, *The Ethics of Big Data as a Public Good: Which Public? Whose Good?*, 374 PHIL. TRANSACTIONS ROYAL SOC'Y A, Dec. 2016.

139. See *supra* notes 1 and 3, and accompanying text.

two or more patent holders to aggregate and share their patents by cross-licensing under a unified agreement at a standard rate.[140] A central administrator typically licenses the collected patent rights, collects royalties, and distributes those sums to the patent holders according to some predetermined formula. Rather than searching for relevant patent holders and negotiating a series of licenses, prospective licensees (typically inventors and researchers) can simply approach a single pool of patents for a license offered at a standard rate. This practice has routinely conserved vast transaction costs within technology markets.[141]

Mayer-Shönberger and Cukier foresee a similar dynamic in the user-generated data domain, arguing that "we'll see the advent of new firms that pool data from many consumers, provide an easy way to license it, and automate the transactions."[142] To facilitate data access and analysis, various formats of data-sharing pools may be considered as a joint initiative of several platform companies, or as a collaboration among platform companies, government entities, and research institutions.[143]

One example of a data pool is a collaboration between Esri, a global mapping company, Waze, a community-based traffic and transport app, and municipal governments that can access real-time traffic data provided by the two companies. As part of the pool, municipal governments share through Esri real-time construction and road closure data, and in exchange Waze shares its community-collected traffic data. All three parties benefit

140. Mattioli, *supra* note 106. *See also* Robert P. Merges, *Institutions for Intellectual Property Transactions: The Case of Patent Pools*, *in* EXPANDING THE BOUNDARIES OF INTELLECTUAL PROPERTY: INNOVATION POLICY FOR THE KNOWLEDGE SOCIETY 123, 129-30, 132, 144 (Rochelle Dreyfuss et al. eds., 2000); Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, *in* 1 INNOVATION POLICY AND THE ECONOMY 119-50 (Adam B. Jaffe, Josh Lerner & Scott Stern eds., 2001). Pooled patents are typically available to all members of the pool for free and to nonmembers on standard licensing terms. Frischmann et al. cite the example of a patent pool facilitated by the Manufacturers Aircraft Association (MAA), formed in 1917. As a result of the U.S. government's urgent need for airplanes during World War I, it facilitated an agreement between the MAA and other airplane manufacturers, through which the manufacturers agreed to cross-license the patents to one another on a royalty-free basis. *See* Brett M. Frischmann et al., *supra* note 113, at 3.

141. Mattioli, *supra* note 106.

142. MAYER-SCHÖNBERGER & CUKIER, *supra* note 31, at 147-48.

143. For an example of one such initiative, see Peter Lee, *Toward A Distributive Commons in Patent Law*, 2009 WIS. L. REV. 917, 990 (2009) (describing a joint effort between the Icelandic government and a private company in the 1990s to build a database of clinical records, DNA, and family histories for the entire country).

as a result of their participation in the data pool because it enables them to easily obtain data that would otherwise be inaccessible to them.[144] Another example of a multi-partner data pool is the California Data Collaborative—a joint effort among a coalition of water utilities, cities, and water retailers that seek to create an integrated, California-wide platform for water policy and operational decision making.[145]

Data pools can be organized along a continuum ranging from the most to the least centralized[146]: under a fully centralized model, data from various sources is aggregated in a single database; an intermediate distributed model contains a central access portal, through which data that is maintained in different locations can be accessed; under a fully distributed model, data repositories are not technically integrated, but share a common legal and policy framework that allows access on uniform terms and conditions; and under a "noncommons" model, data repositories are fully disaggregated and lack technical and legal interoperability. As the data-sharing model is more centralized it also becomes more user-friendly, as it enables users to easily obtain data in compatible formats. At the same time, such models are more difficult to implement due to legal and technical interoperability obstacles.

Accordingly, data pools facilitated by platform companies may vary in their degree of centralization, interoperability, and data anonymization and processing. Some of them may only enable exclusive access to a select group of organizations and researchers that sign strict confidentiality agreements, and thus data anonymization may not be necessary. Others may be open to a broader range of stakeholders, but only contain data that does not raise privacy concerns.

So far, institutionalized data sharing initiatives among platform companies have been relatively rare. However, given the proliferation of data platforms in different areas and the value derived from the combination of distinct data sources,[147] data pools could turn into an attractive opportunity—for both public and commercial purposes. In the

144. *Esri and Waze Open Data-Sharing for Governments*, GOVLAB, https://perma.cc/KFG9-TPSC (archived June 29, 2019).

145. CALIFORNIA DATA COLLABORATIVE, https://perma.cc/3JWA-VHUB (archived June 29, 2019).

146. Jorge L. Contreras & Jerome H. Reichman, *Sharing by Design: Data and Decentralized Commons*, 350 SCIENCE 1312 (2015).

147. See sources cited *supra* note 107.

context of algorithmic decision-making, data pools may create a safe space, where platform companies and independent partner organizations could identify problematic algorithmic outputs, scrutinize and compare algorithmic performance, experiment with different operation modes, and generally collaborate on common issues of concern.

4. Granting Access to Public Actors

The three data sharing modalities outlined above share a common starting point: data platform companies release data under specific conditions to trusted partners. The idea of sharing data more openly or with a broader range of stakeholders is typically met with anxiety by data platform companies—such level of access may imperil the privacy of data contributors and contributions, impair the competitive market advantage of data platforms, weaken their business models, and eventually disincentivize them from aggregating user-generated data in the first place.

One way to open up data, but avoid meddling with privacy and business incentives concerns, is to provide generous data access to a trusted and independent external party. Independent government agencies are well positioned to fulfill this function. The professional independence of statistical offices and their longstanding experience in dealing with personal and confidential data assure that user-generated data would be treated with proper care, caution, and discretion.

The proposal of granting public actors (and in particular national statistical authorities) with access rights to privately-held data has recently been considered in the European Union. The "Communication on Building a European Data Economy," adopted by the European Commission in January 2017, puts forward a proposal that "[p]ublic authorities could be granted access to data where this would be in the 'general interest' and would considerably improve the functioning of the public sector, for example, access for statistical offices to business data, or the optimization of traffic management systems on the basis of real-time data from private vehicles."[148] This proposal follows the logic of the French Digital Act,

148. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Building a European Data Economy § 3.5, COM (2017) 9 final (Jan. 16, 2017), https://perma.cc/82MJ-QQ4Q.

enacted in 2016, which grants statistical authorities the right to access, under certain conditions, privately-held data that can be valuable for public purposes (dubbed "public interest data").[149]

As a model, granting access to public actors is different from the data sharing modalities discussed above. It is no longer the private data platform company that decides what portions of its data would be released, to whom, and under which conditions. It is rather the national regulator who mandates the release of certain data under conditions specified in the law. This model is naturally more intrusive and precarious for the data platform companies. To alleviate the concerns of these private actors, the European Statistical System emphasizes "the longstanding experience of statistical offices in processing confidential or sensitive data as well as the existence of a comprehensive set of rules and measures . . . can provide the necessary assurance that the further re-use by statistical authorities of machine-generated data will not undermine the existing rights of the private data-holders."[150]

A similar approach could be considered in the United States, at least with respect to specific types of user-generated data that have an unequivocal public-interest component. For instance, in the context of scoring algorithms that determine individual credit scores or eligibility for certain benefits, Citron and Pasquale suggest that expert technologists from the FTC or FCC could be granted access to private scoring algorithms "to test them for bias, arbitrariness, and unfair mischaracterizations."[151] This approach could be extended to enable independent government agencies or watchdogs with access to data that can clearly serve the public interest—for instance, aggregated individual mobility data from a disaster-affected area or messages related to elections. Such access would enable the government to swiftly respond to public emergencies, as well as investigate cases of adverse social impacts of algorithmic decisions in politically

149. EUR. STATISTICAL SYS., DATA ACCESS FOR OFFICIAL STATISTICS (2017), https://perma.cc/7F2D-K4YL; Loi 2016-1321 du 7 octobre 2016 pour une République numérique [Law 2016-1321 of October 7, 2016 for a Digital Republic], JOURNAL OFFICIEL DE LA RÉPUBLIQUE FRANÇAISE [J.O.] [OFFICIAL GAZETTE OF FRANCE], Oct. 8, 2016, No. 0235, https://perma.cc/2EYU-YZ7A.

150. Data Access for Official Statistics, *supra* note 149.

151. Citron & Pasquale, *supra* note 62, at 25. *See also* Frank Pasquale, *Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries*, 104 NW. U. L. REV. 105 (2010).

sensitive cases (such as, for instance, Cambridge Analytica's scandal or allegations related to Russia's interference into US elections).

5. Open access

The most permissive pathway on the data commons spectrum is an open access regime, as part of which the data platform company provides free, public, and uncertified access to portions of its user-generated data. This modality creates room for any person to access, analyze, interpret, and re-share the data. The data may still be anonymized and aggregated to protect privacy.

Data platform companies are typically reluctant to share the user-generated data they accumulate without strict access control measures. An open access regime is thus challenging in this context. However, data platforms at times find this regime useful to their own objectives. For instance, all messages posted on Twitter ("tweets") are public by default, unless the user decides to actively protect them.[152] Twitter gives (and in some cases sells) private developers access to public tweets through Twitter's public Application Programming Interface (API) that includes various "premium" plans.[153]

Contrary to other access modalities, Twitter's open access regime only restricts the time frame of its data releases: a sample of the public Tweets published in the past 7 days is accessible for free; and access to all public Tweets published on the platform since 2006 is available for a premium or enterprise fee.[154] Beyond setting out these general conditions, Twitter does not play any active role in defining how its data is used, for what purposes, or by whom.

Twitter's open access API has already been leveraged in a range of international initiatives to better understand the needs and concerns of global communities. In September 2016, the United Nations Global Pulse project[155] announced that it would be working with Twitter "to support

152. *About Public and Protected Tweets*, TWITTER HELP CTR., https://perma.cc/662L-Q9G7 (archived June 29, 2019).

153. *About Twitter's API*, TWITTER HELP CTR., https://perma.cc/H329-WDH2 (archived June 29, 2019). Data generated by private Twitter accounts is not shared through the public API.

154. *Id.*

155. UNITED NATIONS GLOBAL PULSE: ABOUT, https://perma.cc/X479-PAWD (archived

efforts to achieve the Sustainable Development Goals."[156] The rationale for this engagement has been that

> "every day, people around the world send hundreds of millions of Tweets in dozens of languages. This public data contains real-time information on many issues including the cost of food, availability of jobs, access to health care, quality of education, and reports of natural disasters. This partnership will allow the development and humanitarian agencies of the UN to turn these social conversations into actionable information to aid communities around the globe."[157]

As part of the partnership, Twitter reportedly provides the UN Global Pulse with open access to its public API and data analysis tools. Taking advantage of these access rights, the UN Global Pulse used the visualization technique of "word-clouds" to build a Twitter-based crisis monitor that indicates how people are impacted by food prices in Indonesia.[158] These word-clouds get their specific visual shape by balancing human and algorithmic inputs in the data processing. Initially, the UN used experts to "train" the algorithm to recognize specific themes and emotions. However, to handle the speed of data, these word-clouds are then shaped by an automated algorithmic recognition of patterns in the semantic content of tweets.[159] This approach allows UN Global Pulse to transform real-time data feeds from Twitter's open access API into word-clouds, semantic clusters, and color-coded topics.

As part of a different initiative, the UN Global Pulse and the UN Millennium Campaign launched a social media monitor of priority topics

---

June 29, 2019). The project, launched by the Executive Office of the United Nations Secretary-General in 2009, is "based on a recognition that digital data offers the opportunity to gain a better understanding of changes in human well-being, and to get real-time feedback on how well policy responses are working."

156. Anoush Rima Tatevossian, *Twitter and UN Global Pulse Announce Data Partnership*, UNITED NATIONS GLOBAL PULSE BLOG (Sept. 23, 2016), https://perma.cc/83LN-4KQC.

157. *Id.*

158. Anders Koed Madsen, *Tracing Data: Paying Attention*, *in* MAKING THINGS VALUABLE 257-73 (Martin Kornberger et al. eds., 2015).

159. UN Global Pulse, *Monitoring Perceptions of Crisis-Related Stress Using Social Media Data* (Global Pulse Methodological White Paper, 2011), https://perma.cc/QC3T-XJEW.

related to the Post-2015 development agenda, aiming to provide real-time insights on developmental challenges that concern individuals around the world.[160] The monitor relied on a taxonomy of approximately 25,000 terms in English, French, Spanish, and Portuguese that are related to sixteen key development topics, and then filtered Twitter's open API for posts that include these terms. Using the keyword taxonomy, over 295 million tweets about the sixteen Post-2015 SDGs topics were extracted from over 45 million individual Twitter accounts. A world map visualization was then created to show the twenty countries that have proportionately tweeted most about each of the sixteen Post-2015 topics.[161] User-generated data that has been made accessible through Twitter's public API was also used by researchers to spot the dissemination of diseases.[162]

The promise of the open access modality for collaborative explorations of algorithmic process is unequivocal. Twitter's case is illustrative in this respect—it fueled the UN Global Pulse project and also provided the underlying data for countless academic studies.[163] Similar initiatives by other data platform companies could open to the world a trove of data that could be invaluable for addressing a variety of societal challenges, as well as

---

160. UN Global Pulse, *Using Twitter to Understand Post-2015 Global Development Priorities*, (Global Pulse Project Series no. 6, 2014), https://perma.cc/7FPA-AHNS.

161. "For example, Indonesia is one of the top countries that tweeted most about 'better transport and roads,'" reflecting the urgent need to address its "saturated transport infrastructure." *Id.* at 1.

> The visualizations often corresponded with political or development related events. When the Parliament of India passed the Food Security Act, also known as the Right to Food Act, in September 2013, discussion on Twitter increased by almost 300% in the month leading up to Parliament's decision. In Portugal, the outlying topic was jobs, as unemployment rose to a high of 15-17% from 2012-2014. In Spain, people tweeted most about government, with the volume of tweets doubling in November 2014 due to corruption scandals.

Id. at 2.

162. In the field of epidemiology, Chunara et al. have described how Twitter posts were used to spot relevant signals of disease in the 2010 Haitian cholera outbreak two weeks before data from "official" sources became available. Rumi Chunara, Jason R. Andrews & John S. Brownstein, *Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak*, 86 AM. J. TROPICAL MED. & HYGIENE 39. *See also* Joshua Ritterman, Miles Osborne & Ewan Klein, *Using Prediction Markets and Twitter to Predict a Swine Flu Pandemic*, PROC. 1ST INT'L WORKSHOP ON MINING SOC. MEDIA 9 (2009) (employing a similar approach to predict the spread of swine flu pandemics).

163. Klint Finley, *Twitter Opens Its Enormous Archives to Data-Hungry Academics*, WIRED (June 2, 2014), https://perma.cc/VCZ7-AML4.

understanding (and monitoring) otherwise hidden patterns of algorithmic operation.

* * *

In sum, the different sharing modalities showcase how user-generated data can be useful to address a variety of urgent social problems, while granting data platform companies various degrees of control over the data. This section illustrated cases in which data platforms have already engaged in sharing their data with a range of external stakeholders. Nowadays, these are primarily sporadic initiatives that are undertaken according to the company's' discretion for a variety of company-specific reasons and objectives. Some companies have agreed to open up access to their data in response to public criticism, others may be willing to share data to gain positive publicity, and yet others view data releases as part of their general social responsibility or as an income generating activity.

This Article suggests turning these initiatives into a coherent and systematic data sharing policy. Data platform companies should be required or encouraged to provide external stakeholders—government agencies, international organizations, businesses, researchers, journalists, or curious individuals—access to user-generated data under one of the data sharing modalities. The choice of the most appropriate modality can be determined on a case by case basis. Ideally, it would be done in collaboration between the company and the external data-seeking stakeholders, but it could also be determined based on the company's preferences or the regulator's demands.

The next sections discuss the implementation arrangements that should be considered in order to implement the data commons proposal into practice.

C. Regulatory Measures & Incentives

The recognition of a "global commons" regime would naturally impose on data platform companies operational costs, related to the need to de-identify data, structure it, make it available in standard formats, regularly update, etc. Similarly to other cases of "common property" regimes, some of these costs may be borne by large institutional data users (e.g., research universities that sign up to the data pools), but some of them would need to be internalized by the platform companies.

Data platform companies are not likely to welcome such costs. They are known to adamantly oppose any attempts for regulatory intervention in their affairs,[164] and spend "record sums" on corporate lobbying to avoid such regulation.[165] The question of how to implement a global data commons regime thus becomes central. Implementation arrangements could range from direct regulation to softer, collaborative and consent-based methods. Both approaches are delineated below.

1. Sticks: Invoking the Public Utilities Doctrine

Imposing on data platforms the requirement to share access to user-generated data may be justified on the grounds of their *de facto* functioning as "public utilities." In the common law tradition, courts developed the "public utility" doctrine to ensure that industries, which provide goods and services that are considered essential to the public, offer such services "under rates and practices that [are] just, reasonable, and non-discriminatory."[166]

Industries that qualify as public utilities typically meet a double condition: they are considered a "natural monopoly,"[167] and are "affected

164. Solon & Siddiqui, *supra* note 100 (noting that the "main areas of concern [for infomediaries] include the threat of looming action over anti-competitive practices, anything that might lead to higher taxation, net neutrality and privacy").

165. Brian Fung & Hamza Shaban, *To Understand How Dominant Tech Companies Are, See What They Lobby For*, L.A. TIMES (Sept. 1, 2017), https://perma.cc/GZY8-23D4; Jonathan Taplin, *Why Is Google Spending Record Sums on Lobbying Washington?*, THE GUARDIAN (July 17, 2017), https://perma.cc/9CMK-4YNK.

166. Joseph D. Kearney & Thomas W. Merrill, *The Great Transformation of Regulated Industries Law,* 98 COLUM. L. REV. 1323, 1331 (1998). The terms "public utility" and "common carrier" overlap and are sometimes used interchangeably. "Common carrier" refers to publicly accessible entities charged with transporting people, goods, or communications from one point to another for a fee. Common carriers historically faced liability for losses and were required to make their services available to all similarly situated customers on equal terms. *See* Susan P. Crawford, *Transporting Communications,* 89 B.U. L. REV. 871, 878 (2009).

167. For the purposes of identifying a utility, the firm is considered to be a natural monopoly if it "cannot be operated with efficiency and economy unless it enjoys a monopoly of its market." JAMES C. BONBRIGHT, PRINCIPLES OF PUBLIC UTILITY RATES 11 (1961). There can be no close substitutes for the natural monopoly's product or service, and there must be barriers to entry so that the natural monopoly's status persists over time. *See also generally* JAMES C. BONBRIGHT, ALBERT L. DANIELSEN & DAVID R. KAMERSCHEN, PRINCIPLES OF PUBLIC UTILITY RATES (1988).

with public interest."[168] Classic examples include electricity or water industries, which provide their services over a distribution network that they own and operate.[169] As K. Sabeel Rahman explains, the common thread in the public utility discourse is the need to ensure collective, social control over vital private industries that provided foundational goods and services on which the rest of the society depends.[170] The extraordinary economic dominance of monopolistic companies that provided socially necessary services generated the threat of exploitative consumer practices and discrimination. The public utility model offered a way to check—and curb—this form of private power. Courts acknowledged that such firms are quasi-public in character, even if they are formally privately owned. As quasi-public entities, courts imposed on utility companies obligations of a "stricter duty of care," because they had "implicitly accepted a sort of public trust."[171]

In the 21st century, this logic can be extended to the dominant data platforms,[172] which increasingly function like infrastructures: "embedded, largely invisible, often taken-for-granted, highly standardized systems for

168. Munn v. Illinois, 94 U.S. 113, 130 (1877); *see generally* MARTIN G. GLAESER, OUTLINES OF PUBLIC UTILITY ECONOMICS (1927); CHARLES F. PHILLIPS JR., THE REGULATION OF PUBLIC UTILITIES (1993).

169. *See* Thomas B. Nachbar, *The Public Network,* 17 COMMLAW CONSPECTUS 67, 108 (2008) ("[P]ublic utilities . . . are integrated firms that provide both a commodity and the network over which it is carried . . . ."). Title II of the Telecommunications Act of 1996 adapted traditional concepts of public utility to the new telecom reality, seeking to sustain the balance between universal access and market competition by imposing public utility requirements on telecom services. Telecommunications Act of 1996 § 201, 47 U.S.C. § 201 (2012). The Act included the requirements to equally serve all comers, request just and reasonable rates, prohibit unjust or unreasonable discrimination, and establish physical connections with other carriers *Id.* The Act also empowered the FCC with broad authority to oversee the industry, investigate complaints, and enforce these obligations. *See* 47 U.S.C. §§ 204, 205, 208, 215.

170. K. Sabeel Rahman, *The New Utilities: Private Power, Social Infrastructure, and the Revival of the Public Utility Concept*, 39 CARDOZO L. REV. 1621 (2018).

171. *See* Nat'l Ass'n of Regulatory Util. Comm'rs v. FCC, 525 F.2d 630, 641-42 (D.C. Cir. 1976).

172. In the context of the network neutrality debate, scholars have advocated to impose such obligations on internet service providers more generally. For instance, Crawford argues that "High-speed Internet access services are the functional, modern-day equivalent of these earlier networks [of telegraph and telephone services] and must plainly be included in [the common carriage] conceptual framework." Susan Crawford, *First Amendment Common Sense*, 127 HARV. L. REV. 2343, 2369 (2014). Nachbar contends that "it should be self-evident . . . that modern communications networks like [those providing access to] the Internet are prototypical candidates for the imposition of traditional non-discriminatory access obligations." Nachbar, *supra* note 169, at 113.

circulating information."[173] Data platforms nowadays function as marketplaces, public squares, or clearinghouses. These platforms link producers and consumers of goods, services, and information, and the benefits of the platform to its users depends on its consolidated control and network effects. However, these benefits are in fact a double-edged sword—the concentration of control over the platform also creates additional vulnerabilities, as platform companies can exploit the vast amounts of data it collects about its users. Such platform power raises precisely the types of concerns to which historic public utility reformers were attuned to. Unlike public transportation or gas companies these platforms do not necessitate a public franchise for their operation. But they do reflect the two core conditions that historically triggered a public utility recognition—monopoly-like status and high social dependence.

First, information platforms can be functionally analogized to natural monopolies. Traditional natural monopolies like water, electricity, or communications infrastructure typically involve high sunk costs, high barriers to entry, and increasing returns to scale. Dominant platform companies have a market status that is difficult to circumvent: the significant network effects of a consolidated data platform yield similar increasing returns to scale, high entry barriers for competitors, and thus a likelihood towards either concentration among a few private providers on the one hand, or under-provision of the good in a more fragmented industry on the other.[174]

---

173. Mike Ananny & Tarleton Gillespie, *Exceptional Platforms*, INTERNET POL'Y & POL. CONF. (Sept. 22-23, 2016), https://perma.cc/K4AT-KU55.

174. Some scholars contend that the monopoly condition is not critical for the recognition of a "public utility." For instance, Susan Crawford argues that with regards to the regulation of telegraph and telephone companies, "implementation of [the common carriage and public utility] concepts is justified not because of the market power of the actors (after all, inns in olden days and taxis today are both common carriers even though they do not hold monopolies) but because of the status these categories of entities occupy." Crawford, *supra* note 172, at 2368; *see also* Brief of Amici Curiae Reed Hundt, Tyrone Brown, Michael Copps, Nicholas Johnson, Susan Crawford, and the National Association of Telecommunications Officers and Advisors in Support of Appellee at 27 n.8, Verizon v. FCC, 740 F.3d 623 (D.C. Cir. 2014) (No. 11-1355) ("While mature threats to competition can be important reasons for imposing antidiscrimination or equal access obligations, they are not constitutionally required prerequisites." (internal quotation marks omitted) (citing Heart of Atlanta Motel, Inc. v. United States, 379 U.S. 241, 257 (1964)); Nachbar, *supra* note 169, at 97 ("Not only does the market power theory face historical problems, but it also faces jurisprudential ones. The early history of common carrier regulation is devoid of any mention of monopoly, nor is

The second condition is social necessity. Just, fair, and equal access to information and connectivity services are an essential component of modern life, and a discriminatory provision of such services would magnify socio-economic disparities and inequalities of opportunity.[175] The dominant data platforms shape the distribution of and access to news, ideas, and information upon which our economy, culture, and increasingly politics depends. This clearly creates vulnerability among users who could be excluded from access, or more disturbingly, may be consuming a tainted or manipulated information stream.

Against this backdrop, mandating data platform companies to commit to non-discrimination and equal access principles that were traditionally imposed on public utilities fully corresponds with their economic, social, and political status in modern society. Such an approach is particularly warranted given the variety of adverse societal impacts that result from algorithmic decision-making employed by these data platforms. A requirement to provide access and usage rights to (some of) their data to external stakeholders could be a first step in a larger regulatory reform that would recognize data platforms as public utilities under specific circumstances.

2. Carrots: Invoking Financial and Social Incentives

Invoking the public utilities doctrine to support the recognition of a global commons of data may signify a mindset shift in the legal approach to private data platform companies. For this reason, the practical implementation of this approach may be beset with difficulties due to the likely opposition of platform companies. As noted above, one of the benefits of the global data commons approach is that it may serve the interests of the data platform companies, and they may decide to collaborate with it voluntarily. Thus, proper incentives should be provided to these companies to encourage such voluntary collaboration.

The challenge of how to make private entities more socially or environmentally responsible and responsive is far from new. Rather than mandating a certain behavior, government policies may encourage platform companies to assume public obligations on a voluntary basis. For instance,

---

market power an element of modern common carrier regulation of many industries.").

175. Rahman, *supra* note 170.

as noted by the OECD, "U.S. Government agencies, such as, for example, the Consumer Product Safety Commission, the Food and Drug Administration, and the Environmental Protection Agency, may set safety, health, and environmental requirements designed to protect the public, but they rely upon voluntary consensus standards, where possible, to meet their regulatory objectives."[176] Such voluntary behavior may be encouraged, for instance, through financial incentives or corporate social responsibility (CSR) initiatives.

Financial incentives can take a variety of forms. For instance, since 2007, the FDA has been offering drug companies that agree to contribute health data to a patent pool "Priority Review Vouchers," which significantly reduce the time necessary to bring a drug to market.[177] Mattioli suggests that a similar possibility could be for the United States Patent and Trademark Office (USPTO) to offer a fast-track to patent applicants who claim new innovations derived from data pools, thus incentivizing platform companies to participate in data pools and provide access to data to researchers.[178] In a similar vein, the National Institute for Health and other federal agencies that provide research grants could impose stricter data-sharing requirements on grant recipients and, importantly, greater penalties for failure to adhere to such policies. Yet another policy intervention could focus on reducing the risk of liability that data holders face for inadvertent disclosure of personally identifying information.

Another form of incentive could consist of tax benefits. For instance, when Amazon looked for a new location for its second headquarters, which was expected to create 50,000 high-level jobs, more than 200 U.S. cities put in bids. These reportedly included generous tax break offerings from cities such as Newark, Maryland, and Philadelphia.[179] While this approach may be criticized due to the enormous revenues of data platform companies, this approach can nonetheless be helpful to encourage public regarding data-sharing behavior.

---

176. ORG. FOR ECON. CO-OPERATION & DEV., DAF/COMP/WP2.WD(2010)28, COMPETITIVE ASPECTS OF COLLABORATIVE STANDARD SETTING § 2.2(7) (2010), https://perma.cc/SR6B-NPFD.

177. Mattioli, *supra* note 106, at 47; *see also* Michael Mattioli, *Communities of Innovation*, 106 NW. U. L. REV. 103, 126-27 (2015).

178. Mattioli, *supra* note 106, at 47.

179. Kevin Maney, *Big Tech: Hate Amazon, Apple, Facebook and Google? Get in Line*, NEWSWEEK (Nov. 6, 2017), https://perma.cc/7WQX-CBPY.

A range of "soft" approaches to encourage companies to behave in a public-regarding manner have been developed as part of the "corporate social responsibility" (CSR) agenda.[180] Drawing on the CSR experience, an effective, non-costly, and non-controversial approach to encourage platform companies to grant access and usage rights to their data is to publicly praise data commons collaborators, and criticize those who do not wish to engage in data sharing. This "naming and shaming" function can be exercised by non-government institutions. For instance, the Reputation Institute publishes an annual ranking of "socially reputable companies," which reflects companies' performance in three categories: citizenship (support of "good" causes, positive societal influence, environmental responsibility); governance (openness, transparency, and ethical business behavior), and workplace (equal and fair treatment of employees).[181] In 2017, Microsoft was ranked second in this index, and Google arrived third.[182] Similar rankings could be developed to assess the data sharing performance of platform companies, and name and shame them accordingly.

D. Challenges

The recognition of a "global data commons" may generate a host of challenges. Mismanagement of user-generated data can inflict significant social damage—privacy abuses, identity theft, cybercrime, etc. Thus, the recognition of a global data commons does not only imply that access to user-generated data would simply be open under certain conditions. It also requires the introduction of special protection measures that would ensure data integrity and security. This Article does not imply that these challenges could be fully and satisfactorily addressed from the outset. Rather, it seeks to start a conversation on the role and value of user-generated data in our

180. The European Commission defines CSR as "a concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis." *Communication from the Commission Concerning Corporate Social Responsibility: A Business Contribution to Sustainable Development*, COM (2002) 347 final (Feb. 7, 2002), https://perma.cc/VFW3-7HCX; *Promoting a European Framework for Corporate Social Responsibility*, COM (2001) 0366 final (July 18, 2001), https://perma.cc/EJ9D-6NSH.

181. Reputation Institute, 2017 Global CSR RepTrak (Sept. 2017), https://perma.cc/N8A4-F4FM.

182. *Id*. at 22. LEGO Group was ranked first. *Id.*

society, and the range of hurdles that may be associated with opening access to this data to a broader range of stakeholders.

At this initial stage, it may be worth addressing three major hurdles that may be invoked to impede the implementation of this proposal: (i) potential violation of data privacy and protection; (ii) users' consent that their data would be part of a commons regime and (iii) competitive considerations.

1. Addressing Privacy Concerns

One significant critique that can be levied against the usage of user-generated data by any external actor is that the privacy of individual users would be compromised. Sharing user-generated data may result in disclosing personally or demographically identifiable information, which may result in privacy or security abuses. This challenge is, however, not unique to the global data commons model. Selling user-generated data to third-party actors is deeply embedded into the business models of data platform companies. Voluminous scholarship has addressed this concern, highlighting cases in which platform companies and third-party actors violated the privacy of platform users, and examining measures that could be undertaken to prevent privacy abuses.[183]

In fact, a data commons that is managed according to clearly defined protocols could mitigate privacy abuses. Details on whether any privacy protection safeguards are embedded into data platforms' transactions with third parties are hardly available to the public. Conversely, under the global commons regime each sharing modality would include concrete and well-publicized measures that would be undertaken to protect data privacy.

183. *See generally* SEBASTIAN SEVIGNANI, PRIVACY AND CAPITALISM IN THE AGE OF SOCIAL MEDIA (2015); PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT (Julia Lane, Victoria Stodden, Stefan Bender & Helen Nissenbaum eds., 2014); Robert M. Groves & Adam Neufeld, *Accelerating the Sharing of Data Across Sectors to Advance the Common Good*, (Washington, DC: Beeck Center, 2017); Danielle Keats Citron, *Mainstreaming Privacy Torts*, 98 CALIF. L. REV. 1805, 1831 (2010) (contending that courts should invoke established tort remedies to address unwanted intrusions and disclosure of personal information instead of creating new privacy torts); Woodrow Hartzog, *The Scope and Potential of FTC Data Protection*, 83 GEO. WASH. L. REV. 2230 (2015) (arguing that the FTC has the authority to regulate data protection); Daniel J. Solove & Danielle Keats Citron, *Risk and Anxiety: A Theory of Data Breach Harms*, 96 TEX. L. REV. 737 (2018) (discussing courts' treatment of data breach harms).

These measures will primarily consist of data de-identification or "pseudonymization—processing of personal data in such a manner that it can no longer be attributed to a particular data record without additional information.[184] Such processing typically consists of purging data from elements that can identify specific individuals, such as full names, social security or passport numbers, etc. The 2016 EU Data Protection Directive has imposed pseudonymization requirements on all data processing undertaken by Member States.[185]

After pseudonymization, data is no longer directly and easily identifiable, and can only be referred back to a specific individual when combined with other data and statistical analysis.[186] Pseudonymization's rise can be seen as a response to technical advances. Full anonymization—purging of all identifiers so that no linking back to individuals is possible—often requires deleting much of the actual data in favor of tabulated results such as sums and averages. Pseudonymization thus offers a middle ground between directly identifiable personal data, which raises significant privacy concerns, and fully anonymized data, which prevents the full functioning of big data algorithms and thus undermines the business models of platform companies.[187]

The handling of user-generated data by MasterCard provides a handy example of how privacy protection measures could look like. As part of the company's Data Grant Recipients program, targeted data is released to universities and research institutions that conduct their research off-site.

---

184. GDPR, *supra* note 28, art. 4(5). The GDPR defines pseudonymization as "the processing of personal data in such a manner that the data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person." *Id.*

185. GDPR, *supra* note 28, art. 83(1). Pseudonymization also appears to be a method of choice in the context of data protection by design and by default, *id.* at art. 23, data security, *id.* at art. 30, and as part of codes of conduct, *id.* at art. 38.

186. However, it is possible to envisage situations where even de-identified data may be sensitive as individuals can still be tracked as part of groups. In such situations, data that was initially aggregated for disease prediction can, instead, be used to track groups of political interest, such as separatists, smugglers, undocumented migrants, or dissidents. *See* Linnet Taylor, *No Place to Hide? The Ethics and Analytics of Tracking Mobility Using Mobile Phone Data*, 34 ENV'T & PLAN. D SOC'Y & Space, 319 (2016).

187. *See* Viktor Mayer-Schönberger & Yann Padova, *Regime Change: Enabling Big Data Through Europe's New Data Protection Regulation*, 17 COLUM. SCI. & TECH. L. REV. 315, 328-29 (2016).

All the released data is scrutinized by MasterCard's Privacy and Data Protection counsel, and participating research institutions are obliged to develop a procedure that guarantees the destruction of the data at the end of the project. The data is transferred to the researchers via a secure file transfer mechanism, and the researchers are requested to communicate with MasterCard on a quarterly basis. The collaboration between Telenor, a large telecommunications company, and Harvard Public Health researchers was more restrictive in terms of embedded privacy protection. The data provided by the company was de-identified by hashing and encrypting personal identifiers, and some portions of the data were aggregated. Harvard researchers worked with Telenor's research department to identify the maximum spatial and temporal aggregation level for the data that would still make it possible to address their research questions.[188]

While de-identification and aggregation of data clearly imposes costs on the data platform company, these expenses can be shared with the external stakeholders that get access to the data, or they can be subsidized by the government. Since the global data commons seeks to contribute to the achievement of societal objectives and help address public challenges, facilitating the protection of users' privacy as part of the commons through public funding seems well-justified. Eligibility for such funding could be specified for different data sharing modalities, and based on the public value of specific initiatives and projects that will utilize the data commons.

2. Ensuring Users' Consent

A different concern related to the global data commons proposal may be that it opens user-generated data for scrutiny by a variety of third-parties without obtaining the consent of these users. This concern is, in fact, equally applicable to the current data governance regime. The basic principles of copyright suggest that all user-generated data is owned by its creator—the

188. The Telenor researchers aggregated the data to base-station level locally at the Pakistan affiliate's offices, making it possible to move the data out of the country. At the firm's Oslo headquarters, they then aggregated the data up another level owing to the concerns about business sensitivity, so that it showed movements of mobile phones not between base stations but between Tehsils, Pakistani administrative units below the province level. The Harvard team was then given access to the aggregated matrices, rather than the dataset itself. *See* CAROLINE O. BUCKEE & KENTH ENGO-MONSEN, TELENOR REPORT 2/2016, MOBILE PHONE DATA FOR PUBLIC HEALTH: TOWARDS DATA-SHARING SOLUTIONS THAT PROTECT INDIVIDUAL PRIVACY AND NATIONAL SECURITY (2016), https://perma.cc/6BPB-5PF2.

user. Data platforms formally acknowledge this logic, but require users that are interested in their services to accept a broad data usage license. The terms of this license are often remarkably broad. For instance, Facebook's license for using the data created by its users provides a handy example:

> You own the content you create and share on Facebook and the other Facebook Products you use, and nothing in these Terms takes away the rights you have to your own content. . . . To provide our services, though, we need you to give us some legal permissions to use that content.
>
> Specifically, when you share, post, or upload content that is covered by intellectual property rights (like photos or videos) on or in connection with our Products, you grant us a non-exclusive, transferable, sub-licensable, royalty-free, and worldwide license to host, use, distribute, modify, run, copy, publicly perform or display, translate, and create derivative works of your content (consistent with your privacy and application settings).[189]

Legal ownership over data generated on data platforms remains with the user, but this right is often empty. Private data companies are virtually the only entities that have access to it and can make any use of it. Data platform users often do not even know what data is collected about them, let alone reap any social or economic benefits out of it.[190]

A global data commons regime may, in fact, rectify this situation. Platform companies that participate in the data commons may integrate within their "terms of use" a specific query on whether the user accepts that their data would be employed for public purposes with strict privacy protection measures. Companies may also ask their users *ad hoc* whether they are interested in contributing data to a specific research endeavor (e.g.,

189. *Terms of Service*, FACEBOOK, https://perma.cc/M5QZ-ML3M (archived June 29, 2019).

190. This sentiment has been clearly expressed by Sen. Jon Tester (D-Mont.) during the hearing held for Mark Zuckerberg, Facebook's CEO, at the US House Committee on Energy and Commerce on April 10, 2018: "You said—and I think multiple times during this hearing—that I own the data . . . . That sounds really good to me. But in practice . . . [y]ou're making $40 billion bucks a year on the data. I'm not making money on it. It feels like you own the data. . . . [C]ould you give me some sort of idea on how you can really honestly say it's my data . . . ?" *Transcript of Mark Zuckerberg's Senate Hearing*, WASH. POST (April 10, 2018), https://perma.cc/3ELQ-6GJ8.

by issuing notifications when the user logs into the platform). This approach is already part of the law in the European Union, as Art. 13 of the GDPR requires platform companies to inform their users of the categories of recipients with whom their data would be shared.[191] Users may be similarly informed that their data may be used for research purposes.

Facebook's Data Policy, for instance, already includes such provision:

> We use the information we have (including from research partners we collaborate with) to conduct and support research and innovation on topics of general social welfare, technological advancement, public interest, health and well-being.[192]

Other data platforms could follow a similar policy, and thus address potential concerns related to users' consent.

E. Competitive Concerns

Data platform companies may object to the global data commons proposal as sharing their data may threaten their commercial interests or affect their competitive advantage. Smart and efficient data analysis is the core income source for these companies, and a policy that expects them to dilute this income source is, of course, ill-advised. A proper data commons regime should not, by any means, infringe the legitimate commercial interests of platform companies.

User-generated data is a peculiar resource: Its commercial and public-oriented uses can take place concurrently, without devaluing or adversely affecting each other. For instance, the core of Twitter's business model is to collect and analyze its users' data in order to provide these users with targeted and personalized ads, for which Twitter is paid by advertisers. This private, profit-oriented usage of user-generated data does not preclude other, public-regarding uses. The same pieces of data that fuel Twitter's targeted ads, can provide government authorities, organizations, and researchers with invaluable information regarding human behavior, as well as better understand potentially adverse decisions of Twitter's algorithms.

Similarly, individual location data that is accumulated by Waze, the online navigation and mobility platform, contributes to the business model

191. GDPR, *supra* note 28.

192. *Data Policy*, FACEBOOK, https://perma.cc/5VGN-67WF (archived June 29, 2019).

of Waze as it allows the company to provide its users with targeted ads. But these same data points can contribute to the public good if they can help municipal authorities to attune traffic lights operation to the volume of cars that pass municipal junctions during the day. To achieve this objective, municipal authorities do not need to breach Waze's commercial secrets, but rather gain access to some of Waze's data.

Thus, the recognition of the public/private hybridity of user-generated data does not imply that data platform companies, which play a critical role in aggregating and processing their users' data, would lose their commercial benefits and decision-making prerogatives. Rather, it signifies that user-generated data merits a special status and should be utilized for the public good. The spectrum approach proposed in this Article can allay companies' concerns as they would have a significant stake in deciding how data would be shared, and how to protect their trade secrets. An overly restrictive approach can naturally diminish the value of the data to external actors, but it would still constitute an important first step. As noted above, the expenses associated with data sharing should not be borne by data platform companies only. They can be, at least partially, subsidized by government authorities, research institutions, and international organizations.

V. CONCLUSION

Data is commonly regarded as the 21st-century version of oil—an essential resource that fuels the entire global economy, much like oil has fueled the industrial economy of the 19th and 20th centuries.[193] This Article argues that such a critical enabler cannot be accessed and managed by private entities for commercial purposes only. As it would be unthinkable to treat oil and gas as private resources that lack any public functions, it is equally inadequate to approach user-generated data in such a way. Rather, time has come to recognize the public value of user-generated data and open access to it.

Currently, Twitter, Waze, Facebook, or any other data platform company decides to share its data with external stakeholders based on its own priorities and considerations. But they are by no means obliged or expected to do so. The purpose of the global data commons regime is to

193. The Economist, *The World's Most Valuable Resource Is No Longer Oil, But Data*, (May 6, 2017), https://perma.cc/947A-WD25.

systematically prompt companies to engage in such activities and provide them with a variety of structured data sharing modalities that enable the protection of privacy, legitimate commercial interests, and any other considerations.

Contrary to other approaches that seek to curb the algorithmic power of data platforms, the global data commons regime takes a positive, collaborative, and incentives-based stance towards platform companies. It seeks to work together with these companies to derive the highest possible value out of the data that they have amassed to address a wide variety of societal challenges, without infringing the legitimate interests of both platform companies and their users.