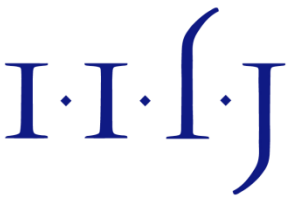

INSTITUTE FOR INTERNATIONAL LAW AND JUSTICE
NEW YORK UNIVERSITY SCHOOL OF LAW

International Law and Justice Working Papers

IILJ Working Paper 2021/1

Confronting Data Inequality

Angelina Fisher & Thomas Streinz

Institute for International Law and Justice
New York University School of Law
139 MacDougal Street, 3rd Floor
New York, NY 10012
www.iilj.org
@nyuiilj

ISSN: 1552-6275

Working Papers are issued at the responsibility of their authors, and do not reflect views of NYU, the IILJ, or associated personnel.

New York University School of Law
New York, NY 10012
U.S.A.

Cite as:
IILJ Working Paper 2021/1

Confronting Data Inequality [*]

Angelina Fisher [†] *& Thomas Streinz* [‡]

Abstract

Data conveys significant social, economic, and political power. Unequal control over data — a pervasive form of digital inequality — is a problem for economic development, human agency, and collective self-determination that needs to be addressed. This paper takes some steps in this direction by analyzing the extent to which law facilitates unequal control over data and by suggesting ways in which legal interventions might lead to more equal control over data. By unequal control over data, we not only mean having or not having data, but also having or not having power over deciding what becomes and what does not become data. We call this the power to datafy. We argue that data inequality is in turn a function of unequal control over the infrastructures that generate, shape, process, store, transfer, and use data. Existing law often regulates data as an object to be transferred, protected, and shared and is not always attuned to the salience of infrastructural control over data. While there are no easy solutions to the variegated causes and consequences of data inequality, we suggest that retaining flexibility to experiment with different approaches, reclaiming infrastructural control, systematically demanding enhanced transparency, pooling of data and bargaining power, and differentiated and conditional access to data mechanisms may help in confronting data inequality more effectively going forward.

Keywords: data inequality, data infrastructure, data law, data ownership, data rights, data governance, digital development.

[*] This paper was written as a background paper for the World Development Report 2021: Data for Better Lives. We are very grateful to Adele Barzelay, Elettra Bietti, Nikolas Guggenberger, Niels ten Oever, Edefe Ojomo, Przemyslaw Palka, David Satola, Dimitri van den Meersche, Christiaan van Veen, and Anna Yamaoka-Enkerlin for their careful comments, incisive questions, and constructive suggestions. We also thank the participants at NYU School of Law's Information Law Institute's Privacy Research Group for their valuable feedback when we presented this project for the first time. Katie Holland, Rachel Jones, and Maxwell Votey provided indispensable editorial assistance. The paper draws on collaborative ideas generated at NYU School of Law's Guarini Global Law & Tech initiative, where the authors teach and research Global Data Law with Benedict Kingsbury: www.guariniglobal.org/global-data-law.

[†] Angelina Fisher is Adjunct Professor of Law and Director for Policy and Practice of Guarini Global Law & Tech at New York University School of Law. Email: angelina.fisher@nyu.edu.

[‡] Thomas Streinz is Adjunct Professor of Law and Executive Director of Guarini Global Law & Tech at New York University School of Law. Email: thomas.streinz@law.nyu.edu.

Table of Contents

Introduction

In economic and human development narratives, the term "data inequality" typically connotes lack of opportunities that having access to digital data might otherwise present, such as better policy making, scientific innovation, and improvements to economic and social conditions. If data is a valuable resource like oil or water, how can it flow from those who have it to those who do not? This framing treats data as a valuable object to be gathered, shared, and exploited, and it tends to be concerned with law insofar as it either impedes or enables these activities.

In this paper, we suggest a different approach to data inequality. Drawing on insights from science and technology studies, media, communications, and information studies, and the emergent discipline of critical data studies, which has been pioneered by feminist and critical race scholars, we take as starting point the position that data is not a naturally occurring phenomenon. Data generation is a social practice. Inequality resides not only in having or not having data but also in having or not having the power to decide what kind of data is being generated and in what form or format, how and where it is amassed and used, by whom, for what purpose, and for whose benefit. We call this the "power to datafy".

We recognize that "data inequality" materializes in different ways, many of which are being explored within the rich scholarship on surveillance, algorithmic discrimination, digital labor, automation of the welfare state, digital mapping, and others.[1] Our focus in this paper is much narrower. We employ the term "data inequality" to refer to unequal control over data, understood both in distributional terms (having or not having data) and in terms of the power to datafy (deciding what becomes or does not become data). These are by no means the only forms of digital inequality. We hope, however, that unveiling the technical, social, organization and legal dynamics that constitute data inequality in this more confined sense will also illuminate alternative regulatory paths for addressing the broader concerns associated with the exercise of power through data.

In this paper, we make two key arguments. First, we posit that consideration of data inequality, as defined in this paper, requires examining the relationship between data and its constitutive infrastructures: those who control key infrastructures necessary for data generation, transfer, and use ("data infrastructures") will be in a better position not only to accumulate data but also to determine how human life and environment become datafied.[2] We take our inspiration for this argument from the burgeoning field of infrastructure studies.[3] One key insight of this interdisciplinary research agenda studying the history, development, operation, maintenance, and decay of infrastructures is to see infrastructures not merely as "objects". Infrastructures are complex, relational, and highly contextual, with their effects being a function of how social, technical, and organizational elements of their

---

[1] *See, e.g.*, Cathy O'Neill, WEAPONS OF MATH DESTRUCTION (2016); Safiya U. Noble, ALGORITHMS OF OPPRESSION (2018); Virginia Eubanks, AUTOMATING INEQUALITY (2018); Rediet Abebe, DESIGNING ALGORITHMS FOR SOCIAL GOOD (2019); Ruha Benjamin, RACE AFTER TECHNOLOGY (2019). *See also* Digital Welfare State and Human Rights Project, NYU SCHOOL OF LAW: CENTER FOR HUMAN RIGHTS AND GLOBAL JUSTICE, https://perma.cc/MEN7-TYER.

[2] Note that our use of the term "data infrastructures" deviates from the industry usage of the term. On the distinction between its meaning in engineering and in infrastructure studies, *see* Florence Millerand and Karen S. Baker, *Data Infrastructures in Ecology: An Infrastructure Studies Perspective*, ENVIRONMENTAL SCIENCE (Aug. 27, 2020).

[3] Paul N. Edwards, Geoffrey C. Bowker, Steven J. Jackson & Robin Williams, *Introduction: An Agenda for Infrastructure Studies*, 10 JOURNAL OF THE ASSOCIATION FOR THE INFORMATION SYSTEMS 364 (2009).

assemblages relate to, intersect with, or are embedded within each other, other infrastructures, the political economy, and law. This kind of "infrastructural analysis" can bring to light the often-less-visible enabling dynamics involved in the generation and subsequent availability of data.[4]

Second, we question the extent to which extant law and institutions are attuned to data inequality.[5] The global "free flow" of data through the Internet relies on physical infrastructures and interoperability standards. The laws facilitating and protecting this "free flow" tend to ignore between whom data flows, where data accumulates, and who ultimately benefits. In protecting the "free flow" of data, international economic law may entrench an unequal status quo by restricting states' ability to localize or redistribute control over data, thereby foreclosing potential pathways towards digital development that are more attuned to data inequality. Data protection and privacy law and intellectual property law tend to treat data as a regulatory object, focusing predominantly on specifically delineated individual rights, but have been largely silent with respect to concentrated control over data-generating infrastructures. These areas of law and the framing they adopt for regulation of data are not able to address data inequality and may even entrench it. Control over data is frequently achieved through control of the relevant data infrastructures. Certain strands of antitrust and competition law and recent regulatory initiatives in the EU seem more attuned towards infrastructural control over data. They tend to incorporate, however, certain assumptions about markets, market efficiencies, and consumer welfare that may ignore broader concerns around data inequality, which ought to be addressed through other means.

We focus on these areas of law — international economic law, intellectual property law, data protection and privacy law, and antitrust and competition law — because they have been most prevalently invoked as regulatory pathways for data. They are also the dominant fields from which regulatory models are being exported to or proposed for developing digital economies. Our aim is not to negate the relevance of these fields and the important contributions they can make but to highlight where their framings might fall short or even undermine development objectives. There are other areas of law that are salient for data regulation — for example, corporate law and tax law — that are beyond the scope of this paper. Overall, we seek to overcome the siloes in which legal regulation of data tends to be discussed, consider regulation of data by other means than law, and hint at alternative conceptualizations of data as a relational construct that implicates the rights and interests of many.

Throughout our analysis we highlight the role of corporate power in constituting, reinforcing, and scaling data inequalities. Corporate control over data and data infrastructure is a complex phenomenon and, depending on one's normative commitments, has various positive and negative externalities. We

---

[4] This approach is inspired by the InfraReg project incubated at NYU School of Law's Institute for International Law and Justice. *See* InfraReg, INSTITUTE FOR INTERNATIONAL LAW & JUSTICE, www.iilj.org/InfraReg. *See also* Benedict Kingsbury, *Infrastructure and InfraReg: On Rousing the International Law 'Wizards of Is'*, 8 CAMBRIDGE INTERNATIONAL LAW JOURNAL 171 (2019). *See also* Geoffrey C. Bowker, Karen Baker, Florence Millerand, David Ribes, *Towards Information Infrastructure Studies: Ways of Knowing in a Networked Environment*, *in* INTERNATIONAL HANDBOOK OF INTERNET RESEARCH 97 (Jeremy Hunsinger, Lisbeth Klastrup, Matthew Allen, eds., 2010).

[5] Our analysis focuses mainly on US and EU law because these jurisdictions have historically shaped legal (non-) regulation of the digital economy and continue to dominate the discourse globally. In the Global Data Law project at NYU School of Law's Guarini Global Law & Tech initiative, we explore alternative pathways towards global data regulation. For more information see Global Data Law, NYU SCHOOL OF LAW: GUARINI GLOBAL LAW & TECH, www.guariniglobal.org/global-data-law.

do not suggest that corporate involvement should be avoided altogether in the interventions aimed at fostering digital economies and societies. On the contrary, we focus on corporations precisely because of the central role they play in such interventions and because of their increasing prominence, through public-private partnerships, in the Sustainable Development Agenda more broadly. Entrepreneurs, software developers, and others may derive benefits from digital infrastructures controlled by large technology companies (e.g., through services offered by cloud computing). Individuals and communities may similarly enjoy certain conveniences and pleasures afforded by data-driven technologies. At the same time, it is imperative to be fully cognizant of the effects that corporate power has on individuals, communities, and countries, particularly where such power is exercised through control over data infrastructures.[6] At the same time, our critical analysis of concentrated corporate control over data infrastructures should also not be misunderstood as an unconditional endorsement for concentrated governmental control over such infrastructures, without due regard to the particular economic, social, and political contexts in which these infrastructures are being created and deployed.[7]

More stringent regulation of existing data infrastructures may be necessary to ensure the development of data-productive rather than data-extractive economies. Any meaningful regulatory intervention, however, is dependent on accurate information about data infrastructures, including provenance of and context within which data is generated, sites and mechanisms of control, and distribution of power and interests among different actors. The notorious and somewhat paradoxical opaqueness of the most important data infrastructures can be countered through more forceful demands of transparency and data-sharing, not as an end in itself but as a precondition for further political and regulatory action.

To avoid reproduction and entrenchment of data inequality, it may be necessary to reduce dependency on data infrastructures controlled by large corporate entities while at the same time resisting adoption of alternative wholesale top-down data governance models that — intentionally or not — may suppress, ignore, or exploit traditionally marginalized groups. This may necessitate the creation, development, and support of alternative data infrastructures. Smaller, local but also potentially transnationally aligned actors could be empowered to make their own choices about which data to collect and how and which data infrastructures to use and to rely on.

The network effects and stark economies of scale that are hallmarks of the digital era might necessitate the pooling of data and power to create the critical mass necessary to counter or at least negotiate effectively with those who possess outsized infrastructural control over data. To equalize asymmetric control over data rather than exacerbate it, conditional and differentiated access to data infrastructures could be developed to ensure adequate compensation and more just recalibration of data access and benefit. This may also be the space where the power of international organizations (and the data they control) could be leveraged for the benefit of developing economies.

---

[6] *See* Linnet Taylor & Dennis Broeders, *In the Name of Development: Power, Profit and the Datafication of the Global South*, 64 GEOFORUM 229 (2015). Many examples in this paper are drawn from Western e-commerce and social media platform companies, but our analysis applies to corporate control over large-scale data infrastructures more generally; *see below* Sections I.C–D.

[7] We emphasize the need for context-specific interventions throughout this paper, particularly in Part III.

Data inequality is not a technocratic problem for which there is either a purely technical or a purely legal or even techno-legal solution. Indeed, by undertaking an infrastructural analysis we hope to illuminate not only the co-constitutive relationship between data law and data infrastructures but also how both are shaped by the political economy of data capitalism.[8] Institutions promoting economic and social development, as well as state actors engaged in digital policies, ought to consider the broader impacts of datafication, including on individual welfare, development freedom, and democratic governance. Addressing data inequality requires recognition of the politics of data. The context-dependency and relativity of data and data infrastructures requires new thinking about ways in which publics affected by datafication can engage in public deliberation, effective contestation, and collective self-governance.

The paper develops these ideas in three parts:

In Part I, we caution against mono-dimensional conceptualizations of "data" as a naturally occurring phenomenon that ought to be exploited as an economic resource, and we emphasize instead the extent to which data is constructed through social and highly political practices. We highlight the changes to data collection, processing, transfer, and use resulting from widespread but uneven adoption of modern digital technologies across the globe. On this basis, we argue that unequal control over data is increasingly a function of highly concentrated corporate control over data infrastructures.

In Part II, we explore the extent to which law has facilitated unequal control over data by not addressing the infrastructural reasons for data inequality. We argue that legal interventions to address unequal control over data need to move beyond the approaches of legal data regulation that are predominantly focused on individual protection of privacy and personal data and property-type protections of data. We acknowledge that competition law and recent regulatory initiatives in the EU may be more attuned towards infrastructural control over data, but also expose their inherent limitations. Embedded in our analysis, we discuss the extent to which international economic law constrains states' ability to localize or redistribute control over data, thereby entrenching data inequality without addressing its root causes.

Part III, then, considers possible solutions to address unequal control over data. We caution against wholesale solutions and suggest targeted interventions to confront data inequality through 1) retaining development freedom, 2) reclaiming infrastructural control, 3) demanding enhanced transparency over data infrastructures, 4) pooling and differentiated access to data mechanisms, and (5) collective governance of data and data infrastructures.

[8] In this regard, we are inspired by the work by Julie E. Cohen, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM (2019); Amy Kapczynski, *The Law of Informational Capitalism*, 129 YALE LAW JOURNAL 1276 (2020); Katharina Pistor, THE CODE OF CAPITAL (2019).

I. Data Inequality as a Function of Infrastructural Control

We begin our analysis by challenging the mono-dimensional conceptualization of data as a resource, and instead emphasize data as a social practice. Seen from that vantage point, data inequality is more than uneven distribution of data. The power to decide what becomes datafied, by whom, where, how, in what form, and with what purpose is unevenly distributed. We then illustrate how the power to make these determinations and the ability to accumulate data is crucially dependent on control over data infrastructures.

A. Conceptualizing Data

The term "data" is ubiquitous. Its meaning, however, differs across fields and disciplines that seek to understand and articulate what data is, what is new or different about digital data, and how data is transforming social, political and economic dynamics.[9] In popular discourse, the use of metaphors is common.[10] Many analogize data to natural resources like oxygen, and of course, the by now proverbial oil.[11]

These metaphors often conceive of data as a *natural kind*, and as a *resource* that "exists in the wild," can be extracted, processed, and consumed through means of industrial production — invoking physical modalities of pipes and hoses to process and move the resource smoothly across space — in order to make something visible, discoverable, traceable, observable, and ultimately calculable. These metaphors of nature operate to evoke images of data as existing *a priori* in the same way that water or air or mineral deposits exist. This imagery is consistent with the etymology of the word data, which derives from the Latin verb *dare* (to give): data as something that is *given*.[12] The givenness of data is thus analogized to the givenness of natural resources which can be extracted.

Data is also often said to *flow*. From a technical perspective digital data is being transmitted through light pulses or electrical signals, depending on the type of cable used, at the behest of humans and their machines. Yet, the imagery of fluidity suggests that in its state-of-nature, data moves smoothly

---

[9] *See, e.g.*, Viktor Mayer-Schönberger & Kenneth Cukier, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013); Rob Kitchin, THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES AND THEIR CONSEQUENCES (2014); Vincent Mosco, TO THE CLOUD: BIG DATA IN A TURBULENT WORLD (2014).

[10] For studies of metaphors of "big data," *see* Cornelius Puschmann & Jean Burgess, *Metaphors of Big Data*, 8 INTERNATIONAL JOURNAL OF COMMUNICATION 1690 (2014); Jan Nolin, *Data as Oil, Infrastructure or Asset? Three Metaphors of Data as Economic Value*, 18 JOURNAL OF INFORMATION, COMMUNICATION AND ETHICS IN SOCIETY 28 (2019).

[11] The Economist popularized this framing in May 2017: "The world's most valuable resource is no longer oil, but data", *see The World's Most Valuable Resource is no Longer Oil, but Data*, THE ECONOMIST (May 6, 2017), https://perma.cc/3ACZ-34VL, followed by a February 2020 special report asking: "Are data more like oil or sunlight?", *see Special Report: Are Data More Like Oil or Sunlight*, THE ECONOMIST, Feb. 20, 2020, https://perma.cc/2SCF-C549. Mathematician Clive Humby coined the "data is the new oil" metaphor in 2006, *see* Charles Arthur, *Tech Giants May be Huge, but Nothing Matches Big Data*, THE GUARDIAN (Aug. 23, 2013), https://perma.cc/VD9A-YFJT.

[12] Daniel Rosenberg notes that during the 18th century, the meaning of "data" shifted from something that is accepted as given to the result of experimentation, discovery, or collection. Daniel Rosenberg, *Data before the Fact in* "RAW DATA" IS AN OXYMORON 15-40 (Lisa Geitelman ed., 2013). The change in meaning is not accidental but rather coincides with the growth and evolution of science and new modes of knowledge production that shifted away from theology to rationality, facts, evidence, and the testing of theory through experiment.

and uninterrupted, without acknowledging anyone's, or anything's, agency in the process.[13] Data from different sources is said to aggregate into *lakes* or *pools*, but when too much data accumulates so as to be unmanageable for the humans or machines who extract and process the resource, the narrative of *torrents*, *floods* and *tsunamis* of data shifts towards a need for control as data must be cleaned, refined, simplified, with "noise" eliminated to reveal its essence.[14] Once its natural force has been curbed, data is finally ready for consumption or can be used to create even more data.

Data metaphors have implications for our analysis of legal and infrastructural dimensions of data inequality. As Cornelius Puschmann and Jean Burges explain, "technological metaphors are 'never innocent' and, when deployed as part of deliberate rhetorical strategies, [they] have the potential to profoundly shape cultural and social practices".[15] Indeed, the metaphors outlined here above made their way into economic and political discourses. The idea of data as a resource focuses attention on data as an object that can be commodified to generate further value.[16] Geopolitical contests and transnational competition between businesses eager to realize data's value propositions play out in debates around the "free flow" of data across borders, on the one hand, and the desire for "data sovereignty" on the other. Seeing data as a valuable resource has also meant that legal interventions have focused on questions such as: Who owns data? Who collects and processes data? How should data be shared and with whom? How should people be protected from uses and misuses of data? How should concentrations of data in the hands of certain commercial actors be regulated? As we will discuss in Part II, these approaches often miss dimensions of data inequality associated with control over the data infrastructures that constitute data.

Treating data as something akin to a *natural* resource has the effect of depoliticizing the processes by which data comes into existence in the first place. It not only removes human agency but also conceals

---

[13] The controversial idea that nonhuman artifacts or objects can have agency to be understood in relation to other nonhuman and human "actants" is usually attributed to the French sociologist and pioneer in science and technology studies Bruno Latour. *See, e.g.*, Bruno Latour, *Where are the Missing Masses? The Sociology of a Few Mundane Artifacts*, *in* SHAPING TECHNOLOGY/BUILDING SOCIETY: STUDIES IN SOCIOTECHNICAL CHANGE 225 (Wiebe E. Bijker & John Law eds., 1992). *See also* Edwin Sayes, *Actor-Network Theory and Methodology: Just What Does it Mean to Say that Nonhumans Have Agency?* 44 SOCIAL STUDIES OF SCIENCE 134 (2014) (clarifying the terminology).

[14] *See* on the importance and implications of classification more generally Geoffrey C. Bowker & Susan Leigh Star, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES (2000).

[15] Cornelius Puschmann & Jean Burgess, *Metaphors of Big Data*, 8 INTERNATIONAL JOURNAL OF COMMUNICATION 1690 (2014).

[16] The value of data can be commercial (e.g., if data itself, or the insight it produces, is commodified and monetized), strategic or informational (e.g., for business management or policy-making), social (e.g., illuminating social problems recognizable only at scale), and so on. Data is of value, for example, to develop artificial intelligence technology reliant on large data sets, and is valued, when there is a market for datasets or when the value of data is embedded in the value of a company, realized upon sale or merger. How data value is produced and measured remains an unsettled, but increasingly studied, question across domains such as taxation, accounting, business management, and macroeconomic analysis. *See e.g.,* Aleksandra Bal, *(Mis)guided by the Value Creation Principle – Can New Concepts Solve Old Problems?*, 72 BUL. INT. TAX 11 (2018); Chiehyeon Lim et al., *From Data to Value: A Nine-Factor Framework for Data-Based Value Creation in Information-Intensive Services*, 39 INTERNATIONAL JOURNAL OF INFORMATION MANAGEMENT 121 (2018); Wendy C.Y. Li, Nirei Makoto & Yamana Kazufumi, *Value of Data: There's No Such Thing as a Free Lunch in the Digital Economy*, RIETI Discussion Paper Series 19-E-022, RESEARCH INSTITUTE OF ECONOMY, TRADE & INDUSTRY (Mar. 2019). *See also* Open Data Watch's "Value of Data Inventory" catalogues different studies on "value" of data. *Understanding the Impact and Value of Data: Ongoing and Upcoming Projects at Open Data Watch*, OPEN DATA WATCH, https://perma.cc/9Z3Y-V8UP. The value of data as a corporate asset is often not being accounted for; *see below* Section III.C.

the socio-technical practices, and the surrounding politics, through which phenomena are being converted into a set of computationally manipulable measurements.[17] Speaking of data in scientific research, Sabina Leonelli observes that data "are the results of complex processes of interaction between researchers and the world, which typically happen with the help of interfaces such as observational techniques, registration and measurement devices, and the re-scaling and manipulation of objects of inquiry for the purposes of making them amenable to investigation."[18] The same observation can be made with respect to the data that is "born digital" (created immediately in binary, hence digital, code), either purposefully or incidentally.[19] The decision to capture (or measure) a particular phenomenon, process, activity or environment is unequivocally made by humans. The means by which data is generated are designed and controlled by humans. Classifications and categorizations, formats, standards, and protocols, media of storage, transport, and dissemination are all integral parts of infrastructures that make data readable, searchable, manipulatable, and transmittable via the Internet. Each of these are themselves assemblages of materialities, social norms, organizational practices, histories, ideologies, and law, in form of legal instruments, practices, and institutions. As Lauren F. Klein and Miriam Posner observe: "data sets never arrive in the world fully formed, but are assembled from tangles of historical forces and ideological motivations, as well as practical concerns."[20]

That data is thus constructed and highly contextual is well-recognized in science and technology studies, media, communications, and information studies, and the emergent discipline of critical data studies, as well as by feminist and critical race scholars and many others who see data not merely as a resource but *as a social practice*.[21] This has allowed these scholars to ask pertinent questions such as: Who gains access to and is able to extract value from data?[22] What mechanisms enable personal data to be controlled by corporations?[23] How does data production and processing shape identities, environments, and our understandings of the world?[24] These questions, in turn, have allowed for examination of data not simply as a resource but also as a site of power that can reinforce (as well as

---

[17] Bruno J. Strasser & Paul N. Edwards, *Big Data Is the Answer . . . But What Is the Question?*, 32 OSIRIS 328 (2017).

[18] Sabina Leonelli, *What Counts as Scientific Data? A Relational Framework,* 82 PHILOSOPHY OF SCIENCE 5, 810 (2015). Leonelli sees data as a relational category; that is "as any product of research activities, ranging from artifacts such as photographs to symbols such as letters or numbers, which is collected, stored and disseminated in order to be used as evidence for knowledge claims." *Id.* at 816. *See also* Bruno J. Strasser & Paul N. Edwards, *Big Data Is the Answer . . . But What Is the Question?*, 32 OSIRIS 328, 329-330 (2017) (citing Bruno Latour, PANDORA'S HOPE: ESSAYS ON THE REALITY OF SCIENCE STUDIES, ch.2 (1999), "[t]o attach the label 'data' to something is to place that thing specifically in the long chain of transformations that moves from nature to knowledge; this act of categorization marks a particular moment in time when someone thought some inscription or object could serve to ground a knowledge claim.").

[19] Rob Kitchin, *Small Data, Data Infrastructures and Data Brokers*, *in* THE DATA REVOLUTION: BIG DATA, OPEN DATA, DATA INFRASTRUCTURES AND THEIR CONSEQUENCES (2014).

[20] Lauren F. Klein & Miriam Posner, *Data as Media*, 3 FEMINIST MEDIA HISTORIES 1 (2017).

[21] Christine L. Borgman, BIG DATA, LITTLE DATA, NO DATA: SCHOLARSHIP IN THE NETWORKED WORLD. (2015); Geoffrey C. Bowker, MEMORY PRACTICES IN THE SCIENCES (2005); Lisa Gitelman (ed.), "RAW DATA" IS AN OXYMORON (2013).

[22] danah boyd & Kate Crawford, *Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon*, 15 INFORMATION, COMMUNICATION & SOCIETY 662 (2012).

[23] Julie E. Cohen, *The Biopolitical Public Domain: The Legal Construction of the Surveillance Economy*, 31 PHILOSOPHY & TECHNOLOGY 213 (2018); Shoshana Zuboff, THE AGE OF SURVEILLANCE CAPITALISM (2019).

[24] David Ribes & Steven J. Jackson, *Data Bite Man: The Work of Sustaining a Long-Term Study*, *in* "RAW DATA" IS AN OXYMORON 147 (Lisa Gitelman ed., 2013); Safiya U. Noble, ALGORITHMS OF OPPRESSION (2018).

subvert) existing inequalities along gender,[25] race,[26] sexuality, and class dimensions,[27] within and across countries. A key theme that weaves through these lines of research is the importance of various practices involved in producing, accumulating, and analyzing data on democracy, freedom, self-governance, and socio-economic and political (in)equality.

These two discourses — one that is concerned with data as a resource that can be commodified to derive value and another that emphasizes data-as-a-social practice — proceed largely in parallel and intersect rarely. We try to bring them into conversation with one another by adopting an infrastructural perspective. Our goal is to illustrate how data inequality is constituted through control over relevant infrastructures and to illuminate the co-constitutive relationship between infrastructures and law. These themes are explored in the ensuing sections.

B. Recognizing Data Inequality

In contrast with metaphors of data lakes, torrents and firehouses, which evoke imageries of abundance, stands the metaphor of data deserts, signifying scarcity and absence of data. This imagery emphasizes the non-existence of data, spotlighting the inequality between those who have data and those who do not. The World Development Report 2021 recognizes this kind of data inequality. It notes that in developing economies the unavailability of data is often due to the absence of necessary infrastructures of connectivity, storage, and processing and prerequisite human labor and expertise.[28] In this paper, we want to advance the inquiry into data inequalities,[29] by foregrounding the less examined but critically important inequality of power to decide *what* data gets produced in the first place, that is the power to decide which phenomenon gets *datafied*, by whom, where, and how.[30] This requires us to tease out the relationship between data and its constitutive infrastructures to understand the root causes for unequal data generation.

So-called data deserts materialize where reality has not been translated (and reduced) into a computational measurement or where proxy data (that could be used to deduce information) exists but is not generally accessible. These kinds of data gaps are neither accidental nor inevitable but are a product of deliberate economic, social and political choices. Catherine D'Ignazio and Lisa Klein

---

[25] *See e.g.,* Catherine D'Ignazio and Lauren F. Klein, DATA FEMINISM (2020).

[26] *See e.g.*, Safiya U. Noble, ALGORITHMS OF OPPRESSION: HOW SEARCH ENGINES REINFORCE RACISM (2018).

[27] *See e.g.*, Christian Fuchs & David Chandler, DIGITAL OBJECTS, DIGITAL SUBJECTS: INTERDISCIPLINARY PERSPECTIVES ON CAPITALISM, LABOUR AND POLITICS IN THE AGE OF BIG DATA (2019).

[28] *World Development Report 2021: Data for Better Lives – Concept Note*, WORLD BANK (May 5, 2020), http://documents.worldbank.org/curated/en/778921588767120094/World-Development-Report-2021-Data-for-Better-Lives-Concept-Note.

[29] Data inequality is a relatively recent term that emerges from the expansive scholarship on digital inequalities. Jonathan Cinnamon has identified three dimensions along which data-specific inequalities (data inequalities) have emerged: *access* to data, *representation of the world* as data, and *control* over data flows. We do not adopt his typology in this paper, although a number of issues discussed by Cinnamon will be also echoed in our analysis. Jonathan Cinnamon, *Data Inequalities and Why They Matter for Development*, 26 INFORMATION TECHNOLOGY FOR DEVELOPMENT 214 (2020).

[30] *See* Ulises A. Mejias & Nick Couldry, *Datafication*, 8(4) INTERNET POLICY REVIEW (2019) for a brief history of the term "datafication". *See also* Katherine J. Strandburg, *Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context*, *in* PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT 11 (Julia Lane et al. eds., 2014); Ira S. Rubinstein & Bilyana Petkova, *Governing Privacy in the Datafied City*, 47 FORDHAM URBAN LAW JOURNAL 755 (2020); Joseph Savirimuthu, *Datafication as Parenthesis: Reconceptualising the Best Interests of the Child Principle in Data Protection Law*, 34 INTERNATIONAL REVIEW OF LAW, COMPUTERS & TECHNOLOGY 310 (2020).

poignantly observe that "[t]he phenomenon of missing data is a regular and expected outcome in all societies characterized by unequal power relations, in which a gendered, racialized order is maintained through willful disregard, deferral of responsibility, and organized neglect for data and statistics about those minoritized bodies who do not hold power."[31]

It is often tempting to turn to mass reserves of data held by large corporate actors in attempts to irrigate data deserts. Data philanthropy, open data, and preservation of "free data flows" are examples of interventions aimed at filling the data void that is due to *non-existence* of data in a particular place (e.g., poverty data in Sub-Saharan Africa) or about a particular phenomenon (e.g., global ambient air pollution mortality estimates). It is important to realize that whatever economic benefit such initiatives might hold, they also risk reproducing and accelerating inequalities of power relations that are embedded in the choices about what became (and what was excluded from becoming) data.[32]

The decision of *what* data to produce rests fundamentally with those who control the means of data production, even as that decision itself is a function of organizational practices, business models, the legal and political environment, market pressures, and the interests of those who control the relevant infrastructures, as much as (or sometimes even more than) the perceived needs and wants of consumers or the public. Data deserts are neither natural nor agentless as the control over the means of data production is unevenly distributed among different actors. The power to determine what becomes datafied is related to the power to accumulate data, which is tied to the control over relevant data-generating infrastructures.[33]

In the age of digitization, proliferation, and enhancement of computational power, the development of the Internet, and the ubiquity of devices, sensors, and platforms, choices about which data is generated are increasingly a function of infrastructures that enable data collection through devices, sensors, and platforms. Deliberate supplying of data — e.g., subjects participating in research projects, individuals applying for government services, employees complying with disclosure requirements under labor and tax laws, patients sharing their medical history to receive adequate health services etc.

---

[31] *See generally* Catherine D'Ignazio & Lauren F. Klein, DATA FEMINISM, ch 1 (2020) ("The Power Chapter"). Examples of this abound. "The Library of Missing Datasets" – a project of artist and educator Mimi Onuoha – presents a list of datasets that *ought* to exist (e.g., because they might illuminate or help address a social problem) but do not. *See* MimiOnuoha, On Missing Data Sets, GITHUB, https://perma.cc/VMX2-YTGL. Further examples: Not one of the forty-two voluntary national reviews – state submissions for review of progress on SDGs – contained data on refugees. *See* Alice Grossman & Lauren Post, *Missing Persons: Refugees Left Out and Left Behind in the Sustainable Development Goals*, RESCUE.ORG, Sept. 2019, https://perma.cc/E5Z8-Y9Z4. In the United States, immigration advocates criticized the Immigration and Customs Enforcement Agency (ICE) for not collecting data on humans contracting the SARS-CoV-2 virus while in ICE detention who might die of the illness either while detained or once released or deported. Dan Glaun, *How ICE Data Undercounts COVID-19 Victims*, PBS, Aug. 11, 2020, https://perma.cc/T2V2-7PB5. Although criminal justice and policing are increasingly data driven, there is a dearth of standardized and rigorous data about police brutality. *See* Lynne Peeples, *What the Data Say about Police Brutality and Racial Bias – and Which Reforms Might Work*, NATURE (June 19, 2020), https://www.nature.com/articles/d41586-020-01846-z.

[32] *See* Linnet Taylor & Dennis Broeders, *In the Name of Development: Power, Profit and the Datafication of the Global South*, 64 GEOFORUM 229 (2015).

[33] Companies that enjoy control over data gathering platforms or devices hold both the power to accumulate data and the power to determine which data is being generated through those data infrastructures. Conversely, actors that have the power to decide what data needs to be generated will often influence which data infrastructures come into existence, which, given infrastructural path dependencies, will in turn determine what will continue to be datafied (and what will not become datafied).

— are not the only or even the primary means of data collection. Data is being collected, at significantly greater scales, from web-browsing activity and electronic communication, as well as through the use of other Internet-enabled products and services. Rather than given, data is being *captured* by cookies and other tracking technologies operated by companies in the data-collecting and selling business.[34] A variety of sensors in mobile phones (and other personal devices like tablets and wearables) collect data about location, positioning, speed of movement, air pressure, light levels, and cellular activity levels (in addition to data about the performance of the device itself). Any software (app) installed on these devices may collect additional types of data, with some collecting data even when the device is not being used. Sensors are also increasingly embedded in products (e.g., cars, refrigerators, smart TVs, etc.), physical infrastructure (e.g., bridges, water meters) and even in biological matter (e.g., biosensors).[35] Rather than being periodic, planned, and purpose-focused, collection of data is increasingly continuous, ubiquitously, and deployed for multitude of uses and re-uses.[36]

Many of the relevant decisions *about what becomes data* (i.e. what is being datafied) are made by commercial actors who exercise, to various degrees, control over data infrastructures through which data is ultimately generated and processed. The scale and speed at which data is generated through platforms and devices is significantly greater, as compared to data that is generated through surveys or scientific samplings. As a result, the various inequalities and power imbalances that exist within a given political economy become reproduced through data on a very large scale. Actors who control data infrastructures are often in a favored position to accumulate data. As data can be reused, including for purposes other than the original, these data reserves become attractive gap-fillers for users who lack their own infrastructures for data collection. The data philanthropy movement encourages data "donations" as a way of remedying situations of data deserts, and calls for open data similarly aim to increase availability of data to wider constituencies. Without rendering normative judgments on the success of these initiatives,[37] we note here that proliferating data produced under circumstances of concentrated control over its means of production puts those who possessed such control in a privileged position to determine how the world is being represented, (re)shaped, and governed.[38]

---

[34] Wolfie Christl, *Corporate Surveillance in Everyday Life: How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions*, CRACKED LABS (June 2017), https://perma.cc/CG29-B6J6.

[35] Patrika Mehrotra, *Biosensors and Their Applications – A Review*, 6 JOURNAL OF ORAL BIOLOGY & CRANIOFACIAL RESEARCH 153 (2016).

[36] The relativity and variability of data poses challenges for data regulation by law, as we explore below with regard to "personal data," *below* Section II.C, and complications for competition law analysis arising from cross-sectoral data use, *below* Section II.D.

[37] *See below* Section II.B (discussing who benefits form making public data available as open data) and Section III.D (contrasting differential access to data solutions with open data).

[38] The politics of knowledge production in a digitalized world is, of course, not a new topic. Critical information studies interrogate "structures, functions, habits, norms, and practices that guide global flows of information and cultural elements". *See* Siva Vaidhyanathan, *Afterword: Critical Information Studies: A Bibliographic Manifesto*, 20:2-3 CULTURAL STUDIES 292 (2006). Critical data studies track "the ways in which data are generated, curated, and how they permeate and exert power on all manner of forms of life". *See* Andrew Iliadis & Federica Russo, *Critical Data Studies: An Introduction*, BIG DATA & SOCIETY (Oct. 2016). We do not reproduce these discussions here since no deservingly full account would be possible. *See also* GOVERNANCE BY INDICATORS (Kevin E. Davis, Angelina Fisher, Benedict Kingsbury & Sally Engle Merry eds., 2012) (examining ways in which quantitative rank-able indicators act as technologies of governance by defining problems, theories of change, and proposed solutions, thereby influencing ways in which political, social, and economic decisions are being made). These critiques of knowledge production are distinct from a critique of "big data" on the grounds that it lacks proper sampling and is thus inherently biased. *Contrast* Nobuo Yoshida, *Revolutionizing Data Collection: From "Big Data" to "All Data"* WORLD BANK BLOGS, Dec. 11, 2014, https://perma.cc/Y97E-VG47.

Those who possess data also often control the terms under which others may (or not) access and use it. Thus, apart from non-existence in the first place, data that is available, in a sense that it *has been collected by someone*, can nonetheless be inaccessible or accessible only to certain constituencies and/or only under certain circumstances and conditions. Accessibility can be regulated by law and technical means, or it can be a function of organizational dynamics. Governments can mandate access to data, for example, through statutory and regulatory mechanisms requiring public reporting or provision of data to specific government agencies, competitors, or consumers. Commercial entities can employ legal instruments like contracts or licenses, often rooted in regimes of personal and intellectual property, to preclude access to data or to establish conditions for such access.[39] But even without recourse to such "legal technologies",[40] access to data can be regulated through technical means.

Companies can use cryptography within their hardware and software to minimize unauthorized access to stored data, and they can implement encryption protocols (e.g. hypertext transfer protocol secure or https) to protect data flows from unauthorized access. Using these tools, companies can, for example, prevent access to or devalue information that may be intercepted by the government.[41] Similarly, however, they can, if compelled by law or otherwise, provide access to users' data to a third party without user' knowledge by adding "backdoors" into software, particularly in cloud-provided services where the users are not offered a choice of whether or not to accept or download an update.[42]

Access to data can also be enabled or blocked through the adoption of particular standards. Making data available only in certain formats can make data practically inaccessible for certain uses or users.[43] For example, data that is not machine readable is harder to aggregate with other data and cannot be as easily used for machine-learning purposes. As Tarleton Gillespie observed, "[a] technology that facilitates some uses can do so only by inhibiting others."[44] Thus, for example, choice of file formats and coding protocols inherently determine not only which data is produced but who participates in its production and use. [45] Similarly, platforms' boundary resources (resources which facilitate the use of core platform functionality to build applications) — such as software development kits (SDKs) and

---

[39] *See below* Section II.B.

[40] This terminology highlights the malleability of law and the role of lawyers in using and adapting legal instruments to further the interests they represent. *See* Kevin E. Davis, *Contracts as Technology*, 88 NYU LAW REVIEW 83 (2013) (analyzing innovation in contractual documents).

[41] For example, Google and Microsoft started deleting many identifiers associated with web searches from their databases after six to nine months to provide some level of anonymity to users and simultaneously diminished the risks associated with unauthorized surveillance and data breaches. *See* Peter Fleischer, Jane Horvath & Alma Whitten, *Another Step to Protect User Privacy*, GOOGLE BLOG (Sep. 8, 2018), https://perma.cc/5G4X-9SZ9; Kevin J. O'Brien, *Microsoft Puts a Time Limit on Bing Data*, N.Y. TIMES (Jan. 19, 2020), https://perma.cc/4XJF-7XAH.

[42] Christopher Soghoian, *Caught in the Cloud: Privacy, Encryption, and Government Back Doors in the Web 2.0 Era*, 8 JOURNAL ON TELECOMMUNICATIONS AND HIGH TECHNOLOGY LAW 359 (2010).

[43] In the United Kingdom, nearly 16,000 Sars-CoV-2 infections went initially unreported, complicating contact tracing efforts, because Public Health England, the national health agency, had used an outdated file format to collate test results. *Covid: Test Error 'Should Never Have Happened' – Hancock*, BBC (Oct. 5, 2020), https://perma.cc/89LG-WMAR.

[44] Tarleton Gillespie, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018).

[45] Content producers and viewers who work with video formats different from those required by YouTube, or who produce videos at length exceeding that allowed by YouTube, may be unable to either post or access content and may be forced to alter or subtly shape the videos themselves. Tarleton Gillespie, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018).

application programming interphases (APIs) — control who can access platforms' hardware, operating systems, and data (metrics, analytics, metadata, etc.) and under what conditions. Such access is critical to creating complementary applications or services.[46]

Protocols can also be deployed to prevent third parties from harnessing users' data. For example, a new version of Apple's Safari browser blocked by default all third-party cookies, preventing other companies from tracking users across multiple websites. Apple also allowed users to see which trackers had been blocked. According to one report, the Safari browser blocked 90 trackers in 5 minutes, a vast majority of them being Google analytics.[47] Google has also announced that it will move towards third-party cookie blocking but over a period of two years. While pressured by concerns over users' privacy, these companies at the same time are seeking to secure ad-based revenue and thus the ability to regulate — via control over infrastructure – who is able to accumulate — and derive value from – certain data about internet users, as well as when and how they are able to access it.[48]

This kind of infrastructural control over data flows is not dependent on legal technologies, though violations of contractually agreed policies are sometimes cited to justify these moves. For example, Apple used its control over the AppStore and the operating systems of iPhones to block Facebook from operating a data collection app, alleging that it violated the terms of its enterprise certificate policy.[49] In the summer of 2020, Apple updated its iOS to limit the data-gathering ability of third parties, making it more difficult for advertisers (and advertising revenue dependent platforms, such as Facebook) to track Apple customers, citing privacy concerns, while benefitting from continued preferential access to user data.[50] In another example, Google used its control over the desktop and mobile versions of its Chrome browser to nudge users towards encrypted domain name system (DNS) services,[51] on which the resolution of domain names (e.g. nyu.edu) into numbers (IP addresses, such as 216.165.47.10) depends. While supported by many privacy and security advocates, this move deprived Internet Service Providers (ISPs), which had traditionally provided unencrypted DNS resolution services, from monitoring the Internet use of their customers and also imperiled regulatory

---

[46] Ahmad Ghazawneh & Ola Henfridsson, *Balancing Platform Control and External Contribution in Third-Party Development: The Boundary Resources Model*, 23 INFORMATION SYSTEMS JOURNAL 173 (2013).

[47] John Koetsier, *Apple's New Browser Blocked 90 Web Trackers In 5 Minutes,* FORBES (Sep. 17, 2020), https://perma.cc/R8PL-6VKM.

[48] For a summary of the debates, *see* Dieter Bohn, *Google to 'Phase Out' Third-Party Cookies in Chrome, But Not for Two Years*, THE VERGE (Jan. 14, 2020), https://perma.cc/R5WG-4SVD.

[49] Josh Constine, *Apple Bans Facebook's Research App that Paid Users for Data*, TECH CRUNCH (Jan. 30, 2019), https://perma.cc/W6DH-Q692.

[50] Reed Albergotti & Elizabeth Dwoskin, *Apple Makes a Privacy Change, and Facebook and Advertising Companies Cry Foul*, WASHINGTON POST (Aug. 28, 2020), https://perma.cc/JZP6-3T4C (citing Nick Jordan, founder of advertising data consulting company Narrative I/O as follows:

> I think there's probably 30 percent truth in that they're doing it for privacy reasons, and it's 70 percent that they're doing it because it's what's good for Apple. It's a question for regulators and courts whether they should be able to wield the power they do over this ecosystem. They created it, but can they rule it with an iron fist?).

[51] Michael Grothaus, *Google Chrome Gets Major Privacy Boost: Here's how to Enable DNS-over-HTTPS*, FAST COMPANY (May 20, 2020), https://perma.cc/F6E4-GTPH. Arooj Ahmed, *Google is Extending Secure DNS to Chrome 85 for Android Soon*, DIGITAL INFORMATION WORLD (Sep. 4, 2020), https://perma.cc/RK3A-JNZ4. Mozilla's Firefox browser made a similar move, but the lobbying campaign by ISPs focused on Google. *See* Jon Brodkin, *Firefox Turns Encrypted DNS on by Default to Thwart Snooping ISPs*, ARS TECHNICA (Feb. 25, 2020), https://perma.cc/WF2V-B59L.

interventions dependent on such ISP-enabled monitoring.[52] All these developments illustrate corporate infrastructural control over data flows that at least complicate the "free flow of data" narrative.[53]

Even where data is available and not made inaccessible through legal or technical means, organizational challenges that result in failures to maintain or update the relevant infrastructures or which, through path dependencies, insist on retaining legacy systems, can reduce the usability of data and in a sense make it "unavailable". The COVID-19 pandemic drove this point home during spring 2020, when US social security authorities struggled to implement increased unemployment benefits, because its mainframe computers ran a 60-year old programming language called COBOL for which programmers were lacking.[54]

The aforementioned examples illustrate how control over data infrastructures can enable the acquiring and amassing of data and conveys power over datafication. In the ensuing section we explore infrastructural control in more detail.

C. Disentangling Infrastructures

Infrastructure is a notoriously ill-defined and ubiquitous term — not unlike data. There is and can be no uniform definition of infrastructure as infrastructures are inherently context-dependent and relational: what is infrastructural for some, might not be infrastructural for others. In this paper, we take our inspiration from the field of infrastructure studies.[55] One key insight of this interdisciplinary research agenda studying the history, development, operation, maintenance, and decay of infrastructures is to see infrastructures as not mere "objects". Fiberoptic cables or data centers are not infrastructures in and of themselves. They become *infrastructural* only when considering the social, economic, and political contexts within which processes, practices, norms, and rules bring about their existence, geographical positions, ownership and control structures, and operations, thereby establishing connections to other components and infrastructures. The uptake of this approach for legal scholars is that all these dimensions are not just entangled with each other — so that their disentanglement might provide analytical value — but also with various legal instruments, whether public or private, international or domestic.[56]

Our reference to *data infrastructures* in this paper calls attention to the technical, social, political, organizational, and legal dimensions of complex assemblages that capture, generate, categorize, standardize, aggregate, modify, (re)assemble, (re)interpret, transfer, or use data for a variety of purposes. The technical dimension of such data infrastructures consists itself of various components

---

[52] Mark Jackson, *Google, UK ISPs and Gov Battle Over Encrypted DNS and Censorship*, ISPREVIEW (Apr. 22, 2019), https://perma.cc/K8E3-A7VQ; IT Pro Team, *DNS Shakeup Could Kill ISP Filters*, IT PRO (June 22, 2019), https://perma.cc/89Y4-S8X4.

[53] *See below* Section II.A.

[54] Ian King, *An Ancient Computer Language Is Slowing America's Giant Stimulus*, BLOOMBERG (Apr. 13, 2020), https://perma.cc/H8W9-R2T9.

[55] *See, e.g.*, Paul N. Edwards, Geoffrey C. Bowker, Steven J. Jackson & Robin Williams, *Introduction: An Agenda for Infrastructure Studies*, 10(5) JOURNAL OF THE ASSOCIATION FOR THE INFORMATION SYSTEMS 364 (2009); Benedict Kingsbury, *Infrastructure and InfraReg: on Rousing the International Law 'Wizards of Is'*, 8 CAMBRIDGE INTERNATIONAL LAW JOURNAL 171 (2019).

[56] The infrastructural analysis here is hence the basis for our discussion *below* in Part II.

of digital infrastructure on different layers (hardware and software, localized or decentralized computing and storage facilities, inter-networking and cloud computing capabilities etc.). The social dimension draws attention towards variegated human (and human-machine) interactions implicated in such data-related activities, such as social practices, community norms, or the individual behavior of data subjects (who intentionally or inadvertently form part of data infrastructures). The organizational dimension, which we emphasize by focusing on corporate actors, asks which structures and processes hold data infrastructures together, and explores how governance and decision-making structures shape decisions about datafication.

Data infrastructures exist in a variety of contexts and on many scales. Some are domain specific (e.g., health information exchanges) whereas others are more general (e.g., behavioral user data); some are accessible by all (e.g., the Humanitarian Data Exchange)[57] whereas others are closed and made available only to a circumscribed group of people (e.g., student performance records available only to parents and teachers); some have a for-profit use purpose (e.g., Facebook's behavioral data infrastructure for targeted ads), others can be used for any purpose (e.g., open government data infrastructures),[58] and yet others are made available on the promise of non-commercial use (e.g., Equinor's data on the decommissioned Volve oil field in the North Sea).[59] Some data infrastructures are managed by government entities (e.g., population census) while others are managed by private commercial or non-profit bodies. Increasingly, the "public" and the "private" become blurred, with many data infrastructures involving both commercial and not-for-profit actors. Some data infrastructures are intensely local (e.g., Barcelona's DECODE project) while others are transnational (e.g., UN Statistical Division's Federated Data Model for SDG data). Often, however, data infrastructures are *both* local (e.g., collecting data about individuals or communities) and transnational (e.g., data can be collected and processed by multinational corporations, stored across data centers in different jurisdictions, or used by constituencies dispersed around the globe). Data infrastructures produce, store, modify, or transport data. They operate in different contexts, with different types of data, and involve different actors and different interests. Many data infrastructures generate data purposefully – sometimes at regular or specified intervals, with the entire process of collection, aggregation, processing and use governed by rules, norms, and laws. Increasingly, however, data is also being generated passively, continuously, and incidentally through platforms, sensors, and devices.[60]

The new means, scope, and speed of data generation has resulted in creation of sprawling jurisdictionally unbounded data infrastructures, connecting sensors and other devices, extracting,

---

[57] The Human Data Exchange, https://perma.cc/FG6Q-PRYH.

[58] The re-purposing of governmental data is a key narrative in the open data movement. *See* Jonathan Gray, *Towards a Genealogy of Open Data*, GENERAL CONFERENCE OF THE EUROPEAN CONSORTIUM FOR POLITICAL RESEARCH IN GLASGOW, WORKING PAPER (Sep. 3-6, 2014), https://dx.doi.org/10.2139/ssrn.2605828. *See further below* Section III.D.

[59] Volve, EQUINOR, https://perma.cc/P89P-4KW3.

[60] Ambient data collection challenges a number of established binaries, including natural/digital (e.g., think digital twins), material/virtual (e.g., think 3-D printing, augmented and virtual reality devices), and bodies/objects (e.g., think wearables and implanted devices). These are human perceptions of cyber-embeddedness that can replicate or fill-in-the-gaps in perceptions of material reality (examples include wearables, biometric ID, digital monitoring, implanted devices). On the relationship between humans and machines more generally, *see* seminal work by Donna Harraway, *A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century*, *in* SIMIANS, CYBORGS AND WOMEN: THE REINVENTION OF NATURE 149-181 (1991).

cleaning, storing, aggregating, and otherwise processing data implicating individuals, communities, and environments. Different entities may exercise control at different points of the process by which a set of measurements about physical phenomena become usable and analyzable data. Thus, different entities may have different types of control. For example, producers of hardware that collects measurements have control over *how* those measurements are collected; that hardware may be sold, leased, or provided as a service to another party, pursuant to a contractual arrangement, and the recipient from that point on will exercise control e.g., over where the hardware is located. A company that designs software used to aggregate, clean, and otherwise process data exercises control over data, including through standards and formats. That software can be similarly either sold or provided as a service, thereby transferring certain elements of control over data to their recipient. Companies that operate data management platforms will further exercise control over how processed data is analyzed, how insights of the analysis are presented, what type of access is given to the users, for what purposes, and so on.

This requires us to confront a paradox: how is it even possible to exercise control over large-scale, highly complex, and often geographically distributed data infrastructures?

D. Identifying Control over Infrastructure

Control can be exercised either over different infrastructural components or over certain infrastructures as whole. While control is never absolute, control over critical elements (e.g., a particular protocol, application, or operating system) can be sufficient to secure overall control both over data and over the power to datafy.

The Internet, a technically distributed but increasingly economically centralized data infrastructure with centralized points of governance and control, cloud computing, and software have been (and continue to be) foundational not only for the development and growth of platforms, but also in enabling control over ever more expansive data infrastructures. The Internet enables data transfers and serves as foundational infrastructure for cloud computing, which provides the overall organization for today's data infrastructures.[61] Cloud computing enables storage and processing of very large data sets in distributed fashion (leading to cost efficient locating of data) and offers enterprise services that open up new possibilities for data use. Software (and specifically software that extracts, cleanses, aggregates, processes, and analyzes data) can modify and extract data to convert it into economic value. This trifecta has not only facilitated and accelerated the extraction of value from data but has also opened up possibilities in the development of artificial intelligence, with wide-ranging applications. This, in turn, has fueled demand for ever larger amounts of data, while often ignoring negative externalities.[62]

---

[61] *Cf.* Nick Couldry & Ulises A. Mejias, THE COSTS OF CONNECTION: HOW DATA IS COLONIZING HUMAN LIFE AND APPROPRIATING IT FOR CAPITALISM ch. 2 (2019) (noting that "the Cloud Empire is the *what*, the overall organization of resources and imagination that emerges from the practices of data colonialism").

[62] Former Google CEO Eric Schmidt famously observed: "big data is so powerful, nation states will fight [over it]". Rob Price, *Alphabet's Eric Schmidt: 'Big data is so Powerful, Nation States Will Fight' Over it*, BUSINESS INSIDER (Mar. 9, 2017), https://perma.cc/2R8Z-VVZW. On the downsides of ever larger data sets for natural language processing, *see* Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, FACCT '21, Mar, 3-10, 2021.

Platform companies are often in a favorable position to generate and accumulate data.[63] The cross-over between platform studies and infrastructure studies illuminates how platform companies have acquired infrastructural significance in a range of economic and social functions and became not only dominant data holders but also significant shapers of the digital world.[64] By serving as intermediaries and by bundling up various services, they may consolidate control over data infrastructures. Their corporate organization gives them legal grounding for the "platformization of infrastructures" and the "infrastructuralization of platforms".[65] They may cement infrastructural control over data by creating "walled gardens" within which data is generated, accumulated, concentrated, and protected.[66]

In other words, platform companies — themselves infrastructures for e-commerce, communication, software services, or other transactions — are also data infrastructures.[67] The congruence between corporate control over dominant platforms and the resulting data generating and shaping capacity leads to concentrated control over data and results in outsized power to datafy. This is not to say that only platform companies enjoy infrastructural control over data. There are other corporate actors that are not platforms – understood as intermediaries for two-sided markets, which control important and large-scale data infrastructures, for example network operators and networking service providers (such as Cloudflare), networking equipment vendors (such as Nokia, Huawei, Ericsson, and Juniper), or manufacturers and operators of IoT devices. Yet, platforms companies are particularly potent data generators, as the following examples illustrate.

E-commerce platforms (e.g., Amazon, Shopify, Alibaba, JD), social media platforms (e.g. Facebook, YouTube, Twitter, WeChat, etc.), search engines (e.g., Google, Baidu), and smartphones' operating systems and app marketplaces (e.g., Apple's iOS with the App Store or Google's Android with Google Play) generate gigantic amounts of data, which their corporate owners hoard and use to shape the world via data.[68] These companies are not only infrastructural for all forms of connectivity (technical, economic, social, political) but they are also data infrastructures that create identities, foster social

---

[63] The term digital "platform" has been used to refer to a variety of different online structures – and corresponding business models – that enable a wide range of activities between different actors for different purposes. *See* for a discursive take on the meaning of platform, Tarleton Gillespie, *The Politics of 'Platforms'*, 12 NEW MEDIA & SOCIETY 347 (2010). In communication studies, platforms denote "sites and services that host public expression, store it on and serve it up from the cloud, organize access to it through search and recommendation, or install it onto mobile devices". *See* Tarleton Gillespie, *Governance of and by Platforms*, *in* THE SAGE HANDBOOK OF SOCIAL MEDIA (Jean Burgess, Thomas Poell & Alice Marwick eds., 2017). In competition law and economics, platforms are often analyzed as 'two-sided markets' "when (1) a single transaction takes place between two different groups of users connected by the platform, and (2) the numerosity of each group of users creates reciprocal inter-side positive externalities." *See, e.g.*, Giacomo Luchetta, *Is the Google Platform a Two-Sided Market?*, 10 JOURNAL OF COMPETITION LAW & ECONOMICS 185 (2013). According to the OECD, a platform facilitates interactions between two or more distinct but interdependent sets of users (whether firms or individuals) who interact through the service via the Internet. *See* OECD, AN INTRODUCTION TO ONLINE PLATFORMS AND THEIR ROLE IN THE DIGITAL TRANSFORMATION (2019), https://doi.org/10.1787/53e5f593-en.

[64] Jean-Christophe Plantin, Carl Lagoze, Paul N Edwards & Christian Sandvig, *Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook*, 20 NEW MEDIA & SOCIETY 293 (2018).

[65] *Id.*

[66] *Id.*

[67] Note that this is sometimes conveyed by saying that X is not just a Y company but also a "data company".

[68] Data brokers – an emerging and highly profitable industry – feed off other platforms' data generating capacity as they scrape websites, buy data captured and/or aggregated by others, reaggregate and repackage them and offer this data or "data insights" derived from it for sale. *See* Leanne Roderick, *Discipline and Power in the Digital Age: The Case of the US Consumer Data Broker Industry*, 40(5) CRITICAL SOCIOLOGY 729 (2014).

practices, shape economic relations, and birth new industries. Importantly, their ability to produce, collect, analyze, and intensely protect "their" data from competitors is supported not solely, or even mainly, by resorting to legal forms of protections, but by exercising control over different layers or components of infrastructure and by acquiring further data infrastructures, as needed.

Consider, for example, e-commerce platform companies, which amass behavioral data by surveilling the commercial activity of sellers and buyers. Amazon's platform operates as an infrastructure for commerce,[69] which allows it to capture data of buyers and sellers. The company uses clickstreams — digital breadcrumb trails — to monitor which sites users come from, how they move through Amazon's pages and where they go to next. [70] Amazon and its peers, like the Chinese giant Alibaba, developed corporate organizations to integrate their operations across markets, which has allowed these platforms to establish advantageous positions in adjacent markets,[71] which in turn, gives them access to more data. E-commerce platforms have invested heavily in logistics, payment, and even microfinance infrastructures, thereby obtaining control over key infrastructures for online trade of goods and services.[72] Alibaba also captured the electronic payment market with creation of Alipay and, through numerous acquisitions, entered the entertainment market and social media markets as well. Retailers and consumers became increasingly dependent on these infrastructures, thereby ensuring continued supply of data about sales, payments, user activity etc. As Alibaba's executives have been declaring since 2016, Alibaba is not a retail company, it's a data company.[73] Similarly, one of Amazon's former executives, James Thomson, said of Amazon "[they] happen to sell products, but they are a data company. Each opportunity to interact with a customer is another opportunity to collect data".[74]

---

[69] K. Sabeel Rahman, *Private Power, Public Values: Regulating Social Infrastructure in a Changing Economy*, 39 CARDOZO LAW REVIEW 5 (2017).

[70] Lina M. Khan, *Sources of Tech Platform Power*, 2 GEORGETOWN LAW TECHNOLOGY REVIEW 325 (2018). Amazon's data gathering practices have been subject to scrutiny by US antitrust investigations, which revealed how Amazon uses data about retailers' value chain to expand into other sectors. To ensure continuous access to consumers' behavioral data (as well as retaining and building customer base for its own products), Amazon reportedly monitors communications between third-party marketplace merchants and consumers and penalizes those merchants who direct consumers to their own sites or other sales channels. Amazon also convinced smaller third-party retailers to sell items via its Marketplace, offering to share with them customer analytics while retaining complete access to and control over that data, effectively "renting the Amazon customer" to third-party sellers. U.S. House Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary, Majority Staff Report and Recommendations, INVESTIGATION OF COMPETITION IN DIGITAL MARKETS (2020); *See also* Leo Kelion, *Why Amazon Knows so Much About You*, BBC, https://perma.cc/P3K7-9AQC. *See below* Section II.D. Amazon has captured a significant (and leading) market share in the United States and is consistently increasing its market share globally. Enjoying market dominance has allowed Amazon to exercise "gatekeeping power" to cement its hold on data against competitors. *See below* Section II.D.

[71] Lina M. Khan, *Sources of Tech Platform Power*, 2 GEORGETOWN LAW TECHNOLOGY REVIEW 325 (2018).

[72] *See* for an interesting case study of Alibaba.com, Seung Ho Park & Ziqian Zhao, ALIBABA GROUP: FOSTERING AN E-COMMERCE ECOSYSTEM (2016). Already in 2011, Alibaba, announced that it was spending close to US$4.5 billion on logistics and on building out integrated warehouse networks across China. Alibaba started to offer loan applications to retailers online, using Alibaba-developed credit assessment models as well as behavioral data generated by the sellers' daily transactions. The resulting credit and risk assessments produced additional data. *See*
Alibaba Establishes Small Loans Lender in Chongqing, TMT CHINA WEEKLY (June 24, 2011).

[73] Alizia Staff, *Five Reasons Why Alibaba is a Data (Not E-Commerce) Company*, ALIZILA (Oct. 17, 2016), https://perma.cc/B5TJ-3ZZJ.

[74] Leo Kelion, *Why Amazon Knows so Much About You*, BBC, https://perma.cc/P3K7-9AQC; U.S. House Subcommittee on Antitrust, Commercial and Administrative Law of the Committee on the Judiciary, Majority Staff Report and Recommendations, INVESTIGATION OF COMPETITION IN DIGITAL MARKETS (2020)

Both Amazon and Alibaba eventually developed independent cloud infrastructures to enable their enterprises. Cloud computing allows platforms (of all kinds) to centralize control over data by establishing parameters for how data is produced, stored, and shared. Cloud providers offering software- and platform-as-a-service and data analytics firms provide specialized services on the basis of collected data, often using data extracted from the very constituencies that subsequently consume the companies' services.[75] Access to additional data generated through cloud computing services can also be used by the providers to train machine learning algorithms, which are subsequently offered as a service as well.[76] Although there are smaller cloud providers and a spectrum of data analytics companies, some of which focus on specific domains such as health or agricultural data, many of the big tech companies — including Alphabet, Alibaba, Baidu and Tencent — perform multiple functions, operating as platforms, cloud service providers, and data analytics companies. Such integration not only concentrates control of such companies over data infrastructures but also creates path dependencies for other businesses, consumers, and the public, thereby commanding "loyalty" when "exit" and "voice" are not viable.[77]

Social media giants like Facebook, YouTube, Twitter, and Tencent, have similarly been able to amass large stores of behavioral data albeit through different business models. Advertising revenue dependent business models rely on platforms' ability to create and manipulate social practices that constitute data. Platforms leverage behavioral data of users by subjecting them to granular analysis, "matching" users' desires and interests with products whose companies than place targeted ads. Any subsequent interaction with that product in turn produces additional data. For many companies, the ability to create, surveil, and affect "data doubles"[78] (i.e., digital representations of individuals, communities, and environments) across networked spaces and territorial borders is central to their business models, leading them to pursue corporate acquisitions of products and other data-generating platforms that could be added on to expand their existing infrastructures (e.g., Facebook's acquisition of social media data-gathering platforms Instagram and WhatsApp, Google's acquisition of health data-gathering device FitBit, Alibaba's acquisition of UC Browser developer UCWeb etc.). The integration of products and services that strengthen and expand data infrastructures through corporate acquisitions also consolidate and entrench control over data via such infrastructures.

Infrastructural control over data allows companies not only to extract their users' data but also to *shape* what data they gather from user behavior. For example, despite content being supposedly user-generated, the extent to which content is mediated by social practices (e.g., by content moderators), manipulated through technical means (e.g., by virtue of algorithmic targeting), and removed through a combination of technical, social, and organizational means (e.g., via blocks, suspensions, and bans) places significant constraints on individuals' agency. The entire engagement can be seen as "manufactured" because users are being nudged to engage (or not) with the advertised material or product, and each iterative (non)engagement is registered as new data inputs. Such behavior-shaping power extends beyond selling products. Users are being targeted with particularized information (e.g.,

---

[75] For examples of types of data collected by cloud service providers, *see* Privacy Notice, AMAZON WEB SERVICES, https://perma.cc/BK9P-Z7DY.

[76] We thank Niels ten Oever for highlighting this point to us.

[77] Albert O. Hirschman, EXIT, VOICE, AND LOYALTY (1970).

[78] Kevin D. Haggerty & Richard V. Ericson, *The Surveillant Assemblage*, 51 THE BRITISH JOURNAL OF SOCIOLOGY 4 (2000); Linnet Taylor & Dennis Broeders, *In the Name of Development: Power, Profit and the Datafication of the Global South*, 64 GEOFORUM 229 (2015).

"suggested posts" on Instagram) deemed of potential interest on the basis of users' demographic or behavioral data. This is not just a binary matter of showing or not showing content, but encompasses subtler forms of manipulation, for example by shifting the order of search results and news feeds.[79]

Technical protocols and centralized control also define and structure spaces within which users can conduct their array of activities.[80] This is poignantly illustrated by Catherine D'Ignazio and Lauren Klein in their account of Facebook's override of user choices. In 2014, Facebook expanded the gender categories available to registered users from the conventional binary (male/female) to over fifty choices, ranging from 'Genderqueer' to 'Neither'. A year later, Facebook replaced the select-from-options model with a blank text field, a decision that was touted as being very progressive. D'Ignazio and Klein note, however, that

> …below the surface, Facebook continues to resolve users' genders into a binary: either "male" or female". Evidently, this decision was made so that Facebook could allow its primary clients – advertisers – to more easily market to one gender or the other. Put another way, even if you can choose the gender that you show to your Facebook friends, you can't change the gender that Facebook provides to its paying customers. …corporations like Facebook, and not individuals…, who have the power to control the terms of data collection."[81]

The ability to control data infrastructures is a powerful form of control over social, political and economic organizations of human life.[82] Contesting that power is challenging. Data infrastructures exhibit a high degree of opaqueness if not invisibility.[83] How Facebook builds user profiles or what algorithm Amazon uses to determine customer's purchasing power is not known. The datasets on which algorithms are trained and the code of internal data management software are rarely revealed. Although companies are increasingly deploying open source software and open standards, for the most part how (and by whom) control over data is exercised in different contexts is neither apparent nor easily ascertained. Similarly, how much and what kind of data companies control might not even be known in a holistic fashion internally, but is entirely inscrutable from the outside.[84] This opacity is partly purposeful, e.g., where the technology is proprietary or when concealing infrastructure provides a market advantage, partly a function of expertise and special knowledge required to understand certain components and layers of data infrastructures, and partly a function of the sheer megalomaniac size of data infrastructures that escape internal and external scrutinization due to their complexity.

---

[79] Robert Epstein & Ronald E. Robertson, *The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcome of Elections*, 112 PNAS E4512 (2015), https://doi.org/10.1073/pnas.1419828112. Platform's penchant for user manipulation can have stark effects on democratic decision-making, as Jonathan Zittrain has observed in *Engineering an Election*, 127 HARV. L. REV. F. 335 (2014). *See also*, Evelyn Douek, *Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability*, 121 COLUMBIA LAW REVIEW 1 (2021), https://dx.doi.org/10.2139/ssrn.3679607.

[80] Julie E. Cohen, BETWEEN TRUTH AND POWER 42 (2019).

[81] Catherine D'Ignazio & Lauren F. Klein, DATA FEMINISM 98-100 (2020). This was discovered by Rena Bivens, *The gender binary will not be deprogrammed: Ten years of coding gender on Facebook*, 19 NEW MEDIA & SOCIETY 880 (2017).

[82] Laura DeNardis, THE INTERNET IN EVERYTHING 451 (2020).

[83] Frank Pasquale, THE BLACK BOX SOCIETY (2015).

[84] *See below* Section II.C on why data protection law does not generate adequate transparency and further below Section III.C. on how more transparency over data could be demanded.

Compounding the problem of contestation is the increasing ubiquity (and in some case addictiveness) of data-generating platforms and devices through which the platforms are accessed, and the heightened degree to which data infrastructures are being embedded and routinized in daily lives, all of which causes the individuals' awareness of "behind the scene" control to recede. These conditions are neither incidental nor coincidental but rather a product of deliberate design choices that render any impression of choice and agency exemplified, among other things, through consent "clicks" or service agreements, ultimately an illusion.[85]

Importantly, and particularly relevant for developing countries, the unequal distributive impacts of data infrastructures are a phenomenon not limited to big tech platforms. The increasing deployment of sensors for data collection, for example in digital agriculture, illustrates similar effects in the context of farming data. Sensors in farming equipment, such as tractors, are often linked to specific data management platforms, like Climate Field View, run by Climate Corporation.[86] The Climate Corporation and other big agriculture companies, like John Deer, have invested heavily in technologies that use detailed data on soil, seed, weather, etc. to provide ostensibly useful insights to farmers (for example, how to increase yields).[87] This has meant in practice that farmers who want to benefit from data-driven farming supply data to data management platforms, who in turn are not only able to dictate the terms (including compensation) due to power imbalances but also to entrench control by creating farmers' dependencies on particular data management platforms, including through proprietary software and the limited interoperability of sensors embedded in farming equipment with their platforms. As Sarah Rotz and her colleagues have aptly summarized:

> … farmers and farm workers continue to carry the material risks and bear the livelihood impacts of agriculture while the capital gains of digitalization are, largely, extracted by data management companies. Indeed, agricultural data have significant use value because they are an essential tool for these companies' platform and predictive algorithm development. As with capitalist modes of banking, farmers deposit their data (money) into the system. These data are then used (reinvested) by the companies to make a profit. In effect, some farmers are becoming 'digital labourers', while data management companies accumulate the economic benefits via the expansion of their knowledge systems—the new digital commodity. This is similar to the capital accumulation models of social media platforms such as Facebook and Google.[88]

---

[85] Critical privacy scholarship has examined the limitations of consent at length. *See e.g.*, Woodrow Hartzog, PRIVACY'S BLUEPRINT: THE BATTLE TO CONTROL THE DESIGN OF NEW TECHNOLOGIES (2018) (arguing that technologies are designed to undermine privacy); Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub & Thorsten Holz, *(Un)informed Consent: Studying GDPR Consent Notices in the Field*, PROCEEDINGS OF THE 2019 ACM SIGSAC CONFERENCE ON COMPUTER AND COMMUNICATIONS SECURITY 973 (2019) (presenting empirical evidence about how users are being nudged towards giving "consent").

[86] Monsanto acquired Climate Corporation for nearly one billion US dollar in 2013. *See* Michael Specter, *Why the Climate Corporation Sold Itself to Monsanto*, NEW YORKER (Nov. 4, 2013), https://www.newyorker.com/tech/annals-of-technology/why-the-climate-corporation-sold-itself-to-monsanto.

[87] Sjaak Wolfert, Lan Ge, Cor Verdouw & Marc-Jeroen Bogaardt, *Big Data in Smart Farming – A Review*, 153 AGRICULTURAL SYSTEMS 69 (2017).

[88] Sarah Rotz, Evan Gravely, Ian Mosby, Emily Duncan, Elizabeth Finnis, Mervyn Horgan, Joseph LeBlanc, Ralph Martin, Hannah Tait Neufeld, Andrew Nixon, Laxmi Pant, Vivian Shalla & Evan Fraser, *Automated Pastures and the Digital Divide: How Agricultural Technologies are Shaping Labour and Rural Communities*, 68 JOURNAL OF RURAL STUDIES 112, 117 (2019).

As data-collecting technologies become more diffuse and embedded in material systems, as their architecture becomes more complex and distributed, as path dependencies become cemented, [89] and as the entities controlling them become larger and more consolidated, without legal, regulatory, policy, and political intervention, opportunities for contestation of oversized control over data infrastructures, and thus of remedying of data inequalities, diminishes.

[89] Geoffrey C. Bowker, Karen Baker, Florence Millerand & David Ribes, *Towards Information Infrastructure Studies: Way of Knowing in a Networked Environment*, *in* INTERNATIONAL HANDBOOK OF INTERNET RESEARCH 97 (Jeremy Hunsinger, Lisbeth Kastrup, Matthew Allen eds., 2010). As Jean-François Blanchette notes, computing infrastructure "developed incrementally, from the progressive laying down of its infrastructural components, including data centers, fiber cables, economic models, and regulatory frameworks. Such incremental development means that early-stage design choices *persist*, often with unforeseen consequences and become increasingly difficult to correct as the infrastructure becomes ubiquitous, its functionality expands, and the nature of the traffic it serves evolves …" Jean-François Blanchette, *Introduction: Computing's Infrastructural Moment*, *in* REGULATING THE CLOUD: POLICY FOR COMPUTING INFRASTRUCTURE (Christopher S. Yoo & Jean-François Blanchette, eds.) 1, 3 (2015).

II. Legal Dimensions of Data Inequality

Extant law does not effectively address unequal control over data and unequal power to datafy, which are being enabled by concentrated control over data infrastructures identified in Part I. At best, law tends to ignore the infrastructural root causes of data inequality. At worst, it may contribute to or entrench data inequality. In this part, we explore legal dimensions of data inequality to substantiate these claims. Our analysis straddles several established domains of law that pertain to relevant components of data infrastructures, with a focus on data protection and privacy law, intellectual property law, and antitrust and competition law. We cannot, however, offer comprehensive legal analysis in this paper. For example, we bracket the important and intricate tax law questions raised by the tax avoidance strategies of globally operating corporations,[90] even though some claim that they are particularly pronounced in the digital economy.[91] We also do not explore corporate law and corporate governance interventions, despite their critical importance for the ways in which multinational corporations can exercise control over complex and transnationally diffused data infrastructures.[92]

Our analysis integrates domestic and international law because separation between these two "levels" is ultimately artificial, especially in the digital domain. Throughout, we highlight the growing importance of international economic law which has acquired characteristics of "megaregulation" in recent "comprehensive" trade and investment agreements.[93] These agreements increasingly create important secondary rules that shape and often constrain states' ability to regulate data flows, data ownership, data protection, and data competition.

Our aim in this part of the paper is to illustrate the complicated relationship between uneven control over data and data infrastructures and various fields of private and public law. In doing so, we are inspired by the analysis of the legal coding of capital that Katharina Pistor has pioneered and extended to the emerging legal coding of data, and the critical accounts by Julie Cohen and Amy Kapczynski about the ways in which platform companies have used legal technologies to advance their interests

---

[90] *See* on the evidence about tax avoidance generally Nadine Riedel, *Quantifying International Tax Avoidance: A Review of the Academic Literature*, 69 REVIEW OF ECONOMICS 169 (2018). *See also* Robert Bird & Karie Davis-Nozemack, *Tax Avoidance as a Sustainability Problem*, 151 JOURNAL OF BUSINESS ETHICS 1009 (2018). *See* for Latin American and Caribbean regional perspectives on taxation and the digital economy the various contributions in the online symposium convened by Monica Victor at AfronomicsLaw. Symposia: Forthcoming Feature Symposium, AFRONOMICSLAW, https://www.afronomicslaw.org/symposia/. *See* on the OECD's multilateral initiatives and their impact on international tax law: Ruth Mason, *The Transformation of International Tax*, 114 AMERICAN JOURNAL OF INTERNATIONAL LAW 353 (2020).

[91] *See*, e.g., Grant Richardson & Grantley Taylor, *Income Shifting Incentives and Tax Haven Utilization: Evidence from Multinational U.S. Firms*, 50 THE INTERNATIONAL JOURNAL OF ACCOUNTING 458 (2015) (finding that multinationality, transfer pricing aggressiveness, thin capitalization and intangible assets are positively associated with tax haven utilization). The European Commission took action against Ireland for enabling such tax avoidance strategies, framing its arrangements with Apple as illegal state aid amounting to EUR 13 billion in unlawful tax advantages. The EU's General Court annulled this decision in Cases T-778/16 and T-892/16, *Ireland and Others v. European Commission*, ECLI:EU:T:2020:338. The European Commission has appealed the judgment to the EU's Court of Justice.

[92] *See above* Section I.C. *See* John Ruggie, *Multinationals as Global Institution: Power, Authority and Relative Autonomy* 12 REGULATION & GOVERNANCE 317 (2018) (contrasting corporate social responsibility with imposition of binding legal obligations on multinational enterprises).

[93] *See* Benedict Kingsbury, Paul Mertenskötter, Richard B. Stewart &Thomas Streinz, *TPP as Megaregulation*, *in* MEGAREGULATION CONTESTED: GLOBAL ECONOMIC ORDERING AFTER TPP ch 2 (Benedict Kingsbury et al eds., 2019). *Contrast* HANDBOOK OF DEEP TRADE AGREEMENTS (Aaditya Mattoo, Madia Rocha & Michele Ruta eds., 2020).

in an information-capitalist economy.[94] The goal is not to blame "the law" for data inequality but to make visible how different domains of law are entangled with data infrastructures and to show how lawyers can use different "legal technologies" to facilitate corporate control over data.[95] We certainly do not think that the many scholars, practitioners, and activists who are engaged in the legal domains of intellectual property, data protection and privacy, or competition law are not addressing important issues. Our argument is much narrower: these domains of law are often not attuned to infrastructural control over data and therefore do not effectively regulate data inequality (at best) or might entrench it (at worst). This is also not to say that all these domains of law need to change to confront data inequality. As we explore in Part III below, remedying data inequality requires carefully calibrated interventions within the law and beyond.

In the following, we show, first, how the legal system has facilitated the global "free flow" of data through the Internet, which crucially relies on physical infrastructures and interoperability standards, without much concern as to between whom data flows and where it accumulates. We then turn, second, to the much-discussed question of legal data ownership. We find that even where the law recognizes no property rights in data, control over data is achieved through control of the relevant data infrastructures and encoded in contractual terms irrespective of formal property rights. However, when this status quo is being challenged by demands for transparency or data sharing, lawyers are likely to invoke property or property-like rights in data and data holders will lobby for the recognition of such rights. Third, we address the dominant approach to contemporary data regulation: rights-based data protection and privacy law. While we recognize the importance of this domain and the global diffusion of data protection and privacy rights, we ultimately conclude that this rights-based approach does not effectively confront data inequality. At best, it raises the costs of data accumulation but has not, at least thus far, effectively curtailed data hoarding. At worst, it privileges those with the means to shoulder increased compliance costs, thereby inadvertently exacerbating data concentration. Fourth and finally, we discuss the law applicable to platform companies. We recognize the evolving debate about whether and how antitrust and competition law should confront platform power, and we concede that this approach is more attuned to dynamics of infrastructural control than the other areas of law we explore. But we also caution that antitrust and competition law come with certain assumptions about market efficiencies and consumer welfare that can make it blind to broader concerns around data inequality, which ought to be addressed through other means.

A. "Free Flow" of Data

The global expansion of the Internet since the early 1990s has enabled an unprecedent degree of inter-connectedness of communication networks and devices. This development was initially hailed as

[94] Katharina Pistor, THE CODE OF CAPITAL: HOW THE LAW CREATES WEALTH AND INEQUALITY (2019); Katharina Pistor, *Rule by Data: The End of Markets?*, 83 LAW AND CONTEMPORARY PROBLEMS 101 (2020); Julie E. Cohen, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM (2019); Amy Kapczynski, *The Law of Informational Capitalism*, 129 YALE LAW JOURNAL 1460 (2020).

[95] We adopt the "legal technologies" term from Kevin Davis, *Contracts as Technology*, 88 NEW YORK UNIVERSITY LAW REVIEW 83 (2013), who defines technology as useful knowledge about how to produce things at low cost (studying contract law innovations). We depart from the focus on innovations and instead highlight how lawyer's knowledge and skill can be used in data contracting, licensing, and property claims.

promoting individual and collective freedom globally,[96] despite persistent "digital divides" between and within countries.[97] Initially mostly perceived as a communication infrastructure, the Internet has evolved into an indispensable infrastructure for data generation, processing, and transfer of data more generally.[98] Its governance has been subject to increasing contestation for a variety of reasons ranging from the traditional but increasingly challenged preponderance of US stakeholders in Internet governance institutions,[99] unresolved tensions between multistakeholderism, intergovernmentalism, and nation states' jurisdiction,[100] and, more recently, concerns about the dominance of US and China based but globally operating digital corporations and the varieties of "surveillance capitalism" they orchestrate.[101] When the US government decided to facilitate the economic exploitation of the Internet during the early 1990s, it fostered a shift from public funding and public management towards commercialization and privatization.[102] Ever since, private sector leadership of the Internet has been a staple and mantra of Internet governance.[103]

The Internet has been remarkably successful in facilitating data transfers across borders. The relevance of data mobility for the unbundling of economic production, creating transnational networks in which goods and services are being exchanged and recombined, often within firms, seems beyond doubt yet remains difficult to quantify.[104] The Internet's ability to route information through inter-connected networks depends on physical connectivity (either through a cable or via the electromagnetic spectrum), interoperable protocols that govern the exchange of data between different networks — the proverbial inter-networking — and the absence of legal limitations imposed by governments or private actors with an interest in blocking, monitoring, or at least limiting cross-border data movements.

---

[96] *See* Ira C. Magaziner, *Creating a Framework for Global Electronic Commerce, Future Insight*, Release 6.1 PROGRESS & FREEDOM FOUNDATION (July 1999) (claiming that the Internet was promoting "individual freedom and individual empowerment" and that it would "bring all the peoples of the world closer together"). *See* critically David Pozen, *The De-Americanization of Internet Freedom*, LAWFARE, June 13, 2018, https://perma.cc/MR7D-X2NK; Jack Goldsmith, *The Failure of Internet Freedom*, KNIGHT FIRST AMENDMENT INSTITUTE AT COLUMBIA UNIVERSITY, June 13, 2018, https://perma.cc/4TVG-DZ3L.

[97] *See generally* Jan van Dijk, THE DIGITAL DIVIDE (2020). *See above* Section I.B on different forms of data inequality.

[98] *See above* Section I.C. *See also* Laura DeNardis, THE INTERNET IN EVERYTHING (2020) (describing the Internet's transformation from a communication into a control network).

[99] Andrew L. Russell, OPEN STANDARDS AND THE DIGITAL AGE: HISTORY, IDEOLOGY, AND NETWORKS (2014); Milton L. Mueller, NETWORKS AND STATES: THE GLOBAL POLITICS OF INTERNET GOVERNANCE (2010); Laura DeNardis. THE GLOBAL WAR FOR INTERNET GOVERNANCE (2014).

[100] Thomas Streinz, *Global Hybrid Internet Governance: The Internet Corporation for Assigned Names and Numbers (ICANN)* (draft paper, on file with author).

[101] The term has been popularized by Shoshana Zuboff, THE AGE OF SURVEILLANCE CAPITALISM (2019). On the geoeconomics and geopolitical interplay between corporations and governments engaged in surveillance activities, *see* Madison Cartwright, *Internationalising State Power Through the Internet: Google, Huawei and Geopolitical struggle*, 9 INTERNET POLICY REVIEW 1 (2020).

[102] *See* Brett Frischmann, *Privatization and Commercialization of the Internet Infrastructure: Rethinking Market Intervention into Government and Government Intervention into the Market*, 2 THE COLUMBIA SCIENCE AND TECHNOLOGY LAW REVIEW 1 (2001).

[103] *See* Madeline Carr, *Power Plays in Global Internet Governance*, 43(2) MILLENNIUM 640. Not creating a government-led or intergovernmental governance structure in its place was one of the conditions the US National Telecommunications and Information Administration (NTIA) postulated before giving up oversight over key administrative functions at the heart of the Internet's domain name system. *See* Kal Raustiala, *Governing the Internet*, 110 AMERICAN JOURNAL OF INTERNATIONAL LAW 491 (2016).

[104] Richard Baldwin, THE GREAT CONVERGENCE (2016); Milton Mueller & Karl Grindal, *Data Flows and the Digital Economy: Information as a Mobile Factor of Production*, DIGITAL POLICY, REGULATION AND GOVERNANCE (Jan. 2019).

In the following, we navigate the Internet stack to illustrate how law has interacted with different components of Internet infrastructure to enable global data flows while also creating, sustaining, or at least ignoring inequalities in terms of data access, control, and governance. As generally within this paper, our examples are meant to be illustrative.

We show how submarine cables' colonial path-dependencies are being accommodated by the international law of the sea. Today, their ownership and operation often implicate the same actors that enjoy concentrated control over complex data infrastructures. We illustrate how protocols enable interoperability between networks while social norms within global Internet standard-setting bodies limit private sector ability to avail themselves of intellectual property rights, thereby preserving inter-connectivity without allowing much space for countervailing interests. We expose how emerging rules in instruments of international economic law protect transnational data mobility against governmental control in form of cross-border data transfer limitations and territorial data localization requirements, thereby limiting states' ability to counteract data inequality in this way. Lastly, we conclude with examples that illustrate evolving internet-networking dynamics and power struggles on the world wide web (www), thereby exposing the power differentials between different networks on the Internet, mediated through contracting or private norms. The emergence of content delivery networks (CDNs), in particular, affects data flows in ways that seems in tension with established net neutrality principles without violating net neutrality laws.

The Internet can be modeled as consisting of several layers that together form a "stack" in which each layer facilitates a discrete function, on which layers above rely. At the bottom, and most fundamentally, is the need to establish a physical network connection. The internet's backbone for cross-border data flows consists of a global network of submarine fiberoptic cables, laid on the seabed.[105] Imperialism, free trade policies, and state subsidies shaped the locations and ownership of early submarine cables (used for telegraph services), connecting British, European, Japanese, and American empires to their colonies and dominions overseas.[106] These routes, laid down in the nineteenth century, remain the most important corridors for modern fiberoptic cables.[107] The extent to which states are able to regulate surveillance, construction, maintenance, and use of modern fiberoptic cables is guided to some extent by the international law of the sea.[108] Within their coastal waters, states' national public

[105] Nicole Starosielski, THE UNDERSEA NETWORK (2015); SUBMARINE CABLES: THE HANDBOOK OF LAW AND POLICY 2 (Douglas R. Burnett, Robert C. Beckman & Tara Davenport eds., 2014). The global cable network is estimated to comprise more than 400 separate submarine cable systems, stretching over 1.2 million kilometres and carrying the bulk of global Internet traffic (estimated at 97% in 2014). *See* TeleGeography, SUBMARINE CABLE MAP, https://perma.cc/4QWW-JFFG.

[106] Roxana Vatanparast, *The Infrastructures of the Global Data Economy: Undersea Cables and International Law*, 61 HARVARD INTERNATIONAL LAW JOURNAL FRONTIERS 1 (2020).

[107] Dwayne Winseck, *The Geopolitical Economy of the Global Internet Infrastructure*, 7 JOURNAL OF INFORMATION POLICY 228, 237 (2017) ("Communication paths … link many of the same 'world cities' now as they did then and some of the same old ornate cable telegraph buildings of the nineteenth century in London and New York have even been retrofitted for fiber optic cables today.").

[108] The UN Convention on the Law of the Sea (UNCLOS) recognizes sovereign control of a coastal state over a 12-nautical-mile belt of sea known (i.e., its territorial sea), which includes the airspace above and the seabed and subsoil below (UNCLOS, Article 2). An archipelagic state has sovereignty over the waters enclosed by its archipelagic baselines (archipelagic waters) with the express obligation to respect existing submarine cables laid by other states and to permit maintenance and replacement of such cables on receiving due notice. United Nations Convention on the Law of the Sea art.51(2) (Montego Bay, Dec. 10, 1982) 1833 U.N.T.S. 3, 21 I.L.M. 1261 (1982), *entered into force* 28 July 1996.

regulatory systems (via licenses and permits) apply. Private law technologies of contracts, corporate ownership, and project financing facilitate the industrial construction of submarine cables.

Uneven access to transnational connectivity and disparate control over submarine cables are largely functions of "historical legacies" and contemporary market dynamics but can also be attributed to the centrality of national sovereignty and territorial jurisdiction in the regulatory regime for submarine cables as that privileges coastal (and archipelagic) states and states that host land cable routes. States dependent on others for Internet connectivity are at a natural disadvantage. The relative power balance between states and cable operators mediates their respective legal rights and de facto control over physical Internet infrastructure.[109] Although government ownership and financing of fiber optic cables is rising overall, it often uses public-private partnerships, with investments made by telecommunications operators, non-telecommunication companies with high capacity demands for their private networks, and investment banks. Internet platform companies have joined traditional telecommunications companies to become cable co-owners in consortia, reflecting their growing power in Internet governance as they increase their control over the scarce resource of transnational bandwidth. Despite these trends, relative ownership and control over core elements of the global Internet infrastructure is shifting away from US firms towards new Internet infrastructure providers located in emerging markets, especially in Asia.[110]

If a physical link is established, names and numbers are assigned to the nodes of the network to make them uniquely identifiable, and protocols coordinate transport of data between them. This requires interoperable standards that are being developed by various standard-developing organizations (SDOs), illustrated in a simplified manner by the following figure,[111] which identifies key standards and SDOs, but does not account for the data infrastructures built on top of the Internet's application layer and also conceals who the dominant stakeholders in SDOs are.

---

[109] Dwayne Winseck, *The Geopolitical Economy of the Global Internet Infrastructure*, 7 JOURNAL OF INFORMATION POLICY 228, 236 (2017):

> The monopoly landing rights that [states] typically gave in the early years of development varied considerably, as did the terms of service they demanded with respect to privileges to be provided to local state officials and interconnection with local telegraphs, as well as their need to monitor (surveillance) and block (censorship) messages perceived as threats to public morality or national security. These landing licenses typically reflected the strength of the state that negotiated them. The stronger the state, the less likely it was to grant monopoly rights, as was the case in Britain and the United States, whereas the weaker the state, the longer the right to a monopoly, the more restrictive the terms of service obligations, and the less likely companies were to cooperate in ways other than those that advanced their business interests.

[110] Dwayne Winseck, *Internet Infrastructure and the Persistent Myth of U.S. Hegemony, in* INFORMATION, TECHNOLOGY AND CONTROL IN A CHANGING WORLD 93, 101–115 (B. Haggart et al. eds., 2019).

[111] Jorge L. Contreras, *Patents and Internet Standards*, GCIG Paper No. 29, GLOBAL COMMISSION ON INTERNET GOVERNANCE PAPER SERIES (Apr. 15, 2016), https://www.cigionline.org/publications/patents-and-internet-standards.

Layer	Standards	SDOs
4. Application	XML (data exchange)	W3C, OASIS
	HTTP, HTML (Web)	IETF, W3C
	IMAP, POP, MIME (email)	IETF
3. Transport	TCP, UDP	IETF
2. Internet	IPv4, IPv6, ICMP, ARP	IETF
1. Network	Ethernet, DSL, Wi-Fi, X.25	IEEE
	3G/4G	ETSI

Note: Acronyms used in this table: ARP — Address Resolution Protocol; DSL — digital subscriber line; HTTP — Hypertext Transfer Protocol; HTML — Hyper Text Markup Language; ICMP — Internet Control Message Protocol; IMAP — Internet Message Access Protocol; IPv4, IPv6 — IP version 4, IP version 6; MIME — Multi-Purpose Internet Mail Extensions; POP — Post Office Protocol; UDP — User Datagram Protocol; XML — Extensible Markup Language.

The Internet's open standards enable everyone to connect as long as participants' machines comply with these protocols.[112] In other words, standards "regulate" how data "flows" on the Internet.[113] This turns ostensibly technical standard-setting organizations such as the Internet Engineering Task Force (IETF) into global regulators of data flows, which raises inevitable questions about interests and politics.[114] These conflicts have materialized in attempts to integrate international human rights law as a substantive standard or at least a discursive toolkit to guide the decision-making of Internet standard-setting organizations.[115] Internet standard-setting organizations' commitment to global connectivity serves as a powerful meta-norm that perceives interests in tension with this goal as irritants.[116]

---

[112] *See* Andrew L. Russell, OPEN STANDARDS AND THE DIGITAL AGE: HISTORY, IDEOLOGY, AND NETWORKS (2014) (contrasting standards for previous telecommunication networks with the standards that enabled the Internet).

[113] The scare quotes around "regulates" are meant to indicate that standards are formally voluntary but often de facto unavoidable, whereas governmental regulation is formally binding but not always complied with. *See* on the complex relationship between formal law and technical standards Benedict Kingsbury, *Preface*, *in* CAMBRIDGE HANDBOOK OF TECHNICAL STANDARDIZATION LAW (VOL. 2): FURTHER INTERSECTIONS OF PUBLIC AND PRIVATE LAW xv (Jorge L. Contreras, ed. 2019). The scare quotes around "flows" echo our discussion of data metaphors above, *above* Section I.A, which emphasized that data does not move without agency but is being sent and received.

[114] *See, e.g.*, Corinne Cath and Luciano Floridi, *The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights*, 23 SCIENCE AND ENGINEERING ETHICS 449 (2017).

[115] *See* the work by the IETF's Human Rights Protocol Consideration Research Group which is tasked with researching whether Internet standards and protocols can enable, strengthen or threaten human rights generally, albeit with a focus on traditional civil and political rights. Human Rights Protocol Considerations, DATATRACKER, https://perma.cc/B7E2-RYHL. *See also* Monika Zalnierute & Stefania Milan, *Internet Architecture and Human Rights: Beyond the Human Rights Gap*, 11 POLICY & INTERNET 6 (2019); Monika Zalnieriute, *Human Rights Rhetoric in Global Internet Governance: New ICANN Bylaw on Human Rights*, 10 HARVARD BUSINESS LAW REVIEW 1 (2020).

[116] Niels ten Oever, WIRED NORMS: INSCRIPTION, RESISTANCE, AND SUBVERSION IN THE GOVERNANCE OF THE INTERNET INFRASTRUCTURE (2020).

Their commitment to connectivity explains why the Internet's core standard-setting organizations have been remarkably successful in preventing encroachment by intellectual property law. Technologies manufactured in accordance with the protocols and parameters specified by standards can, in principle, enjoy patent protection. Complex products may implement dozens or even hundreds of standards, each of which may in turn be covered by numerous "standards-essential patents" (SEPs).[117] Most patents are typically owned by firms themselves engaged in the standards-development process, thus making governance structures of the standards-setting organizations and the opportunity to participate in them an important site of infrastructural control.[118] In the context of Internet-related standards, however, the dominant standard-setting organizations have developed policies and norms requiring the licensing of relevant patents on a royalty-free basis, treating standards as a type of public good that should benefit everybody without restrictions, or at least at rates that are "fair, reasonable, and nondiscriminatory".[119] While participation in nominally "global" Internet governance institutions remains uneven and is dominated by actors — mainly company representatives and academics — who can afford attendance, they have succeeded in maintaining globally uniform standards that are open for anyone to adopt to enable transnational connectivity.

At the same time, for reasons of institutional design and ideology, Internet governance institutions have largely refrained from critically examining the distributive outcomes and power dynamics that their creation has enabled as it transformed into a foundational infrastructure for data creation, processing, and transfers. In an arrangement that resembles the tenuous balance between transnational economic integration and domestic societal safeguards that John Ruggie has theorized as "embedded liberalism",[120] nation states remain primarily responsible for the well-being of their citizens in the Internet era. At the same time, and in contrast to prior telecommunication technology (from the telegraph to the telephone), states do not enjoy a comparable level of control over the institutions that control Internet infrastructure. States may, however, resort to measures that limit the cross-border transfer of data, thereby challenging the Internet's foundational logic and most celebrated achievement for a variety of reasons, ranging from the pursuit of societal objectives such as data protection or economic welfare to national security concerns or political self-preservation qua censorship.[121]

---

[117] SEPs are patents that will always be infringed by a product conforming to a particular standard. The existence of patents covering standards has, in some cases, led to patent wars, royalty stacking (which makes it prohibitively expensive for competitors to develop standard-complying products), and patent hold-ups (instances where SEP holder demands excessive royalties after product manufacturers have made significant investments in standardized technology, thus resulting in lock-in effects).

[118] *See* Panos Delimatsis, Olia Kanevskaia Whitaker & Zuno Verghese, *Exit, Voice and Loyalty: Strategic Behavior in Standards Development Organizations*, TILEC Discussion Paper No. DP 2019-022 (Dec. 2, 2019), https://dx.doi.org/10.2139/ssrn.3487466.

[119] *See, e.g.*, Open Stand initiative, supported by IEEE, the Internet Society (ISOC), the Internet Engineering Task Force (IETF), the Internet Architecture Board (IAB), and W3C, and other standard-setting organizations, which provides that "Affirming standards organizations have defined procedures to develop specifications that can be implemented under fair terms. Given market diversity, *fair terms may vary from royalty-free to fair, reasonable, and non-discriminatory terms (FRAND).*" (emphasis added) Principles, OPEN STAND, https://perma.cc/6TYS-DEVE. For an in-depth discussion of patents, internet and standards, *see* Jorge L. Contreras, *A Tale of Two Networks: Patents, Standards and the Internet*, 93 DENVER LAW REVIEW 833-95 (2016).

[120] *See* Rawi Abdelal & John G. Ruggie, *The Principles of Embedded Liberalism: Social Legitimacy and Global Capitalism*, *in* NEW PERSPECTIVES ON REGULATION 151–162 (David Moss and John Cisternino, eds., 2009).

[121] The extensive literature on such restrictions often worries about "Internet fragmentation" or a "splinternet". *See, e.g.,* Milton Mueller, WILL THE INTERNET FRAGMENT? (2017); Mark Lemley, *The Splinternet*, DUKE LAW JOURNAL (forthcoming 2021), https://ssrn.com/abstract=3664027. *See also* Daniel Lambach, *The Territorialization of Cyberspace*, 22 INTERNATIONAL

One prominent example of such measures is EU's General Data Protection Regulation (GDPR) that limits the cross-border transfer of personal data by default.[122] The most coveted way to overcome this limitation is the "adequacy assessment" for which the European Commission determines whether another jurisdiction provides for an "essentially equivalent" level of protection.[123] The EU's Court of Justice later extended this requirement to the other legal technologies, especially standardized contractual clauses, that are available to "export" personal data from the EU to other jurisdictions.[124] The EU's data protection regime hence discriminates between jurisdictions with "essentially equivalent" data protection laws and those without and between entities that are able to provide prerequisite "additional safeguards" and those who are not able to do so.[125] The backlog that the European Commission has accumulated and discrepancies between countries that received an adequacy finding in the past and those that arguably provide a more robust level of data protection without being granted that status, raise questions about the EU's compliance with its non-discrimination commitments and applicable regulatory disciplines under the law of the World Trade Organization (WTO), in particular the General Agreement on Trade in Services (GATS).[126] The EU recognizes the fundamental tension between its restrictions on transfers of personal data and the commitment towards unimpeded "data flows" that animates many Internet governance institutions and the US "digital trade" agenda.[127]

India is another prominent jurisdiction that has experimented with a variety of data localization requirements.[128] Some of these requirements are sectoral and ostensibly motivated by safety and security concerns, for example the obligation imposed on payment system providers, issued by the

---

STUDIES REVIEW 482 (2020) (describing how different actors – not just states – territorialize and reterritorialize "cyberspace"); Niels ten Oever, *The Metagovernance of Internet Governance*, *in* CONTESTED POWER AND AUTHORITY IN INTERNET GOVERNANCE: RETURN OF THE STATE? ch. 3 (Blayne Haggart, Natasha Tusikov & Jan Aart Scholte eds., 2021) (differentiating between a private and multistakeholder internet governance regime and a multilateral internet governance regime, the latter of which seeks to accommodate national and regional norms and values).

[122] Regulation (EU) 2016/679 of Apr. 27, 2016, On the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J.(L.119/1) art. 44. This restriction on cross-border transfers of personal data goes back to the 1995 Data Protection Directive and even earlier national data protection laws. *See* Paul M. Schwartz, *European Data Protection Law and Restrictions on International Data Flows*, 80 IOWA LAW REVIEW 472 (1995); Christopher Kuner, TRANSBORDER DATA FLOWS AND DATA PRIVACY LAW (2013).

[123] This standard was clarified by the EU's Court of Justice (ECJ) in Case C-362/14 *Maximillian Schrems v. Data Protection Commissioner (Schrems)*, ECLI:EU:C:2015:650 (Oct. 6, 2015).

[124] E.C.J., Case C-311/18 *Data Protection Commissioner v. Facebook Ireland Ltd. & Maximillian Schrems (Schrems II),* ECLI:EU:C:2020:559 (July 16, 2020).

[125] Some entities might hence choose not to transfer personal data from the EU at all. *See* Anupam Chander, *Is Data Localization a Solution for* Schrems II*?*, 23 JOURNAL OF INTERNATIONAL ECONOMIC LAW 771 (2020).

[126] Svetlana Yakovleva & Kristina Irion, *Towards Compatibility of the EU Trade Policy with the General Data Protection Regulation*, 114 AJIL UNBOUND 10 (2020). *See* on potential justifications Neha Mishra, *Privacy, Cybersecurity, and GATS Article XIV: A New Frontier for Trade and Internet Regulation?*, 19 WORLD TRADE REVIEW 341 (2020).

[127] Svetlana Yakovleva & Kristina Irion, *Pitching Trade Against Privacy: Reconciling EU Governance of Personal Data Flows with External Trade*, 10 INTERNATIONAL DATA PRIVACY LAW 201 (2020).

[128] Aridrajit Basu, Elonnai Hickock & Aditya Singh Chawla, *The Localisation Gambit: Unpacking Policy Measures for Sovereign Control of Data in India*, CENTRE FOR INTERNET & SOCIETY, Mar. 19, 2010; Aridrajit Basu, *The Retreat of the Data Localization Brigade: India, Indonesia and Vietnam*, THE DIPLOMAT, Jan. 10, 2020.

Reserve Bank of India, to store data relating to payment systems "only in India".[129] India's protracted effort to reform its data protection law featured different forms of data localization in subsequent draft of its data protection bills, applicable to different categories of personal data.[130] In February 2019, India published its draft e-commerce policy, which adopts a nationalized version of the "data as a resource" framing and views data localization requirements with regard to certain categories of data as an important instrument to retain data-generated value within India.[131] In this way, India is embracing an openly data protectionist approach to digital development that challenges the "free flow of data" paradigm celebrated by Internet governance institutions and most trade economists.

To preserve "free data flows" against such governmental interference, the US developed a new model of rules for the digital economy during the negotiations for the Trans-Pacific Partnership (TPP).[132] Even though the US ultimately withdrew from TPP, these rules are now in effect through the Comprehensive and Progressive Agreement for Trans-Pacific Partnership (CPTPP).[133] Subsequent agreements, including the new NAFTA between the US, Mexico, and Canada (USMCA), the US-Japan Digital Trade Agreement (USJDTA), and the Digital Economy Partnership Agreement (DEPA) between Chile, Singapore, and New Zealand adopted essentially the same model.[134] However, it would be wrong to think that these commitments to free data flows materialize only between the countries that sign on to these agreements. Multi-national corporations can easily avail themselves of a corporate nationality that protects them from "unnecessary" data transfer limitations and data localization requirements. This kind of regulatory arbitrage can lead to the de-facto multilateralization of the model first instantiated in TPP. While beneficial for the global preservation of free data flows, it deprives countries from alternative digital development models under which they might favor their homegrown digital economy, as India tries to do, and complicates differentiated approaches under which only data flows to certain jurisdictions are allowed as under the EU's regime for outward transfers of personal data.[135]

Even in the absence of governmental limitations on data transfers, the Internet exhibits multiple stratifications, hierarchies, and ultimately inequalities that contradict the narrative that it is an inherently egalitarian infrastructure. International economic law is tilted against governmental distortions of what is perceived as the natural and market-efficient transnational flow of factors of production, including data. It is largely silent, however, about the hierarchies and power differentials that persists within the private sector. One such hierarchy concerns the relationship between different

---

[129] *See* Reserve Bank of India, Directive on Storage of Payment System Data, RBI/2017-18/153 (Issued on April 6, 2018), https://perma.cc/B85X-XAGB; *and* Payment and Settlement Systems Act, 2007, §18.

[130] Draft Personal Data Protection Bill, 2018, § 33, https://perma.cc/VB7T-EFHX; Draft Personal Data Protection Bill, 2019, Bill No. 373 of 2019 § 33, https://perma.cc/2LSU-MN72.

[131] Draft National e-Commerce Policy: India's Data for India's Development, https://dipp.gov.in/whats-new/draft-national-e-commerce-policy-stakeholder-comments.

[132] Thomas Streinz, *Digital Megaregulation Uncontested? TPP's Model for the Global Digital Economy*, *in* MEGAREGULATION CONTESTED: GLOBAL ECONOMIC ORDERING AFTER TPP ch 14. (Benedict Kingsbury et al. eds., 2019).

[133] CPTPP entered into force for Australia, Canada, Japan, Mexico, New Zealand, and Singapore in December 2018, and for Vietnam in January 2019. Brunei, Chile, Malaysia, and Peru have signed the agreement but did not ratify it. A consolidated version of CPTPP is available at https://www.iilj.org/megareg/materials/.

[134] Thomas Streinz, *Digital Megaregulation Continued: The Regulation of Cross-Border Data Flows in International Economic Law*, JAPAN SPOTLIGHT (June 2020).

[135] Thomas Streinz, *Data Governance in International Economic Law: Non-Territoriality of Data and Multi-Nationality of Corporations* (draft paper, on file with author).

types of ISPs and their relationship to Internet content providers. The biggest backbone networks (tier one) ISPs own operating infrastructure and interconnect with other networks of a similar size under only thinly (if at all) legalized "peering" arrangements.[136] Tier two ISPs, often regional, utilize a combination of paid transit under contractual terms and peering to deliver Internet traffic. Tier three ISPs are dependent on higher tier ISPs and purchase access rights from them. Internet Exchange Points (IXP) serve as the physical venue where ISPs can interconnect with one another, thereby constituting central entities in an otherwise distributed network.[137]

All these arrangements are largely governed by private law technologies (especially contracting), if not mere social practice, within public law frameworks of telecommunications law.[138] Net neutrality laws have become the most prominent intervention to guard against data flow discrimination on the Internet.[139] However, the scope of conventional net neutrality is limited as it only concerns the "last mile" relationship between ISPs and end users where ISPs are barred from deliberately favoring certain Internet traffic.[140] The resulting market dynamics are complicated and subject to much debate.[141] For the purposes of this paper, we focus only on the ways in which conventional net neutrality laws fail to account for disparate infrastructural control over data flows. Net neutrality laws solidify fundamental decisions enshrined in the Internet's foundational protocols. These protocols were designed to deliver packets across inter-connected networks regardless of addressee or content.[142] Net neutrality laws

---

[136] The basic idea is that computer networks mutually benefit from connecting with each other; *see* Leonard Kleinrock, *Creating a Mathematical Theory of Computer Networks*, 50 OPERATIONS RESEARCH 125 (2002). Hence, in principle, there is no need for extensive bargaining (and contracting), at least not between large networks. *See* further on the economics of peering (including between networks of different sizes) Pio Baaka & Thorsten Wichmann, *On the Economics of Internet Peering*, NETNOMICS 89 (1999); Jean-Jacques Lafont, Scott Marcus, Patrick Rey & Jean Tirole, *Internet Peering*, 91 AMERICAN ECONOMIC REVIEW 287 (2001).

[137] Nikolaos Chatzis, Georgios Smaragdakis & Anja Feldmann, *On the Importance of Internet eXchange Points for Today's Internet Ecosystem*, arXiv:1307.5264v2.

[138] *See* Uta Meier-Hahn, *Internet Interconnection: How the Economics of Convention can Inform the Discourse on Internet Governance*, GIGANET: GLOBAL INTERNET GOVERNANCE ACADEMIC NETWORK, ANNUAL SYMPOSIUM (2014), http://ssrn.com/abstract=2809867 (analyzing internet interconnection arrangements).

[139] Tim Wu is usually credited with inventing the term in *Network Neutrality, Broadband Discrimination*, 2 JOURNAL OF TELECOMMUNICATIONS & HIGH TECH LAW 141 (2003); *see also* Barbara van Schewick, INTERNET ARCHITECTURE AND INNOVATION (2010).

[140] In the US, the Federal Communications Commission (FCC) issued the Open Internet Order in 2015, which classified Internet Service Providers as "common carriers" under Title II of the Communications Act of 1934 and Section 706 of the Telecommunications Act of 1996, thereby imposing net neutrality obligations on ISPs. This decision was reversed in 2017. In September 2018, California enacted its own net neutrality legislation, the California Internet Consumer Protection and Net Neutrality Act of 2018, which clarifies unlawful conduct by ISPs, including the controversial practice of zero rating (not charging for a certain kind of Internet traffic). Net neutrality in the EU is established through Regulation (EU) 2015/2120 laying down measures concerning open internet access (2015) OJ L 310/1. The EU's Court of Justice clarified in Cases C-807/18 and C-39/19 *Telenor Magyarország Zrt.*
v Nemzeti Média- és Hírközlési Hatóság Elnöke ECLI:EU:C:2020:708 (Sept. 15, 2020) that the regulation covered zero rating practices.

[141] *See, e.g.*, Gerald R. Faulhaber, *Economics of Net Neutrality: A Review*, 3 COMMUNICATIONS & CONVERGENCE REVIEW 53 (2011); Shane Greenstein, Martin Peitz & Tommaso Valletti, *Net Neutrality: A Fast Lane to Understanding the Trade-Offs*, 30 JOURNAL OF ECONOMIC PERSPECTIVES 127 (2016).

[142] For the basic idea behind the end-to-end principle, *see* J.H. Saltzer, D. O. Reed, & D. D. Clark, *End-to-End Arguments in System Design* 2 ACM TRANSACTIONS ON COMPUTER SYSTEMS 4 (1984); *see* generally for the design implications David D. Clark, DESIGNING AN INTERNET (INFORMATION POLICY) (2018). *See* Nick Doty, ENACTING PRIVACY IN INTERNET STANDARDS, (2020), https://npdoty.name/writing/enacting-privacy/ (analyzing controversies in the W3C); Michael Rogers & Grace Eden, *The Snowden Disclosures, Technical Standards, and the Making of Surveillance Infrastructures*, 11

reinforce this technological decision by preventing ISPs from leveraging their central role for Internet access and traffic to charge their customers on either side (content providers and end users) more for delivering certain content faster. While this may seem egalitarian, it also lends itself to unequal data flow dynamics. If all Internet traffic is to be delivered equally to end users, those with much Internet traffic stand to gain more than those with little.

As major platforms transformed the Internet from a communication infrastructure into a data-gathering infrastructure, they found ways to leverage their infrastructural advantage over traditional ISPs (user access providers and data transit providers) to achieve preferential treatment for their data flows. The "last mile" treatment of Internet traffic by ISPs is not the only factor that determines how quickly and reliably data is being transmitted on the Internet. As platform companies' own content networks grew, ISPs entered into direct peering arrangements with them, which leads to superior network performance. In addition, platforms increasingly relied on content delivery networks (CDNs), especially to facilitate streaming services and cloud computing applications. CDNs replicate content on servers that are physically or virtually closer to end users, by maintaining a presence in, or close to, many large edge networks, thereby enhancing user experience. In sum, privileged access to fiber optic submarine cables (especially in regions suffering lack of bandwidth), peering arrangements with ISPs, and CDNs with data centers at the edge of the networks enabled large platforms to deliver their content faster and more reliably than their competitors.[143] These infrastructural dynamics are essentially not regulated by existing telecommunications law. Despite widespread adoption of net neutrality laws and protocols that, in principle, do not discriminate between Internet traffic, data does not "flow" equally on the Internet.

B. Data Ownership

While the inequality of data flows tends to be underappreciated, the unequal distribution of data and the associated possibilities for value generation are well established.[144] Some have suggested that property law may rectify this situation by recognizing data ownership rights and by facilitating more efficient data markets.[145] However, as we have argued above, concentrated control over data is often a function of concentrated control over data infrastructures. Most data accumulation occurs irrespective of legal property rights in data, and technological means can be deployed to prevent data access by others, thereby entrenching data inequality in a way that is not dependent on property law.[146]

INTERNATIONAL JOURNAL OF COMMUNICATION 802 (2017) (scrutinizing the role of intelligence agencies in technical standard-setting organizations); Niels ten Oever, *This is Not How we Imagined it": Technological Affordances, Economic Drivers, and the Internet Architecture Imaginary*, 23(2) NEW MEDIA & SOCIETY 344 (2021) (describing the prioritization of corporate interests over the interests of end users in Internet governance bodies).

[143] *See* Sravan Patchala, Seung Hyun Lee, Changhee Joo & D. Manjunath, *On the Economics of Network Interconnections and Net Neutrality*, 11TH INTERNATIONAL CONFERENCE ON COMMUNICATION SYSTEMS & NETWORKS (COMSNETS) (2019).

[144] *See, e.g.*, Dan Ciuriak, *Economic Rents and the Contours of Conflict in the Data-driven Economy*, CIGI PAPERS NO. 245 (July 2020); Eric A. Posner & E. Glen Weyl, RADICAL MARKETS ch 5 (2018).

[145] *See, e.g.,* Jeffrey Ritter & Anna Mayer, *Regulating Data as Property: A New Construct for Moving Forward,* 16 DUKE LAW & TECHNOLOGY REVIEW 220 (2018); Néstor Duch-Brown, Betin Martens & Frank Mueller-Langer, *The Economics of Ownership, Access and Trade in Digital Data*, JRC DIGITAL ECONOMY WORKING PAPER 2017-01 (Feb. 17, 2017), https://perma.cc/UU6G-RL2G.

[146] *See above* Section I.C.

This does not mean, however, that the law of data ownership is irrelevant.[147] Certain categories of data are protected by established intellectual property (IP) rights, namely copyright, trade secrecy, and *sui generis* database rights recognized in some jurisdictions. Assertions of property claims are often invoked, and become contentious, in response to demands for transparency, calls to share data with broader constituencies, and if data sharing is being required by law. Not establishing or recognizing legal ownership rights in data is insufficient to address data inequality because unequal control over data can be asserted infrastructurally; at the same time, proactively establishing or recognizing legal property rights in data can *further* entrench infrastructural control with the authority of law by preventing redistributive measures because data holders would use property rights as an additional shield to exclude others from access. International IP law and international investment law may lead to a further commodification of data qua international law even as domestic data ownership law remains contested and unsettled.

In light of considerable legal uncertainty around data ownership, some have suggested establishing new ownership rights over data for "data creators" to facilitate contracting over data and to incentivize data generation.[148] This idea, however, ignores the not-IP-like incentive structure under which most data gets generated and rewards those who have treated data essentially as a *res nullius*: "things that belong to no one but can be claimed by whoever catches them first".[149] The comparison with established IP rights over data is instructive to validate this critique as it clarifies the limited extent to which data is protected as property under existing IP law while also raising the question whether IP law indeed is the right legal framework for a discussion of data generation and its distributive effects.[150] The reason why IP law has framed this debate to date is likely due to path dependencies arising from data being intangible, as certain intangibles are subject to IP protection. The discourse is often plagued by conflating the normative case for recognizing property rights in personal data to address concerns around individual privacy and the excesses of "surveillance capitalism", with the broader questions about whether data, both personal and non-personal, already lends itself to property protections under existing law. For this reason, we first discuss the salience of "data ownership" under (domestic and international) IP law as well as contract and tort law before exploring interventions that adopt a property framing to counteract distributive data inequality and to rebalance control over different components of data infrastructures.

Copyright law, which is internationally harmonized qua the Berne Convention and the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), protects the original *expression* of an idea in creative works, including in digital form, but not ideas, procedures, methods of operation, or

---

[147] *See* Teresa Scassa, *Data Ownership*, CIGI PAPERS NO. 187, 2 (Sept. 2018) (table 1: contexts in which data ownership issues arise). *See also Study on Emerging Issues of Data Ownership, Interoperability, (Re-) usability and Access to Data, and Liability*, EUROPEAN COMMISSION (2018) (prepared for the European Commission's DG CONNECT).

[148] *See, e.g.*, Jeffrey Ritter & Anna Mayer, *Regulating Data as Property: A New Construct for Moving Forward*, 16 DUKE LAW & TECHNOLOGY REVIEW 220 (2018).

[149] Katharina Pistor, *Rule by Data: The End of Markets?*, 83 LAW AND CONTEMPORARY PROBLEMS 101, 107 (2020).

[150] Rochelle C. Dreyfuss, *The Challenges Facing IP Systems: Researching for the Future*, 4 KRITIKA: ESSAYS ON INTELLECTUAL PROPERTY 1, 3 (2020):

> Are the changes to the creative environment so extensive that the terms on which traditional IP law operates are no longer functioning effectively? Are the piecemeal legal responses seen to date a first-best solution or are there better ways for the law to support, manage, and structure innovation in this new Age? Are the right parties profiting? What are the distributive effects of these changes and have they been properly taken into account?

mathematical concepts as such.[151] Traditional copyright, initially developed for the creation of artistic and literary works by individuals (authors and other "creators"), is an imperfect fit for data generation. A photograph, for example, captures reality (in some way) but copyright does not attach to the facts contained within the picture but to the way in which these facts are being represented (expressed). Indeed, as the US Supreme Court has stated, "facts—scientific, historical, biographical, and news of the day, may not be copyrighted and are part of the public domain available to every person."[152] Whether or not data thus constitutes fact becomes a critical question.[153] The TRIPS agreement clarified that compilations of data, whether in machine readable or other form, which by reason of the selection or arrangement of their contents constitute intellectual creations shall be protected *as such*, but also stated that such protection shall not extend to *the data itself* and shall be without prejudice to any copyright subsisting in the data itself.[154] Hence, certain categories of data can be subject to copyright — if the general standard for creative works is satisfied — and compilations of data (databases) can be subject to copyright, too, if they constitute intellectual creations.[155] However, in

---

[151] Berne Convention for the Protection of Literary and Artistic Works (as amended on Sept. 28, 1979) art. 2, Nov. 18, 1984, 1161 U.N.T.S. 3; Agreement on Trade-Related Aspects of Intellectual Property Rights art. 9(2), Apr. 15, 1994, World Trade Organization, https://perma.cc/KJC3-STWN [hereinafter TRIPS].

[152] *Feist Publications, Inc. v. Rural Telephone Service Co.,* 499 U.S. 340, 348 (1991). *See also CH Canadian Ltd v. Law Society of Upper Canada*, [2004] 1 S.C.R. 339 (Can.) ("copyright protection only extends to the expression of ideas as opposed to the underlying ideas or facts").

[153] In *New York Mercantile Exchange Inc. v. Intercontinental Exchange Inc.*, the US Court of Appeal for the Second Circuit was asked to consider whether settlement prices generated by an algorithm created by the plaintiff were subject to copyright protection. The court approached the question by framing it in terms of "whether the plaintiff was the *author* of the settlement prices or merely their *discoverer*." (emphasis added). *N.Y. Mercantile Exch, Inc. v. Intercontinental Exch., Inc.*, 497 F.3d 109 (2d Cir. 2007). *See also RBC Nice Bearings, Inc. v. Peer Bearing Co.*, 676 F.Supp. 2d 9, 21 (D. Conn. 2009) (data derived from a series of calculations carried out by the plaintiffs was unprotectable facts). Similarly, in *BanxCorp v. Costco Wholesale Corp.*, the court focused on whether the formula used to convert raw data into the final value has "the degree of consensus and objectivity" that renders the final value "fundamentally a 'fact'." In that case, the court went on to say that: "If the data purports to represent *actual objective* prices of actual things in the world—the actual price of an actual settlement contract on a particular day—it is an unprotectable fact; if the data purports to represent an estimated price of a kind of idealized object—for instance, what a hypothetical, mint condition 2003 Ford Taurus with approximately 60,000 miles might be worth—then the hypothetical price may be eligible for some form of copyright protection in the right circumstances" (emphasis added) *BanxCorp v. Costco Wholesale Corp.*, F.Supp. 2d 280 (S.D.N.Y. 2013). These cases thus suggest that, in some circumstances, producers of indicators and other extrapolations may claim ownership, and enjoy copyright protection, in the indicators themselves. However, as Teresa Scassa points out, "to the extent that [such derived data] represent the idea behind the analytics that led to their creation [and thus] reflect a merger of idea and expression…, then it would seem that derived data must necessarily remain in the public domain, except where there is no merger between idea and expression. The challenge will be in determining when no merger occurs." *See* Teresa Scassa, *Data Ownership*, CIGI PAPERS NO. 187 (Sept. 2018).

[154] TRIPS art. 10(2) (emphasis added).

[155] The Copyright Act, 17 U.S.C. § 103(b) (U.S.), for example, defines a compilation as a "collection and assembling of preexisting materials or of data that are selected in such a way that the resulting work as a whole constitutes an original work of authorship." It also clarifies, in line with TRIPS, that the copyright in a compilation extends only to the compilation itself, and not to the underlying data. Thus, for example, a compiler of genetic sequence data can ensure that her database is copyrightable if she chooses an original set of genes or proteins for inclusion in the database or arranges the database in an original manner. *See* M. Scott McBride, *Bioinformatics and Intellectual Property Protection*, 17 BERKELEY TECH. L. J. 1331, 1349 (2002). The copyright protection afforded to a compilation of facts, however, only applies to the elements that are deemed sufficiently creative or original. Therefore, in the case of a genetic sequence database deemed sufficiently creative/original under the *Feist* standard (note 152) to merit copyright, the protection afforded by that copyright would extend only to the compiler's original selections or arrangement of data, not the data as such.

reality, much data and most databases do not fulfil these requirements, as much data generation consists in the recording of facts and most databases do not satisfy the threshold for creative works.[156]

The EU responded to this (perceived) problem by creating a *sui generis* right for databases through its 1996 Database Directive[157] and tried to entice other jurisdictions to reciprocate,[158] but most jurisdictions, including Canada and the US, have refrained from expanding copyright protection in this way.[159] In deviation from the creative works standard traditionally deployed in copyright law, the *sui generis* right applies when a database maker can show that "that there has been qualitatively and/or quantitatively a *substantial investment* in either the obtaining, verification or presentation of the contents to prevent extraction and/or re-utilization of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database."[160] The EU's own review of its Database Directive found no evidence that the creation of a *sui generis* right had any impact on the production of databases.[161] This striking finding indicates that IP incentives are neither necessary nor sufficient to explain database creation, or indeed data creation more broadly. The fact that the EU refrained from reforming the Database Directive by revoking *sui generis* data protection rights also illustrates how difficult it is to take away property protections once established, even if they do not achieve their intended objective, as beneficiaries can defend their established property rights under international investment and even human rights law.[162]

Protection of undisclosed information, as required by TRIPS and further ratcheted up in recent preferential trade agreements,[163] operates under a different IP logic than copyright and lends itself to potentially vast protection of data as "trade secrets". To acquire protection, the information in question must be secret in the sense that it is not generally known or accessible, must have commercial value due to being secret, and the entity lawfully in control of the information must have made reasonable steps to keep it secret. The value proposition under which "big data" operates, lends

---

156 As Teresa Scassa observes, "compilations of fact present many challenges when it comes to copyright. 'Whole universe' sets of fact may not reflect an original selection; similarly, where facts are arranged according to industry norms or standards, the compilation may lack originality. A data set that is constantly growing (for example, streamed sensor data) may similarly be incapable of being a compilation since there is never a completed work. Even if a selection or arrangement is original, the principle that facts are in the public domain means that only the original selection or arrangement of the compilation will be protected; anyone who extracts facts from the compilation using an independent selection and arrangement of those facts has not infringed copyright." *See* Teresa Scassa, *Data Ownership*, CIGI PAPERS NO. 187 (Sept. 2018).

157 Council Directive 96/9/EC of Mar. 11, 1996, On the Legal Protection of Databases, 1996 O.J. (L 77/20) [Database Directive].

158 Database Directive art. 11(3) foresaw international agreements between the EU and third countries, under which the EU's *sui generis* database protection would be extended to databases created in third countries, if those countries instituted *sui generis* database protection domestically.

159 "Sui generis protection" means the protection that is afforded to a particular object by virtue of that object's unique characteristics.

160 Database Directive art. 7 (emphasis added). The EU's Court of Justice clarified in Case C-46/02 *Fixtures Marketing* [2004] ECR I-10365 that the investment must pertain to the creation of the database, not the data contained therein.

161 European Commission Staff Working Document, Evaluation of Directive 96/9/EC on the Legal Protection of Databases of Apr. 25, 2018, SWD(2018) 146 final.

162 *See* Martin Husovec, *The Fundamental Right to Property and the Protection of Investment: How Difficult Is It to Repeal New Intellectual Property Rights*, *in* RESEARCH HANDBOOK ON INTELLECTUAL PROPERTY AND INVESTMENT LAW (Christophe Geiger (ed., 2019).

163 TRIPS art. 39.

credence to the claim that (potentially any) data has commercial value. Changes in the ways data is being generated, stored, and processed under recourse to cloud computing infrastructures have made it easier to limit access to data and to keep it secret. In other words, control over data infrastructures that generate and store data of commercial value in secure fashion is being rewarded by trade secrecy protection of such data, thereby solidifying data inequality and creating questionable incentives for competition and innovation.[164] Unlike copyright and patent law, which afford protection for only a limited period of time, trade secrecy affords potentially unlimited protection. One limiting principle is that trade secrecy protection ends once the information in question is no longer secret. Unlike patent law, which requires public disclosure and thereby facilitates the dissemination of knowledge (while retaining exclusive rights of commercial exploitation), trade secrecy law incentivizes data holders to keep information secret to enable exclusive commercial exploitation (potentially forever).[165] Data holders can try to retain secrecy qua contractual arrangements (e.g., nondisclosure agreements) but run the risk of losing trade secrecy protection if such safeguards fail. Trade secrecy is hence not an IP right that would lend itself to the kind of market-driven data transactions that proponents of new property rights for data creators have in mind. Its significance lies mainly in fending off attempts to get access to data, for example in attempts to scrutinize datasets which are suspected to contribute to bias in machine-learning.[166]

Property protections of data under IP law are not comprehensive, but data holders may still use contracts as legal technologies for regulating access to and control over data. In data contracts, data providers often assert "data ownership" even where arguably no recognized property right in data exists outside the contractual arrangement. The language in "data licensing agreements" is often borrowed from or at least influenced by IP law (especially copyright), even when "data ownership" claims are tenuous.[167] The terms of such license agreements vary but typically include: provisions delineating ownership and use rights, including use restrictions, purpose limitations, stipulations regarding the (in)ability of the licensee to aggregate or modify the data or create or use other derivative data, treatment of derived data, provisions regarding data delivery, security obligations, audit rights, risk allocations (via warranties and indemnities), dispute resolution provisions, and other provisions standard to contracts (such as termination, assignment, choice of law, etc.).[168] In the absence of explicit property protections, data providers run the risk that third parties that are not bound by the contract eventually acquire the data. This creates an incentive to resort to technological means to control access to data at a distance. In other words, where legal control is insufficient or contested, infrastructural control might still suffice. "Digital rights management" (DRM) was initially developed for copyright protection, achieved significant support from content industries and obtained legal protections against

---

[164] Jeanne Fromer, *Machines as the New Oompa-Loompas: Trade Secrecy, the Cloud, Machine Learning, and Automation* 94 NYU LAW REVIEW 706 (2019); David S. Almeling, *Seven Reasons Why Trade Secrets Are Increasingly Important,* 27 BERKELEY TECHNOLOGY LAW JOURNAL 1091, 1092-1095 (2012).

[165] *See* Mark Lemley, *The Surprising Virtues of Treating Trade Secrets as IP Rights*, 61 STANFORD LAW REVIEW 311, 352 (2008) (arguing that trade secrets should "expire" after a certain period).

[166] *See, e.g.,* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STANFORD LAW REVIEW 1343 (2018); *see generally* David S. Levine, *Secrecy and Unaccountability: Trade Secrets in our Public Infrastructure*, 59 FLORIDA LAW REVIEW 135 (2007).

[167] *See, e.g.*, Daniel J. Gervais, *The Protection of Databases*, 82 CHI.-KENT L. REV. 1109, 1148 (2007).

[168] *See, e.g.*, Daniel Glazer, Henry Lebowitz & Jason Greenberg, *Data as IP and Data License Agreements*, Practice Note, THOMPSON REUTERS (2019).

circumvention in many jurisdictions.[169] However, even without copyright and anti-circumvention protection of data by law, DRM technologies can still be deployed to control access to data.

Infrastructural analysis illuminates the ways in which legal technologies constitute, enable and otherwise intersect with control over data.[170] There is no need for full property protection of each component or layer of data infrastructures to sustain data inequality — as evidenced by the relatively widespread adoption of open source software, open standards, and open data — as long as control over key components of the infrastructure as a whole is not being challenged. Even where legal ownership over data is in question, control over data can be exercised through control over data infrastructures. Conversely, even when data rights and obligations are allocated in contracts, effective monitoring and enforcement is often equally dependent on infrastructural control (e.g., by allowing only access on certain machines).

Data infrastructures have both physical and digital components. The law of property, as it relates to real and personal (both tangible and intangible) property, protects physical components of data infrastructures and thus influences who has control over different components at different layers of data infrastructures. Laws regarding possession, use, and control of real and personal property allow for ownership of physical components of data infrastructures (e.g., cables, cell towers, data centers, computers, etc.) and protect owners' right to decide who has access to these objects (e.g., through the law of trespass). Data is always stored somewhere, though not necessarily in one place,[171] and the prerequisite hardware (hard drives) enjoy property protection (e.g., against theft or destruction) as does the real property on which data centers reside. In many jurisdictions, computer systems enjoy additional protections against unauthorized access, which can lead to claims functionally equivalent to the right to exclude others from accessing data on the basis of a property right. LinkedIn has advanced an interpretation of the Computer Fraud and Abuse Act (CFAA), which prohibits intentionally accessing a computer without authorization, according to which web scraping and use of publicly available information are without authorization under CFAA if in violation of the terms of the platform.[172] Tort law may provide a similar protection under the common law tort theory of "trespass to chattels". Courts in the US have held that online providers of information can protect their databases from unauthorized use and copying under this theory.[173] The protection that this theory

---

[169] *See, e.g.*, Digital Millennium Copyright Act, 17 U.S.C. § 1201 (circumvention of copyright protection systems). *See* for a sharp critique of DRM from a development perspective: Cory Doctorow et al., *Digital Rights Management (DRM): A Failure in the Developed World, A Danger to the Developing World*, *ITU-R Working Party 6M Report on Content Protection Technologies* (Mar. 11, 2005), http://hdl.handle.net/10760/6917.

[170] The distinction between data as an object or resource and data infrastructures that produce, shape, store, and transfer data is helpful in clarifying where ownership rights exist and where such legal protection is non-existent or at least contested.

[171] Paul M. Schwartz, *Legal Access to the Global Cloud*, 118 COLUMBIA LAW REVIEW 1681 (2018) (discussing the implications of "data sharding" for governmental access to data stored in transnational cloud infrastructures).

[172] *HiQ Labs, Inc. v. LinkedIn Corp*, 938 F.3d 985 (9th Cir. 2019). The case is pending as *LinkedIn Corp. v. hiQ Labs Inc.* before the US Supreme Court, *LinkedIn Corp. v. HiQ*, No. 19-1116 (U.S. Mar. 12, 2020), https://perma.cc/LK4B-BEM2. *See generally* on cyber-trespass laws as quasi-property regimes Thomas Kadri, *Platforms as Blackacres*, 68 UCLA Law Review (forthcoming 2021), https://papers.ssrn.com/abstract=3742086.

[173] *See* Charles C. Huse, *Database Protection in Theory and Practice: Three Recent Cases*, 20 BERKELEY TECH LAW JOURNAL 23, 29 (2005). According to this theory, "trespass to chattels" occurs when "an intentional interference with the possession of personal property has proximately caused injury." *See Thrifty-Tel v. Bezenek*, 54 Cal. Rptr. 2d 468, 473 (Ct. App. 1996).

appears to offer only relates to the data infrastructure used to store or publish the information, not to the data as such.[174]

Ownership over the digital components of data infrastructures varies. Those who exercise corporate control over data infrastructures do not necessarily own every component at every layer. Copyright law can be used to assert control over software, encompassing both source and object code, as well as structure, sequence, organization, and features generated by code.[175] Certain companies leverage their proprietary control over data management software and data formats to retain control over data. Farming data platforms in digital agriculture, introduced above,[176] are a case in point. Both Monsanto and John Deer use proprietary software which locks-in farmers and "their" data due to a lack of data portability and interoperability.[177] This has distributive effects as larger farms may be able to internalize the rising costs associated with dependencies on corporate data infrastructures better than their smaller competitors, thus further empowering "supersized farms" while marginalizing smaller farms, farmers, and their workers.[178] To the (limited) extent to which software enjoys patent protection in some jurisdictions, patent law may have comparable effects, albeit potentially counterbalanced by reasonable and non-discriminatory licensing terms. While software is generally not patentable in most jurisdictions, the increasing integration of software into machines (e.g. robots) complicates separation of software from hardware, and may lend patent protection to complex data infrastructures in which software and hardware are inseparably intertwined.

Proprietary control over software is no longer the norm. Open source software, where copyright law is leveraged to make software freely available, has become increasingly important, acquiring

---

[174] The party invoking the theory must be the owner of the "chattel" that has been interfered with, thus reverting to the question of what type of legal regime could confer rightful ownership interests over data as such.

[175] TRIPS art. 10(1) provides that computer programs, whether in source or object code, shall be protected as literary works under the Berne Convention of 1971. This is reflected in the copyright laws of jurisdictions around the world, e.g. Copyright Act, R.S.C. 1985, c. C-42, s. 5(1) (Can.); Art. 2021 L. 112 et seq C. propriété intellectuelle (Fr.); § 2(2) Gesetz über Urheberrecht und verwandte Schutzrechte **(**Urheberrechtsgesetz) 28.11.2018, BGBI. I S. 2014 (Ger.); Copyright, Designs and Patents Act 1988, c.48, s. 1(1) (U.K.); and Copyright Act, 17 U.S.C. § 102 (2012) (U.S.). In the US, if one wants to obtain copyright registration, source code for a specific version of the computer program must be deposited as part of the copyright registration process. Source code can be uploaded to the electronic registration system, as a PDF file, or as any other file type accepted by the U.S. Copyright Office. Alternatively, the source code can be printed out on paper and mailed it to the Office. *Cf.* Circular 61, Copyright Registration of Computer Programs.

[176] *See above* Section I.B.

[177] Sarah Rotz, Evan Gravely, Ian Mosby, Emily Duncan, Elizabeth Finnis, Mervyn Horgan, Joseph LeBlanc, Ralph Martin, Hannah Tait Neufeld, Andrew Nixon, Laxmi Pant, Vivian Shalla & Evan Fraser, *Automated Pastures and the Digital Divide: How Agricultural Technologies are Shaping Labour and Rural Communities*, 68 JOURNAL OF RURAL STUDIES VOLUME 112-122 (May 2019). As one of the farmers interviewed by Rotz and her colleagues observed:

> Everything is connected to the internet, I don't think you have any control over it anymore. That is a tricky one right, because like I said they [data management companies] have access to everything, yet we still get the bills all the time. So when do we get to issue a bill and get a little bit of a kick-back for the information that we are generating on a daily basis? Because, the supplier companies are like 'we need to find our R&D programs, to make it better for you guys'. But every time you make a new investment then the price of your equipment just went up because now it is the newest, latest, greatest, so you figure you [the company] can charge another 10% or another 5% or whatever amount it might be. So, you [the company] took all my information to do that.

[178] For broader effects of digital farming, *see* e.g., Evagelos D. Lioutas, Chrysanthi Charatsari, Giuseppe La Rocca & Marcello De Rosa, *Key Questions on the Use of Big Data in Farming: An Activity Theory Approach*, WAGENINGEN JOURNAL OF LIFE SCIENCES, 90-91 (2019).

infrastructural importance for a variety of use cases.[179] Examples range from encrypted data transfers (e.g., OpenSSL) over operating systems of web servers (e.g., Apache) to machine-learning algorithms (e.g., TensorFlow) and the internal infrastructures of large tech companies (Google runs a version of the open source operating system Linux on its desktop computers and servers.). Corporate control over data infrastructures is commensurate with certain components of those infrastructures being available to everyone else.

In the United States, there is ongoing litigation over whether copyright protection extends to certain APIs.[180] APIs are an important infrastructure for software as they facilitate interaction and interoperability between different components within and across programs and systems. APIs can also be used for data transfers online where web-based APIs have emerged as important vehicles for transnational access to data. Copyrightability of APIs is largely a question of rent-seeking and -distribution. Infrastructural control over APIs exists irrespective of copyright protection.[181] The designers and operators of webAPIs decide who can access the dataset they make available in this way and under what terms (e.g., download, just view, perform queries, etc.).[182] This is of particular concern when the resulting gatekeeping role can be abused to shut down access to data or to discriminate between different users. For related reasons, web APIs are sometimes seen as inferior vehicles for effectuating rights to data portability compared to user download and upload.[183] As webAPIs assume an increasing role as gateways between data infrastructures, those who control access to data qua webAPIs assume an infrastructural role regardless of copyright protection. For example, Ushahidi, a non-profit software company based in Nairobi, Kenya, provides its data collection, visualization, and interactive mapping service based on APIs provided by Google and Twitter. If these companies decided to terminate or alter the API, Ushahidi's ability to continue its service would be negatively impacted.

We now turn to potential remedies to challenge unequal control over data (de facto control over data) through legal instruments operating within a property law framework. One such idea seeks to leverage IP law to denounce property rights in data, an idea that "open data" advocates have promoted with remarkable success.[184] In distinction to open source software licensing, open data licensing cannot operate under the assumption that data is subject to copyright and has to account for the disparate *sui generis* IP protection of databases across jurisdictions. While modern "open data" licenses

---

[179] Guarini Global Law & Tech, Open Source Software as Digital Infrastructure: Legal Technologies and Institutional Design, www.guariniglobal.org/digital-infrastructure.

[180] Peter S. Menell, *Rise of the API Copyright Dead? An Updated Epitaph for Copyright Protection of Network and Functional Features of Computer Software*, HARVARD JOURNAL OF LAW & TECHNOLOGY 305 (2018); Jonathan Band, *The Global API Copyright Conflict*, 31 HARVARD JOURNAL OF LAW & TECHNOLOGY 615 (2018).

[181] Such control can be challenged by demanding interoperability, *see* Ian Brown, *Interoperability as a Tool for Competition Regulation*, OPEN FORUM ACADEMY (2020), https://osf.io/preprints/lawarxiv/fbvxd/.

[182] Public webAPIs tend to vest decision-making powers over the terms of data access solely with the data host. In private transactions, APIs are sometimes used alongside contracts to regulate access to data via legal and technical means.

[183] *See* Gabriel Nicholas & Michael Weinberg, *Data Portability and Platform Competition: Is User Data Exported from Facebook Actually Useful to Competitors?*, NYU SCHOOL OF LAW: ENGELBERG CENTER ON INNOVATION LAW & POLICY (2019), https://perma.cc/4X4E-QJPV.

[184] *See e.g.*, Beth Simone Noveck, *Rights-Based and Tech-Driven: Open Data, Freedom of Information, and the Future of Government Transparency*, 19 YALE HUMAN RIGHTS & DEVELOPMENT LAW JOURNAL (2017).

can account for these variations (at the expense of being more complicated and cumbersome),[185] the unwillingness of data hoarders to make data available as "open data" may be more difficult to tackle. In reality, with some exceptions,[186] "open data" strategies have largely been deployed to make data generated by the public sector data available as "open data". Recent instruments of international economic law actively encourage this approach.[187] Corporate data infrastructures often serve as gateways to find "open" data sets or include "open data" into their cloud offerings.[188]

Open data licenses make data freely available. They do not guard against de-facto appropriation and exploitation of "open data" by those with concentrated control over data infrastructures. This can lead to questionable distributive outcomes, especially when this model is only being deployed to make public sector data available to the private sector but not vice versa. Despite the plethora of reports, articles, and documents professing the value of "open data", there do not seem to be rigorous studies comparing the value derived from open data between public and commercial actors, although a few scholars have suggested that open data disproportionately empowers commercial entities already holding a socially and economically advantageous position.[189] An example from a report by NYU's GovLab on "open data" in Mexico illustrates the complexities of addressing data inequality via open data:

> … one of [Google's] most well-known products in Mexico is Google Maps, which offers a foundation created upon various geographic data and on which information layers can be constructed, including Points of Interest, Natural Resources, and Transportation Routes, among many others. Recently, the tool "Transit" has begun to offer transportation options using public transportation data from the Department of Mobility of Mexico City. Of great importance to the company has been the creation of a public Open Data Policy by the Government of Mexico, which removes barriers to accessing reliable data from officially authorized sources. The company explains that since the adoption of this policy, any user who seeks to use open data only has to accept the terms and conditions of use established in each case, reducing the need for additional agreements that impede, slow, or bureaucratize the access and use of public information, which is crucial for business' different initiatives. For example, two projects that have benefitted from the project are the GDELT Project, an open platform that monitors world news, and Google Public Data Explorer, a global directory of public data that also includes Mexican demographic data.[190]

---

[185] Examples include the Open Database License (ODbL) by Open Data Commons, which belongs to Open Knowledge International. The Creative Commons (CC) licensing system, initially developed for creative works, has been adjusted to accommodate use for data. *See generally* Alexandra Giannopoulou, *Understanding Open Data Regulation: An Analysis of the Licensing Landscape*, *in* OPEN DATA EXPOSED 101 (Bastiaan van Loenen, Glenn Vancauwenberghe & Joep Crompvoets eds., 2018).

[186] In 2016, Google released Open Images, a dataset of more than 9 million images labelled according to over 6000 categories. *See* Ivan Krasin & Tom Duerig, *Introducing the Open Images Dataset*, GOOGLE BLOG, Sep. 30, 2016, https://perma.cc/R9JL-C7QS.

[187] *See, e.g.,* USMCA, Article 19.18.

[188] *See, e.g.*, Dataset Search, GOOGLE, https://perma.cc/V5LR-LHXE; Registry of Open Data, AWS, https://perma.cc/39VQ-YBXU.

[189] Michael B. Gurstein, *Open Data: Empowering the empowered or effective data use for everyone?*, 16(2) FIRST MONDAY (2011), https://doi.org/10.5210/fm.v16i2.3316; Bianca Wiley, *Open Data Endgame: Countering the Digital Consensus,* CIGI PAPERS NO. 186 (Aug. 2018), https://www.cigionline.org/publications/open-data-endgame-countering-digital-consensus.

[190] New Study: The Open Data 100 Mexico, GovLab, Oct. 7, 2014, https://perma.cc/P73N-3TYT.

The report proceeds to note that "the company has allowed the use of open data in Mexico to positively impact the daily experience of its end users", enabling people to receive notifications about natural disasters, safe zones, and other relevant information which are being "pushed" by Google using data from National Meteorological Service and the National Water Commission, among other government datasets.[191] Although the report may be correct about user benefits, it fails to acknowledge how Google benefits from this arrangement: Google receives access to data at zero cost and can derive additional data from people's use of new functionalities, including new behavioral data that Google can monetize without any obligation to share the data or profits with the public.[192] This is not the only example in which public data generation has been repurposed to subsidize highly profitable and data-rich companies. Recent advances in machine translation can be attributed, in part, to human translation work that employees at International Organizations such as the UN or the EU have carried out for decades, and which companies in the machine learning business could exploit to train their machine-learning algorithms.[193]

Forcing private data holders to give up data by way of mandatory data sharing poses a direct legal challenge to de facto control over data. If faced with such requests, data holders are likely to resort to legal data ownership claims to counter data access rights of others. While access to data rights can lead to effective data redistribution, they also entrench the choices data generators made in the process of datafication. Whether entities make data available as "open data" or are forced to share data under mandatory data sharing laws, many choices regarding the terms of data access (e.g., file formats) and the possibilities of data use (e.g., with regard to categorization and structuring of data) have already been made by data producers who have the power and means to determine which data is collected in the first place and how. The decision of *what* data to produce and *which* data to grant access to often rests entirely with data-generating entities, absent specific regulatory interventions. This can lead to discrepancies in data utility that are further compounded if those who generate data also have access to other data and the means to aggregate and process data from different sources.

Another idea that seeks to deploy property law to challenge corporations' factual control over data proposes to award new property rights over personal data to individuals. This idea is not new,[194] but

---

[191] *Id.*

[192] Global data/tech corporations pay notoriously little tax and lobby heavily against digital services taxes. There is some evidence that intangible asset intensity increases tax avoidance possibilities for multinationals. *See* Roberto Crotti, *Does Intangible Asset Intensity Increase Profit-Shifting Opportunities of Multinationals?* (IHEID, Working Paper 02-2021), https://ideas.repec.org/p/gii/giihei/heidwp02-2021.html.

[193] *See* Ido Ramati & Amit Pichevski, *Uniform Multilingualism: A Media Genealogy of Google Translate*, 20 NEW MEDIA & SOCIETY (2018) (analyzing the underlying power structure of algorithmic and human collaboration in Google translate)

> When Google Translate was launched in 2006, it began utilizing texts like United Nations documents, international treaties, and multilingual corporate websites, all of which were accessible through its various services: in the words of Translate's first architect Franz Josef Och, its algorithms started mining 'everything that's out there'.

On broader concerns about machine translations see, e.g., Shlomit Yanisky-Ravid & Cynthia Martens, *From the Myth of Babel to Google Translate: Confronting Malicious Use of Artificial Intelligence—Copyright and Algorithmic Biases in Online Translation Systems*, 43 SEATTLE UNIVERSITY LAW REVIEW 99 (2019).

[194] *See e.g.*, Pamela Samuelson, *Privacy as Intellectual Property* (1999) 52 STAN. L. REV. 1125.

resurfaces regularly.[195] Individual data ownership rights, derived from property law, and data rights stemming from data protection laws ought to be distinguished. Certain data protection rights (such as the right to erasure) can be conceptualized as akin to property rights (which commonly also include the right to destroy one's own property).[196] However, such comparisons risk obscuring the fundamentally different logics animating the respective legal regimes.[197] At the same time, the idea to accord individual property rights over one's personal data shares some of the conceptual difficulties that also plague data protection law. One such problem concerns the distinction between personal and non-personal data: data inequality also accrues due to concentration of infrastructural control over non-personal data but approaches that depend on individuals exercising *their* rights over *their* data cannot account for this. A related problem is that data is inherently relative, which makes it difficult to delineate rights and obligations.[198] The case of genetic data made available to genetic data analytics companies such as 23andMe is illustrative of this phenomenon: does or should one "own" one's genetic data and have the right to make it available to others if this also reveals genetic information over one's relatives? Even if these conceptual difficulties could be overcome, an individualistic data property rights regime seems unlikely to be effective at challenging infrastructural control over data. One reason for this is the value proposition of "big data", which suggests that each individual's personal data are far less valuable individually compared to a dataset aggregating the data of many people, which complicates bargaining over adequate compensation for access to data. Moreover, the established power asymmetries and collective action problems that would pit individual interests, even if protected by property claims, against concentrated corporate interests and infrastructural control would be difficult if not impossible to overcome.[199]

If control over data is largely a function of control over data infrastructures and not entirely dependent on recognizing legal ownership rights in data, then interventions that would recognize such ownership rights are unlikely to be effective in remedying uneven control over data, and might produce adverse effects by rewarding those who already enjoy infrastructural control with additional legal safeguards against redistributive and other regulatory measures. Indeed, there is reason to think that a property framing is inapposite to questions of uneven control over data, precisely because it ignores the infrastructural dimensions of control over data and the related power to datafy. The most radical proposition along these lines would be to declare the world of data a *res communis*, a public good that cannot be owned by anyone, rather than a *res nullius* that is up for grabs. The economic argument for this approach is a continuation of information economics that has identified information asymmetries as harmful for economic growth and advocated for knowledge as a global public good, with only relatively thin IP protections. While this approach has its own complications, as it would need to be

---

[195] *See e.g.*, Eric A. Posner & E. Glen Weyl, RADICAL MARKETS: UPROOTING CAPITALISM AND DEMOCRACY FOR A JUST SOCIETY (2018); Katrina Miriam Wyman, *Property in Radical Markets*, 87 THE UNIVERSITY OF CHICAGO LAW REVIEW 126 (2019).

[196] *See e.g.*, J M Victor*, The EU General Data Protection Regulation: Toward a Property Regime for Protecting Data Privacy*, 123 YALE LJ 513 (2013).

[197] *But see* Diana Liebenau, *What Intellectual Property Can Learn from Informational Privacy, and Vice Versa*, 30 HARVARD JOURNAL OF LAW & TECHNOLOGY 285 (2016).

[198] Sebastian Benthall, *Situated Information Flow Theory*, PROCEEDINGS OF THE 6TH ANNUAL HOT TOPICS IN THE SCIENCE OF SECURITY (HOTSOS) (2019). *See also* Salome Viljoen, *Democratic Data: A Relational Theory for Data Governance*, YALE LAW JOURNAL (forthcoming 2021), https://ssrn.com/abstract=3727562 (Nov. 23, 2020).

[199] This problem also plagues the effectiveness of individual data rights that that data subjects commonly enjoy under data protection laws; *see below* Section II.C.

reconciled with established IP and data protection rights, it strikes us as the right baseline to develop appropriate data ownership and control regimes going forward.

In the meantime, it seems likely that international IP and investment law will be used to frame this debate in terms of commodification and asset protection. As Rochelle Dreyfuss and Susy Frankel have explored in detail,[200] the incorporation of the international IP regime into international trade law turned what once developed to coordinate incentives for individual IP creation into a system to protect the transnational interests of those who trade IP as a commodity (including within firms as part of sophisticated tax avoidance strategies). International investment law came to provide additional safeguards under which IP is being conceptualized as an asset to be protected against direct or indirect expropriation or regulation that can be challenged as a violation of the notoriously vague "fair and equitable treatment" standard. As international investment law also shapes private law,[201] it risks imposing a data-as-a-resource framing onto the evolving debate around data ownership, and it is likely to be mobilized against attempts to redistribute data qua mandatory data sharing.[202]

C. Data Rights

The rights-based approach to data regulation in form of data protection and privacy laws has dominated the discourse around legal regulation of the digital transformation of economies and societies around the globe. Data protection and data privacy law emerged in the 1970s in response to advances in computation technology.[203] The OECD's Privacy Guidelines of 1980 and the Council of Europe's Data Protection Convention of 1981 created two early models, respectively, for internationally harmonized privacy principles and data protection laws.[204] When the EU harmonized its Member States' data protection laws through its 1995 Data Protection Directive,[205] it created a new template for data protection law, which exercised a significant compliance pull even beyond the EU's borders ("Brussels Effect").[206] The EU's General Data Protection Regulation (GDPR) carried forward this legacy.[207] The GDPR is routinely touted as a template for jurisdictions around the world that do not yet have data protection laws or are planning to reform their existing laws. The Council of

---

[200] Rochelle Dreyfuss & Susy Frankel, *From Incentive to Commodity to Asset: How International Law Is Reconceptualizing Intellectual Property*, 36 MICHIGAN JOURNAL OF INTERNATIONAL LAW 557 (2015).

[201] Julian Arato, *The Private Law Critique of International Investment Law*, 113 AMERICAN JOURNAL OF INTERNATIONAL LAW 1 (2019).

[202] Thomas Streinz, *International Economic Law's Regulation of Data as a Resource for the Artificial Intelligence Economy*, *in* ARTIFICIAL INTELLIGENCE AND INTERNATIONAL ECONOMIC LAW: DISRUPTION, REGULATION, AND RECONFIGURATION ch. 9 (Shin-yi Peng, Ching-Fu Lin & Thomas Streinz eds., forthcoming 2021).

[203] *See* Przemyslaw Palka, *Data Management Law for the 2020s: The Lost Origins and the New Needs*, 68 BUFFALO LAW REVIEW 559, 572-589 (2020).

[204] OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, OECD https://perma.cc/9CRF-4NPW; Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data, Oct. 1, 1985, ETS No. 108.

[205] Directive 95/46/EC of Oct. 24, 1995, On the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, 1995 O.J. (L 281/3).

[206] Anu Bradford, THE BRUSSELS EFFECT: HOW THE EUROPEAN UNION RULES THE WORLD ch. 5 (2020).

[207] Paul Schwartz, *Global Data Privacy: The EU Way*, 94 NYU LAW REVIEW 771 (2019).

Europe's reformed Convention 108 is open for non-European countries to join.[208] By the end of 2019, 142 countries had data privacy laws on the books, 62 countries more than in the previous decade.[209] These are important developments as the world ostensibly gravitates towards comprehensive data protection regulation.

One should be cautious, however, not to overestimate the extent of convergence as differences persist even if data protection law on the books may look similar. The EU's data protection regime was reinforced when the EU's Charter of Fundamental Rights enshrined data protection and privacy as fundamental rights,[210] which the EU Court of Justice used for a right protective interpretation of GDPR.[211] Jurisdictions without such constitutional safeguards and activist courts will operate under different conditions and will likely generate different outcomes. Furthermore, as data protection and privacy laws proliferate both in places where such legal frameworks had been previously absent and in places that had determined their existing laws to be in need of updating, legislators and the public not only learn from each other but benefit from observing data controversies, debates, and litigation around the world and in different contexts. As a result, new features continue to appear.[212] The current extent of and the potential for future global harmonization of data protection and privacy laws ought not to be overstated.

In the US, which does not have a federal data protection law, California passed a consumer privacy act (CCPA) in 2018, augmented in 2020 by the California Privacy Rights Act.[213] There are important differences between European-style data protection laws and the conceptions of data privacy that dominate the discourse in the US and elsewhere.[214] We focus on data protection law in this paper as this concept does seem to have more purchase globally. Our main claims should apply *mutatis mutandis* to data privacy laws as well.

---

[208] Lee A. Bygrave, *The "Strasbourg Effect" in Data Protection: Its Logic, Mechanics and Prospects in Light of the "Brussels Effect",* 38 COMPUTER LAW & SECURITY REVIEW (2020); Graham Greenleaf, *How far can Convention 108+ 'Globalise'? Prospects for Asian Accessions*, 38 COMPUTER LAW & SECURITY REVIEW (2020).

[209] Graham Greenleaf & Bertil Cottier, *2020 Ends a Decade of 62 New Data Privacy Law*, 163 PRIVACY LAWS & BUSINESS INTERNATIONAL REPORT 24 (2020).

[210] Articles 7 and 8 of the Charter of Fundamental Rights of the European Union, (2012) OJ C 326/391. The EU is encouraging other jurisdictions to subscribe to this notion in its template for data governance provisions in trade agreements. *See* horizontal provisions for cross-border data flows and for personal data protection in EU trade and investment agreements, article B.1, https://trade.ec.europa.eu/doclib/html/156884.htm.

[211] Thomas Streinz, *The Evolution of European Data Law*, *in* THE EVOLUTION OF EU LAW ch. 29 (Paul Craig & Gráinne de Búrca eds., 3rd edn. 2021, forthcoming), https://ssrn.com/abstract=3762971.

[212] *See e.g.*, a provision to permit disclosure of individuals' personal information without their knowledge or consent where disclosure is for "socially beneficial purposes" in the newly proposed Canadian bill to enact the Consumer Privacy Protection Act and the Personal Information and Data Protection Tribunal Act, Bill C-11, tabled in the House of Commons, Dec. 2, 2020, https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c11.html.

[213] The California Consumer Privacy Act of 2018 (CCPA), CAL. CIV. CODE § 1798.100–199 took effect on January 1, 2020. *See* Anupam Chander, Margot E. Kaminski, and William McGeveran, *Catalyzing Privacy Law*, https://ssrn.com/abstract=3433922 (arguing that the CCPA represents a distinct approach to information privacy that differs in key aspects from GDPR). The California Privacy Rights Act of 2020 (CPRA) was passed via ballot initiative and will take effect on January 1, 2023. In addition to enshrining certain data protection rights in state law, it will create a California Privacy Protection Agency.

[214] Paul M. Schwartz & Karl-Nikolaus Pfeifer, *Transatlantic Data Privacy Law* 106 GEORGETOWN LAW JOURNAL 115 (2017); Paul M. Schwartz & Daniel J. Solove, *Reconciling Personal Information in the United States and the European Union* 102 CALIFORNIA LAW REVIEW 877 (2014).

Data protection and privacy laws do not appear to be effective in challenging unequal control over data. This is partly by design and partly due to persistent underenforcement, even within the EU, which is often perceived as the jurisdiction with the world's most stringent data protection law. While data protection laws can achieve some rebalancing between individuals and data controllers by granting rights to the former against the latter, and may even achieve systemic change by requiring "data protection by design",[215] data protection laws were not designed to address data inequality effectively. One design feature that further limits data protection law's ability to confront data inequalities is its limited scope of application. This may seem counterintuitive for those who hail the comprehensiveness of the EU's data protection regime. But even an umbrella framework such as the GDPR only applies to "personal data"; "non-personal data" is outside its scope of application. Naturally, this creates an incentive to avoid the strictures of data protection law by shifting focus towards non-personal data. The binary distinction between personal and non-personal data is tenuous due to data's relativity (data about myself might also be data about others) and increased technical ability to re-identify anonymized personal data and to infer personally identifiable information from large and variegated datasets even in instances when no "personal data" had been provided in the first place.[216] Even the GDPR's relatively broad conception of "personal data" recognizes categories of data that remain outside its scope of application (e.g., highly aggregated data and anonymized data). Stretching its scope of application even further risks turning data protection law into a data "law of everything".[217] From a traditional data protection law perspective, this is by design as non-personal data, on its face, does not appear to raise questions of informational self-determination in the same way as personally identifiable information does. The fallacy of this assumption, however, can be illustrated by the prospect of synthetic data: artificially generated data with characteristics suitable for machine learning purposes but without connection to any particular individual.

For some, synthetic data, and other "privacy preserving" technologies solve the "problem" of data protection law by allowing for datafication and data-driven decision-making without recourse to "personal data".[218] Note, however, that synthetic data aspires to reflect reality, thereby shaping and re-shaping the world as it is being represented, however imperfectly, through data. Concerns about uneven data-shaping power hence remain; the same is true for concerns about a possible concentration of synthetic data in the hands of few. Neither concern is being addressed by data protection law as it stands. What is true for synthetic data is true for other kinds of non-personal data being gathered by ever expanding data infrastructures, especially through smartphones and other IoT devices.[219] Data about the environment, for example, while highly salient for policy-making to address climate change, is not governed by data protection law at all, unless it is tied to an individual. The individualistic

---

[215] *See* Mireille Hildebrandt, SMART TECHNOLOGIES AND THE END(S) OF LAW 220-221 (2016).

[216] *See, e.g.*, Michèle Finck & Frank Pallas, *They Who Must Not Be Identified – Distinguishing Personal from Non-Personal Data under the GDPR*, 10 INTERNATIONAL DATA PRIVACY LAW 11 (2020); Inge Graef, Raphael Gellert & Martin Husovec, *Towards a Holistic Regulatory Approach for the European Data Economy: Why the Illusive Notion of Non-Personal Data is Counterproductive to Data Innovation*, 44 EUROPEAN LAW REVIEW 605 (2019).

[217] Nadezhda Purtova, *The law of everything. Broad concept of personal data and future of EU data protection law* 10 LAW, INNOVATION AND TECHNOLOGY 40 (2018).

[218] *See* Khaled El Emam, Lucy Mosquera & Richard Hoptroff, PRACTICAL SYNTHETIC DATA GENERATION: BALANCING PRIVACY AND THE BROAD AVAILABILITY OF DATA (2020). For a critical take from a privacy perspective *see* Theresa Stadler, Bristena Oprisanu & Carmela Troncoso, *Synthetic Data – A Privacy Mirage* (December 11, 2020), arXiv:2011.07018.

[219] *See* above Section I.C.

approach to data protection law has thus rightly been recognized as a problem in confronting data-related harms that accrue collectively.[220]

A related design feature that limits data protection law's ability to confront data inequality lies in the way in which an individual rights-based approach is being effectuated. Certain data protection rights enshrined in comprehensive data protection laws in the mold of the GDPR could, in theory, reduce data control asymmetries, at least with regard to personal data. For example, under GDPR, data subjects have a relatively broad, though not unconditional, right to request the erasure of personal data.[221] If significant numbers of data subjects exercised this right, they would wrest control over personal data from data collectors. But they do not. This is a general weakness of an individual rights-based approach to data regulation. It depends on individuals' willingness and ability to exercise their rights. If they do not, the law remains ineffective. In the literature, the phenomenon that individuals profess strong interest in data protection but do not seem to act accordingly as they routinely "give up" personal data with little regard to privacy has been described as the "privacy paradox". As Daniel Solove has shown, the privacy paradox is actually not a paradox at all.[222] Managing one's privacy is a time-consuming and potentially nerve-wrecking exercise. It is hence rational for individuals to proclaim an interest in data protection generally, especially with regard to societal risks stemming from systemic surveillance, while not contributing to this effort individually by challenging such practices themselves. This suggests the need for a shift towards more collective and systemic enforcement of data protection law.

The GDPR has made some steps in this direction by improving the institutional infrastructure on which effective data protection law depends. EU law pioneered the idea of embedding data protection officers (DPOs) within companies to affect the corporate culture towards data collection. [223] Independent data protection authorities (DPAs) are tasked with investigating data protection violations and sanctioning them, if need be. The GDPR's novel regime regulating algorithmic decision-making provides an example for collaborative governance, in which companies' data protection impact assessments may provide systemic governance and suitable safeguards of individual rights implicated by algorithmic decision-making.[224] How effective these institutional upgrades will turn out to be remains to be seen. The enforcement record of the GDPR so far does not inspire confidence. Underfunded data protection authorities struggled to fulfil their task to monitor resource-rich data controllers. [225] The continued existence of certain targeted advertising business models, long

[220] Martin Tisné, *The Data Delusion: Protection Individual Data Isn't Enough When the Harm is Collective*, https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/the_data_delusion_formatted-v3.pdf .

[221] GDPR, article 17.

[222] Daniel J. Solove, *The Myth of the Privacy Paradox* 89 GEORGE WASHINGTON LAW REVIEW 1 (2021), https://ssrn.com/abstract=3536265.

[223] Kenneth A. Bamberger & Deirdre K. Mulligan, PRIVACY ON THE GROUND: DRIVING CORPORATE BEHAVIOR IN THE UNITED STATES AND EUROPE (2015); *but see* Ari Ezra Waldman, *Designing Without Privacy*, 55 HOUSTON LAW REVIEW 659 (2018) (finding that privacy conceptions are narrow and limited and barely factor into the design of products).

[224] Margot Kaminski & Gianclaudio Malgieri, *Algorithmic Impact Assessments under the GDPR: Producing Multi-layered Explanations*, INTERNATIONAL DATA PRIVACY LAW (2020), https://doi.org/10.1093/idpl/ipaa020.

[225] Even the otherwise largely self-congratulatory evaluation of the GDPR by the European Commission admits that "[g]iven that the largest big tech multinationals are established in Ireland and Luxembourg, the data protection authorities of these countries act as lead authorities in many important cross-border cases and may need larger resources than their population would otherwise suggest. However, the situation is still uneven between Member States and not yet satisfactory

shown to be fundamentally incompatible with GDPR, indicate how difficult it is to upend established data infrastructures deeply engrained in the digital economy.[226]

Even if individuals were willing and able to exercise their data protection rights, these rights might not go far enough to challenge uneven control over personal data effectively. The GDPR's novel right to data portability,[227] now being replicated in similar laws around the world, is illustrative of a rights-based approach that turned out to be too limited in scope and ignored important infrastructural dimensions. The right to data portability is an amalgam of data protection, competition, and telecommunication law rationales. Traditional data protection law recognized the right to access and erase one's personal data. Data portability extends this idea to retrieve and/or transfer personal data from one data controller to the other in a way that can be understood as an expression of informational self-determination. Competition law knows certain doctrines (e.g., the essential facilities doctrine)[228] that grant mandatory access to certain categories of data when, without access to such data, market entry and effective competition become impossible. Data portability can be understood in similar terms to confront the pervasive network effects that have led to extreme platform concentration.[229] These data portability rationales intersect when reduction of switching costs leads to increased competition, which is not only being carried out on price terms but also leads to a levelling up of data protection standards.

Data portability could in theory lead to a redistribution of personal data and potential de-concentration of infrastructural over such data. Unfortunately, there is little evidence that this is actually happening. Two key reasons can be identified for data portability's limited impact. One concerns the scope of the right: the right to data portability under GDPR is explicitly restricted to personal data that the data subject *provided* to the controller. Personal data that the data controller inferred about the data subject is not covered.[230] Moreover, as stated before, the value of personal data is highly contextual and relative. If only the comment that a user provided on a social media platform has to be transferred to another platform but not the picture provided by another user to which the comment was attached, the value of data portability remains limited.[231] The other reason for data portability's limited impact (at least so far) is infrastructural: while the right to data portability under GDPR requires that personal data is being made available in a structured, commonly used, and machine-readable format, clearly recognizing that unstructured data in hard-to-access proprietary formats can constitute

---

overall." *See* European Commission, Data protection as a pillar of citizens' empowerment and the EU's approach to the digital transition – two years of application of the General Data Protection Regulation, COM(2020) 264 final.

[226] *See, e.g.*, Johnny Ryan, *Two years on from complaint to the Irish Data Protection Commission, the RTB data breach is the largest ever recorded, and appears to have worsened*, Submission to the Irish Data Protection Commission (Sep. 21, 2020).

[227] GDPR, Article 20.

[228] *See further below* Section II.D.

[229] The idea to prevent lock-in effects is also behind the idea, reflected in domestic and international telecommunications law, that one has a right to number portability to enable switching between different telecommunication services providers.

[230] The CCPA's right to access personal information in a portable and readily usable format can be construed as a data portability right, which – unlike GDPR – also covers inferred data. *See* for a further comparison between GDPR and CCPA Anupam Chander, Margot E. Kaminski & William McGeveran, *Catalyzing Privacy Law*, http://ssrn.com/abstract=3433922.

[231] *See* Gabriel Nicholas & Michael Weinberg, *Data Portability and Platform Competition: Is User Data Exported from Facebook Actually Useful to Competitors?*, NYU SCHOOL OF LAW: ENGELBERG CENTER ON INNOVATION LAW & POLICY (2019), https://perma.cc/4X4E-QJPV.

insurmountable obstacles to data access and use,[232] GDPR fails to specify how the personal data actually ought to be transferred. This is a missed opportunity to mandate and regulate private data transfer infrastructures (e.g., web APIs) with public oversight or to create alternative public data transfer infrastructures to ensure regulatability and generate interoperability.[233] In the absence of such infrastructural interventions, Apple, Facebook, Google, Microsoft, and Twitter developed an open source data transfer infrastructure — the Data Transfer Project — which they control, thereby indirectly affecting the effectiveness of the right to data portability.[234]

So far, we have discussed why data protection law has not been effective at challenging data inequality. In some instances, data protection law can even exacerbate data inequality. One way data protection law can come into conflict with attempts to redistribute control over personal data is when data controllers invoke data protection obligations to refuse data-sharing. The right to data portability, for example, is explicitly conditioned on not adversely affecting the data protection rights and freedoms of others,[235] a fact that is routinely stressed by platform companies in their discussion and practice of data portability.[236] Even though contemporary data protection laws recognize exemptions to facilitate data sharing for public benefit,[237] platform companies often adopt overly restrictive interpretations of these exemptions and refuse to provide meaningful access to data for researchers studying the impact of platforms on peoples' lives and livelihoods.[238] Another way data protection law may exacerbate data inequality concerns the increased compliance costs (assuming at least good faith efforts at compliance) that may hand an inadvertent advantage to powerful and resource-rich accumulators of personal data. This question is conventionally only discussed under a competition and innovation policy framing, and views about a potential disparate impact of data protection law differ widely.[239] We suggest that it is a question worth asking not just from the perspective of competition and innovation policy but also from a perspective of data inequality.[240]

By highlighting the inadequacy of data protection and privacy laws in resolving data inequality, we nonetheless remain sympathetic to its underlying objective of reclaiming the centrality of individual choice and autonomy against over-intrusive power of corporate data collectors and processors. Still, we think it is worthwhile contemplating whether the overemphasis on individual rights, even if

---

[232] *See above* Section I.B.
[233] *See below* Sections III.B. and III.D.
[234] Data Transfer Project, https://datatransferproject.dev/ and https://github.com/google/data-transfer-project.
[235] GDPR, article 20(4).
[236] *See, e.g.*, the paper by Facebook's Chief Privacy Officer Erin Egan, *Data Portability and Privacy: Charting a Way Forward* (September 2019), https://about.fb.com/wp-content/uploads/2020/02/data-portability-privacy-white-paper.pdf.
[237] GDPR, articles 85 and 89.
[238] Jef Ausloos, Paddy Leerssen & Pim ten Thije, *Operationalizing Research Access in Platform Governance: What to learn from other industries?*, ALGORITHMWATCH (June 25, 2020), https://algorithmwatch.org/en/governing-platforms-ivir-study-june-2020/.
[239] *See, e.g.*, Michal Gal & Oshrit Aviv, 16 *The Competitive Effects of the GDPR*, JOURNAL OF COMPETITION LAW AND ECONOMICS 349 (2020) (arguing that GDPR limits competition in data markets, creating more concentrated market structures and entrenching the market power of those who are already strong; and that GDPR limits data sharing between different data collectors, thereby preventing the realization of data synergies which may lead to better data-based knowledge); Yafit Lev-Aretz & Katherine J. Strandburg, *Privacy Regulation and Innovation Policy* (2020) 22 YALE JOURNAL OF LAW AND TECHNOLOGY 256 (2020) (arguing that carefully designed privacy regulation can provide societally beneficial incentive structures for innovation).
[240] *See below* II.D. for discussion of what a competition law framing might overlook.

beneficial in the short term, may in the long term legitimize the default position that datafication is both an acceptable and a desirable commercial activity so long as certain concessions, in the form of enumerated rights, are made to individuals whose lives and environments are being datafied and affected by datafication. This concern echoes similar sentiments levelled by scholars like Samuel Moyn against human rights law (as ultimately not being effective at challenging rising economic inequality)[241] and Jessica Whyte, who has traced the co-constitution of human rights discourse (emphasizing individual freedoms against governmental intrusion) with the rise of the neoliberal project.[242] We are raising these parallels to caution against a perception of extant data protection as an effective check on data inequality.[243] Data protection law operates under the assumption that if data controllers have legally acquired personal data, they may control that data as long as legitimate grounds for data processing exist. Certain limiting principles contained in data protection law — such as purpose limitation and data minimization — may have a dampening effect on data accumulation and repurposing and are at odds with the value proposition of "big data".[244] Safeguards against data collection (e.g. strict consent requirements for sensitive data) can amount to difficult-to-overcome obstacles to datafication. But ultimately, data protection law does not challenge concentrated control over data infrastructures nor does it meaningfully constrain the power to datafy.[245] The COVID-19 pandemic revealed global corporations' control over data infrastructures, when Apple and Google collaborated to make available within their mobile operating systems bluetooth-powered COVID-19 contract tracing for public health authority-sanctioned apps, but refused to adapt their system to enable centralized collection of data favored by French and British health authorities. Their refusal was applauded by privacy advocates but also made visible that data protection law does not confront centralized corporate control over large scale data-generating infrastructures, in this case the operating systems running on billions of mobile devices.[246]

Competition law is increasingly seen as a supplement or corollary to data protection law to address "platform power", including their control over data infrastructures. In the next sub-section, we explore the potential and limits of this approach.

---

[241] Samuel Moyn, NOT ENOUGH: HUMAN RIGHTS IN AN UNEQUAL WORLD (2018); *but see* the sharp critique by Gráinne de Búrca, 16 INTERNATIONAL JOURNAL OF CONSTITUTIONAL LAW 1347 (2018).

[242] Jessica Whyte, THE MORALS OF THE MARKET: HUMAN RIGHTS AND THE RISE OF NEOLIBERALISM (2020).

[243] *See also* Angela Daly, *Neo-Liberal Business-As-Usual or Post-Surveillance Capitalism With European Characteristics? The EU's General Data Protection Regulation in a Multi-Polar Internet* (July 19, 2020), https://ssrn.com/abstract=3655773.

[244] *See, e.g.*, Tal Zarsky, *Incompatible: The GDPR in the Age of Big Data*, 47 Seton Hall Law Review 995 (2017).

[245] There is also a risk that, through practices of implementation, data protection law is being transformed into a mere compliance exercise ("check list") or a problem to be solved rather than a philosophy to be embraced. This, in turn, may contribute to the notion that mere compliance with data protection law is sufficient from a public policy perspective. *See* for the opposite approach Mireille Hildebrandt, *Primitives of Legal Protection in the Era of Data-Driven Platforms, 2* GEORGETOWN LAW AND TECHNOLOGY REVIEW 252 (2018). *See generally* Mireille Hildebrandt, SMART TECHNOLOGIES AND THE END(S) OF LAW: NOVEL ENTANGLEMENTS OF LAW AND TECHNOLOGY (2015).

[246] Michael Veale, *Sovereignty, privacy, and contact tracing protocols*, *in* DATA JUSTICE AND COVID-19: GLOBAL PERSPECTIVES 34 (Linnet Taylor et al (eds), 2020).

D. Regulating Platform Power

As we have seen, certain platform companies have created expansive data infrastructures.[247] The rise of platform power has led to increased regulatory and scholarly scrutiny. Such scrutiny has proceeded broadly along three tracks: by focusing on laws protecting platforms from liability for hosted content (within the relatively recent field of intermediary liability law),[248] regulations specifically aimed at regulating relationships between platforms, businesses, and consumers (either through dedicated platform regulation or consumer protection law), and through tools of antitrust and competition law. While the discussion about platform liability laws is often framed through the lens of communicative freedoms and related harms, ranging from copyright violations to hate speech, antitrust and competition laws have focused on platform companies as dominant market actors and their impact on dependent commercial actors, potential competitors, and consumers. Platform regulation consists of a complex and disparate set of laws deeply intertwined with the rise of platform companies and informational capitalism.[249] In line with the more limited ambition of this paper, we focus on the ways platform regulation has enabled data inequality, and we explore ways such regulation could mitigate or reduce data inequality going forward.

We first turn to established intermediary liability laws and newly emerging platform regulation. We then address the evolving debate in antitrust and competition law around growing platform power, highlighting their respective salience for questions of data inequality. Each of these legal frames asks its own unique sets of questions that all touch on data inequality but rarely focus on it.

Liability shields for user-generated content were erstwhile seen as critical for Internet freedom.[250] The US pioneered this regulatory approach for user-generated content generally (stating that providers of platforms are not to be treated as publishers or speakers of information provided by others) and for copyright protected content specifically (instituting a notice and takedown regime).[251] These legal mechanisms have been replicated, albeit in more limited form, in jurisdictions around the world.[252] They also feature in recent US-designed "digital trade" agreements.[253] Yes, criticism is mounting that the liability shield for user-generated content, and its expansive interpretation by US courts, amounts to a subsidy for platform power by effectively disabling liability as a legal technology inducing appropriate corporate behavior. The main focus of this debate, at least in the US, is on content-related harms and platforms' responsibility for enabling such harms.[254] One tangent within the debate

---

[247] *See above* Section I.C.

[248] *See* the contributions in Giancarlo Frosio (ed.), OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY (2020).

[249] Julie E. Cohen, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM (2019).

[250] *See* Electronic Frontiers Foundation, *CDA 230: The Most Important Law Protecting Internet* Speech, https://perma.cc/ZQ7L-538R; Jeff Koseff, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019).

[251] Section 230 of the Communications Act of 1934 as amended by the Telecommunications Act of 1996, 47 U.S.C. § 230; Digital Millennium Copyright Act (DMCA), 17 U.S.C. § 512.

[252] *See, e.g.*, Directive 2000/31/EC on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market (e-commerce directive) (2000) OJ L 178/1, articles 12–15. A global overview over intermediary liability laws is available at https://wilmap.law.stanford.edu/.

[253] *See, e.g.*, USMCA, Article 19.17.

[254] *See, e.g.*, Danielle Keats Citron and Mary Anne Franks, *The Internet As a Speech machine and Other Myths Confounding Section 230 Reform* (2020) BOSTON UNIVERSITY SCHOOL OF LAW PUBLIC LAW RESEARCH PAPER NO. 20-8, https://ssrn.com/abstract=3532691; Olivier Sylvain, *Discriminatory Designs on User Data* (April 1, 2018) KNIGHT FIRST AMENDMENT INSTITUTE EMERGING THREATS ESSAY SERIES, https://perma.cc/3GVP-22GE.

concerns the question of whether small or big platforms benefit more, and in what ways, from intermediary liability protection.[255] From the perspective of data inequality that this paper is concerned with, intermediary liability seems only indirectly related to uneven control over data and unequal power to datafy.[256]

However, platform companies have also emerged as targets for dedicated platform regulation that transcends the debate around intermediary liability. One such example is the EU's regulation on platform-to-business relations (p2b).[257] This type of platform regulation singles out a particular type of intermediary, namely platforms that allow business users to offer goods or services to consumers.[258] The regulation supplements the contractual relationships between platforms and business users. Its main intervention is to demand a certain degree of transparency to guard against information asymmetries between platforms and business users.[259] The regulation shies away, however, from demanding the sharing of the data that e-commerce platforms generate about business and consumer behavior. Instead, it requires platforms to disclose in their terms of conditions which business or consumer data is being generated and who has access to it (or not).[260] The idea seems to be that such disclosure might enable businesses to negotiate more favorable terms for access to data from platforms. Although moves to increase transparency should be viewed favorably and, as we argue in Part III below, should be pursued more aggressively, it remains to be seen whether such an intervention in itself is sufficient to effectively redistribute control over data through contractual means, given power asymmetries between platforms and businesses and the increasing business dependency on ecommerce platforms, exacerbated by the COVID-19 pandemic.[261]

Dedicated platform regulation to address information asymmetries is a relatively recent phenomenon. The EU's ambitious digital single market strategy includes proposals not only to revamp the liability regime established under the e-ecommerce directive, but also to provide a new regulatory framework for digital services through the proposed Digital Services Act (DSA) and the Digital Markets Act (DMA).[262] Both instruments are meant to complement each other but adopt different concepts and logics to achieve their respective regulatory objectives. The DSA is designed around different categories of "intermediary services" and regulates providers' respective liability and due diligence obligations as well as their enforcement.[263] The DMA singles out certain "core platform services" as

---

[255] *See, e.g.*, Eric Goldman, *Want to Kill Facebook and Google? Preserving Section 230 Is Your Best Hope*, Balkanization, New Controversies in Intermediary Liability Law (2019), https://ssrn.com/abstract=3398631.

[256] How platform companies would have developed in the absence of protections against intermediary liability over the last quarter century is a difficult to answer hypothetically. Conversely, it is not clear that platform companies' dominance would be curbed effectively if their liability would increase significantly today. The main impact would likely be on the creation and dissemination of user-generated content, making platforms more wary of content that might expose them to liability.

[257] Regulation (EU) 2019/1150 on promoting fairness and transparency for business users of online intermediation services (2019) OJ L 186/57 (p2b regulation).

[258] *See* the composite definition of "online intermediation services" in p2b regulation, article 2(2).

[259] For further ideas and even more assertive ways of demanding transparency from data infrastructure controllers *see below* Section III.C.

[260] *See* p2b regulation, article 9 (misleadingly labelled "access to data").

[261] For ideas how to counter asymmetric control over data through pooling strategies *see below* Section III.D.

[262] These are together referred to as the Digital Services Package: https://perma.cc/SM8H-U48A.

[263] Proposal for a regulation on a single market for digital services (Digital Services Act) and amending directive 2000/31/EC, COM(2020) 825 final.

gatekeepers on which additional obligations are being imposed, including prohibitions to engage in certain business practices.[264] As under the p2b regulation, a certain emphasis is on increased transparency obligations, with a focus on illegal content and platforms' content moderation practices (in the case of the DSA), the price building mechanisms for online advertising, and platforms' ranking decisions to guard against self-preferencing (under the DMA).[265] But the DMA goes further by subjecting "gatekeepers" to various obligations designed to facilitate or mandate data sharing with businesses and end users.[266] Institutions tasked with enforcing the DSA can request access to any data necessary to monitor and assess compliance with the DSA from "very large" online platforms.[267] The European Commission may also request access to databases when conducting market investigations or enforcing the DMA.[268] This suggests a shift towards more centralized and more robust and intrusive enforcement compared to the GDPR. In what form (if at all) these proposals will be adopted by the EU's legislative bodies remains to be seen. What is relevant for purposes of this paper is that the DSA and DMA can be seen as further evidence towards dedicated yet differentiated platform regulation that complements and goes beyond the established intermediary liability regimes.

In recognizing the central (infrastructural) role of "very large" online platforms (under the DSA) and acknowledging the widespread tracking and profiling activities of many large platforms (under the DMA), the European Commission focuses its regulatory attention directly on the role of platform companies in Europe's digital economy (and beyond). At the same time, the Commission's proposals do not take issue with the amassing of data[269] as such and do not question the desirability of ever-increasing datafication.[270] Whether the increased regulatory burdens under a potential DSA or DMA would have a dampening effect on data generation by their main regulatory targets is difficult to predict.

The proposed DMA prohibits "gatekeepers" explicitly from combining personal data sourced from core platform services with other services offered by the gatekeeper, but the prohibition can be overcome if end users are being provided with a specific choice and provide consent under GDPR.[271] This "solution", which seems likely to reenforce the importance of "consent", resembles the remedy

---

[264] Proposal for a Regulation on contestable and fair markets in the digital sector (Digital Markets Act), COM(2020) 842 final.

[265] *See, e.g.*, DSA proposal, article 13; DMA Proposal. Note also the European Commission's guidelines on ranking transparency pursuant to the p2b regulation, (2020)OJ C 424/1.

[266] *See, e.g.*, DMA proposal, article 6.1(g) giving advertisers and publishers a right to request, free of charge, access to performance measuring tools and ad inventory verification information); article 6(h) reenforces the GDPR's right to data portability, discussed above II.C., by requiring "tools for end users to facilitate the exercise of data portability, in line with [GDPR], including by the provision of continuous and real-time access."

[267] DSA proposal, article 31. "Very large" online platforms are providing services to 45 million (or more) average monthly active users within the EU; *see* DSA proposal, article 25, for details.

[268] DMA proposal, article 19.

[269] The DMA states explicitly on page 1, footnote 1 that comprehensive tracking and profiling of end users online as such is not necessarily an issue if done in a controlled and transparent manner, in respect of privacy, data protection and consumer protection.

[270] Indeed, the European Commission's European strategy for data, COM(2020) 66 final celebrates growing data volumes around the world. On the questionable relevance of mere quantitative metrics about data, *see* Bruno J. Strasser and Paul N. Edwards, *Big Data Is the Answer . . . But What Is the Question?* (2017) 32 OSIRIS 328.

[271] DMA proposal, article 5(a). Gatekeepers are being defined as providers of core platform services that have a significant impact on the internal market, serve as important gateways for business users to reach end users, and enjoy an entrenched and foreseeably durable position. *See* DMA proposal, article 3 for further details.

that the German antitrust authorities imposed against Facebook Inc. by requiring it to no longer aggregate personal data collected on its platforms Facebook, Instagram, and WhatsApp.[272] The difference is that the proposed DMA would impose such a non-aggregation obligation outright against "gatekeepers", while the German Bundeskartellamt had to justify its intervention within the established categories of European competition law, alleging that Facebook had abused its dominant market position in a way that rendered users' formal consent (by agreeing to the platforms' respective terms of service) moot. The regulatory prohibitions under the DMA proposal are thus illustrative of the differences between regulatory intervention and competition law enforcement and the limited ambition of the latter concerning data inequality.

Competition authorities in the EU have been conducting extensive investigations into the conduct of platform companies, triggering a debate about the proper scope and purpose of competition law.[273] Antitrust authorities elsewhere, including in the US, have ramped up their investigations into platform companies' conduct.[274] Questions of access to and control over data have featured prominently in these investigations and debates. Indeed, antitrust and competition law seem conceptually better equipped to address systemic issues of data inequality than property rights or otherwise individual rights-based approaches. Yet, as we shall see, there are also important limitations inherent in antitrust and competition law that need to be acknowledged. As with data protection and privacy law, there are commonalities but also important differences between the respective legal regimes in the US and Europe, with most jurisdictions elsewhere gravitating towards the EU's approach to competition law.[275]

The platform power that US-based corporations have accumulated has emerged as one main target for increased antitrust scrutiny in the US.[276] Such inquiries grapple with the questions of concentrated control over data that our paper is concerned with. However, neither data accumulation as such nor the power to datafy is the concern. Even under resurgent yet highly contested "Neo Brandeisian" framing of antitrust law advocated by Lina Khan and others, there needs to be some kind of anti-

---

[272] Decision of 6 February 2019, B6-22/16, https://perma.cc/TZR5-KFB9; upheld in preliminary proceedings by the German Federal Court of Justice, decision of 23 June 2020, KVR 69/19.

[273] *See, e.g.*, the working paper by the French autorité de la concurrence and the German Bundeskartellamt on competition law and data (May 10, 2016), https://perma.cc/YF9U-6BKT; the Stigler Committee on Digital Platforms: Final Report (2019); the Report of the UK's Digital Competition Expert Panel, Unlocking digital competition (March 2019); Jacques Crémer, Yves-Alexandre de Montjoye & Heike Schweitzer, *Competition policy for the digital era* (2019).

[274] U.S. House of Representatives, Subcommittee on Antitrust, Commercial and Administrative Law on the Judiciary, Majority Staff Report and Recommendations, Investigation of Competition in Digital Markets (2020), fhttps://judiciary.house.gov/uploadedfiles/competition_in_digital_markets.pdf. In October 2020, the US Department of Justice, together with several states, filed a lawsuit against Google alleging that the company unlawfully maintained monopolies in the markets for general search services, search advertising, and general search text advertising in the US through anticompetitive and exclusionary practices: https://www.justice.gov/opa/pr/justice-department-sues-monopolist-google-violating-antitrust-laws. In December 2020, the Federal Trade Commission sued Facebook alleging that the company illegally maintained a monopoly for personal social networking through anticompetitive conduct: https://www.ftc.gov/news-events/press-releases/2020/12/ftc-sues-facebook-illegal-monopolization.

[275] Anu Bradford, Adam Chilton, Katerina Linos & Alex Weaver, *The Global Dominance of European Competition Law Over American Antitrust Law* (2019) 16 JOURNAL OF EMPIRICAL LEGAL STUDIES 731. *See also* Thomas K. Cheng, *Convergence and Its Discontents: A Reconsideration of the Merits of Convergence of Global Competition Law* (2012) 12 CHICAGO JOURNAL OF INTERNATIONAL LAW 433.

[276] *See* footnote 274. *See further* Lina Khan, *Sources of Tech Platform Power* (2018) 2 GEORGETOWN LAW TECHNOLOGY REVIEW 325.

competitive conduct to justify interventions by antitrust authorities.[277] The harm that antitrust law, even in this reimagined form, remains concerned about is the potential harm to competition caused by concentrated market power.[278] This concern may overlap to a significant extent with the data inequality concerns outlined above, especially when market power enables platform companies to attain control over expansive data infrastructures.[279] But certain data inequality concerns will remain out of focus of antitrust analysis. For starters, data concentration is not always a function of market concentration and may in fact occur across markets, even transnationally. Moreover, data concentration as such is not a concern. Data concentration is only seen as an enabling factor that might lead to a monopoly position, which is in itself not a problem from an antitrust law perspective, unless a monopolist uses its position to engage in anticompetitive conduct. Likewise, when scrutinizing transactions through which companies' data or data infrastructures are being acquired, antitrust law only intervenes if there are anticompetitive effects.[280] Antitrust law is only concerned with distributional effects when they are the result of anticompetitive conduct. Crucially, however, data inequality harms are not necessarily caused by harms to competition. Moreover, the non-existence or non-availability of data — for example, data that is necessary to measure sustainable development or to craft well-tailored public policies — is not an issue that antitrust law is designed to address at all.[281]

Competitors, however, could have a chance to gain access to data and data infrastructures under US antitrust law, if the US were to revive, renew, and expand the "essential facilities doctrine"[282] — despite

---

[277] What is known as antitrust law in the US harks back to the Sherman Antitrust Act of 1890, 15 U.S.C. §§ 1–7 which was passed in response to the concentration of corporate power during the Gilded Age in the US. *See* Tim Wu, THE CURSE OF BIGNESS: ANTITRUST IN THE NEW GILDED AGE (2018). From the 1970s onwards, the so called "Chicago School" reoriented US antitrust law successfully towards a law and economics driven analysis of "consumer welfare". *See, e.g.*, Herbert Hovenkamp and Fiona Scott Morton, *Framing the Chicago School of Antitrust Analysis*, UNIVERSITY OF PENNSYLVANIA LAW REVIEW (forthcoming). More recently, this shift has been criticized for casting aside certain political dimensions of antitrust law. *See, e.g.*, Ariel Katz, *The Chicago School and the Forgotten Political Dimension of Antitrust Law* (2020) 87 THE UNIVERSITY OF CHICAGO LAW REVIEW 413. *See* for a more general critique of the economic efficiency paradigm Jedediah Britton-Purdy, David Singh Grewal, Amy Kapczynski & K. Sabeel Rahman, *Building a Law-and-Political-Economy Framework: Beyond the Twentieth-Century Synthesis* 129 YALE LAW JOURNAL 1784, 1800–1802 (2020). On the "Neo Brandeisians" *see* Lina Khan, *The New Brandeis Movement: America's Antimonopoly Debate*, JOURNAL OF EUROPEAN COMPETITION LAW & PRACTICE 131 (2018).

[278] *See, e.g.*, Lina Khan, *Amazons Antitrust Paradox* 126 YALE LAW JOURNAL 710 (2017) (arguing that the predominant framework in US antitrust—specifically its pegging competition to "consumer welfare," defined as short-term price effects—is unequipped to capture the architecture of market power in the digital economy and risks missing potential harms to competition).

[279] *See above* Section I.C.

[280] *See, e.g., U.S. v. Google Inc. and ITA Software, Inc.* (Oct. 5, 2011). The court approved Google's acquisition of ITA Software, thereby allowing Google to acquire data and algorithms used to combine and parse flight information from airlines, including pricing and availability data. The court imposed a time limit remedy which required Google to license ITA's data infrastructure to other websites for a period of five years.

[281] *See also* Orla Lynskey, *Regulating 'Platform Power'* LSE LAW, SOCIETY AND ECONOMY WORKING PAPERS 1 (2017) (arguing that platform power is over and under-inclusive and that certain issues need to be addressed outside a competition law framework). *See also* Ariel Ezrachi, *EU Competition Law Goals and the Digital Economy*, OXFORD LEGAL STUDIES RESEARCH PAPER NO 17/2018 (asking: is this a competition problem?).

[282] Arguing for such a move Nikolas Guggenberger, *Essential Platforms*, STANFORD TECHNOLOGY LAW REVIEW (forthcoming 2021), https://ssrn.com/abstract=3703361. *See also* Zachary Abrahamson, *Essential Data*, 124 YALE LAW JOURNAL 867 (2014). *See generally* Brett Frischmann & Spencer Weber Waller, *Revitalizing Essential Facilities* (2008) 75 ANTITRUST LAW JOURNAL 1. The basic idea behind the doctrine is that a monopolist who owns "a facility essential to other competitors" must grant reasonable use of that facility under certain conditions.

US Supreme Court case law that has gradually restricted its potential scope of application and that seems, at least so far, largely oblivious to digital platform dynamics and related harms.[283] The EU's analogue to the "essential facilities doctrine" seems better equipped to address this particular type of data inequality — lack of access to data and data infrastructures by competitors — because the relevant thresholds to establish abuse of a dominant market position seem more favorable to competitors seeking access than the corresponding thresholds under US antitrust law.[284] In the absence of competition law claims, the EU has established certain access to data rights beyond personal data portability for businesses and consumers in certain sectors.[285] Note, however, that such mandatory data sharing may perpetuate the initial determinations made under recourse to the data infrastructure controller's power to datafy, unless the regulation itself mandates what kind of data has to be generated, retained, and shared.[286] Crafting such regulatory intervention is complicated, and the EU has so far refrained from far reaching mandatory data sharing, instead settling for sectoral access to data regimes (e.g. for electricity data in the context of "smart meters")[287] and targeted access to data rights also common in other jurisdictions (e.g. for access to automotive data for repair shops).[288]

The corollary to provisions that enable access to and transfer of data (data portability) are interventions that require interoperability between data infrastructures.[289] Interoperability can be facilitated through private standard-setting organization (e.g. with regard to data standards),[290] required under antitrust and competition law,[291] or mandated by regulators.[292] The latter approach is the one traditionally adopted by telecommunication regulators concerned about lack of interoperability between different telecommunication network providers (common carrier doctrine), and some have suggested that similar principles should apply to certain Internet platforms. The proposed DMA makes some steps in this direction by requiring "gatekeepers" to allow business users and providers of ancillary services

---

[283] *See Verizon Communications Inc. v. Law Offices of Curtis V. Trinko*, LLP, 540 U.S. 398 (2004) (holding that antitrust remedies are not available for regulated industries) and *Ohio v. Am. Express Co.*, 2018 138 S. Ct. 2274 (upholding the restrictive measures of a credit card company for lack of harm). *See* Tim *Wu, Ohio v American Express - The American Express Opinion, the Rule of Reason, and Tech Platforms*, 7 JOURNAL OF ANTITRUST ENFORCEMENT 104, 117 (2019).

[284] For an in-depth analysis of EU competition law from an essential facilities perspective, *see* Inge Graef, EU COMPETITION LAW, DATA PROTECTION AND ONLINE PLATFORMS: DATA AS ESSENTIAL FACILITY (2016); Inge Graef, *When Data Evolves into Market Power–Data Concentration and Data Abuse under Competition Law, in* DIGITAL DOMINANCE: THE POWER OF GOOGLE, AMAZON, FACEBOOK, AND APPLE (Martin Moore & Damian Tambini (eds.), 2018).

[285] *See, e.g.,* Directive (EU) 2019/944 of 5 June 2019 on common rules for the internal market for electricity, OJ (2019) L 158/125.

[286] Transparency obligations regarding data infrastructures, including disclosure of choices made and methodologies involved in data generation is another regulatory avenue. *See further below* Section III.C.

[287] For a summary of EU law on sharing of non-personal data, *see* Support Centre for Data Sharing, *Analytical report on EU law applicable to sharing of non-personal data*, DG CONNECT (Jan. 24, 2020), https://eudatasharing.eu/index.php/legal-aspects/report-eu-law-applicable-sharing-non-personal-data.

[288] The Parliament of the Commonwealth of Australia, Competition and Consumer Amendment (Motor Vehicle Service and Repair Information Sharing Scheme) Bill 2020, Exposure Draft. The draft bill is currently open for public comments until 31 January 2021. Full text of the proposed legislation is available at https://treasury.gov.au/consultation/c2020-128289.

[289] *See, e.g.*, Jacques Crémer, Yves-Alexandre de Montjoye & Heike Schweitzer, *Competition policy for the digital era* (2019) (distinguishing between protocol interoperability, data interoperability, and full protocol interoperability); *see also* Przemysław Pałka, *The World of Fifty (Interoperable) Facebooks*, 51 SETON HALL REVIEW (forthcoming 2021), https://ssrn.com/abstract=3539792.

[290] *See, e.g.*, Michal Gal & Daniel L. Rubinfeld, *Data Standardization* 94 NYU LAW REVIEW 737 (2019).

[291] *See, e.g.*, Chris Riley, *Unpacking interoperability in competition* 5 JOURNAL OF CYBER POLICY 94 (2020).

[292] *See* Ian Brown, *Interoperability as a Tool for Competition Regulation* (2020), https://osf.io/preprints/lawarxiv/fbvxd/.

access to and interoperability with the operating system, hardware, or software features used by the gatekeeper for its own ancillary services.[293] Such a regulatory intervention would diminish the infrastructural control of developers of operating systems and device manufacturers (such as Apple). While envisaged as complementary to competition law remedies, including the essential facilities doctrine, the proposed DMA shares with competition law a framing that is chiefly concerned with gatekeepers' impact on innovation and competition.[294] Although the proposed DMA is not challenging data inequality as such, it may still have positive effects in this regard by making visible and redistributing infrastructural control.

While the focus on competition and innovation is a commonality, regulatory *ex ante* intervention as proposed under the DMA differs from *ex pos*t remedies under competition law also in so far as the latter requires engagement with core competition law concepts such as market definition and market dominance. The ways in which antitrust and competition regulators in both the US and the EU have analyzed mergers in the tech sector (at least so far) are illustrative of the resulting blind spots. The acquisition of additional data infrastructures as such is of no concern from an antitrust or competition law perspective, unless it leads to a monopoly or dominant market position with adverse effects on competition. Such prognosis is of course dependent on accurate information about what the companies plan to do with the acquired infrastructures (or the data). This was not the case when Facebook misled the European Commission about the possibility of aggregating Facebook data with WhatsApp data.[295] Even without outright deception, the estimation of prognostic use is susceptible to miscalculations. In some merger cases, companies have made data sharing concessions to ease data concentration concerns.[296] In many other cases, however, companies managed to survive merger control scrutiny because the authorities were only focused on the effects in particular markets and were not interested in broader concerns around increased data concentration, including those that might have resulted from service integration across different markets (see discussion *below*).[297] This purist view of competition law can lead to a systemic overlooking of large-scale data accumulation across complex data infrastructures.[298] For some, this limited and clearly delineated scope of

---

[293] Proposed DMA, article 6(f) and recital 52.

[294] *See, e.g.,* proposed DMA, recital 54: "Gatekeepers benefit from access to vast amounts of data that they collect while providing the core platform services as well as other digital services. To ensure that gatekeepers do not undermine the contestability of core platform services as well as the innovation potential of the dynamic digital sector by restricting the ability of business users to effectively port their data, business users and end users should be granted effective and immediate access to the data they provided or generated in the context of their use of the relevant core platform services of the gatekeeper, in a structured, commonly used and machine-readable format. This should apply also to any other data at different levels of aggregation that may be necessary to effectively enable such portability. It should also be ensured that business users and end users can port that data in real time effectively, such as for example through high quality application programming interfaces. Facilitating switching or multi-homing should lead, in turn, to an increased choice for business users and end users and an incentive for gatekeepers and business users to innovate."

[295] EU Commission, *Mergers: Commission fines Facebook €110 million for providing misleading information about WhatsApp takeover*, IP/17/1369 (May 18, 2017).

[296] Relevant cases under European competition law include Case COMP/M.4854 *TomTom/TeleAtlas* [2008] OJ C237/53; Case COMP/M.6314 *Telefonica/Vodafone/EE* [2012] OJ C66/122; Case COMP/M.7023 *Publicis/Omnicon* [2014] OJ C84/112.

[297] *See, e.g.*, the European Commission's decision in Case COMP/M.7217 *Facebook/WhatsApp* C(2014)7239 final, which focused exclusively on a potential strengthening of Facebook's position in online advertising and dismissed privacy-related concerns as outside the scope of EU competition law.

[298] *See, e.g.*, Reuben Binns & Elettra Bietti, *Dissolving Privacy, One Merger at a Time: Competition, Data, and Third Party Tracking*, 36 COMPUTER LAW & SECURITY REVIEW (2020) (analyzing data accumulation qua user tracking).

competition law analysis is preferable to an expansion in contradiction to established tenets of the regime.[299] For others, a recalibration of the established regimes is unavoidable due to the versatility and variability of data, thus advocating for integration of data protection and privacy law.[300] These different views have implications for competition law's suitability to address data inequality.

One final, not always fully acknowledged, limitation of antitrust and competition law deserves to be foregrounded: despite some structures for international coordination and a global agenda to install antitrust and competition law regimes in jurisdictions around the world,[301] antitrust and competition law remain fundamentally concerned with the anticompetitive effects on domestic markets. US antitrust law is concerned with the US market, European competition law focuses on the European market, and so on. This statement is true regardless of occasional invocation of "extraterritorial" jurisdiction when anticompetitive conduct occurs outside their jurisdiction but materializes within their jurisdiction (under the effects doctrine), as the analysis remains confined to domestic effects and ignores global implications.[302] The US Supreme Court has explicitly dismissed the idea that US antitrust law should remedy anti-competitive conduct abroad.[303] And even if antitrust and competition authorities were to police anti-competitive conduct abroad, conventional economic analysis would be focused on neatly delineated domestic markets. All this might seem unremarkable and rather "normal", but it is increasingly out of sync with the reality of global data generation and transnationally distributed yet interconnected data infrastructures. As Tim Mitchell and Hugo Radice have shown, respectively, the idea of a national economy that is congruent with the nation state is a construct whose creation can be attributed to (perceived) econometric necessities.[304] Most economists' traditional focus on national markets corresponds to most lawyers' traditional focus on national law. This framing leads to a misalignment between economic and legal frameworks and a data reality where conventional territorial borders are far less relevant to delineate the relevant components and dimensions of data infrastructures. International law or some other form of inter-public or even "global" law could potentially be used to remedy such misalignment.[305] The competition chapters in contemporary trade agreements achieve nothing to that effect, as they remain mainly concerned with procedural rights of

---

[299] *See* Nicolas Petit, BIG TECH AND THE DIGITAL ECONOMY: THE MOLIGOPOLY SCENARIO (2020) (arguing for regulation, not competition, as the appropriate tool to address non-competition harms).

[300] *See* Orla Lynskey, *Grappling with "Data Power": Normative Nudges from Data Protection and Privacy*, 20 THEORETICAL INQUIRIES IN LAW 189 (2019).

[301] *See, e.g.*, Eleanor M. Fox and Amedeo Arena, *The International Institutions of Competition Law: The Systems' Norms*, *in* THE DESIGN OF COMPETITION LAW INSTITUTIONS (Eleanor Fox & Michael Trebilcock eds., 2012)
(discussing WTO and other international institutions).

[302] *See, e.g.*, Eleanor M. Fox, *National Law, Global Markets, and* Hartford*: Eyes Wide Shut* (2000–2001) 68 ANTITRUST LAW JOURNAL 73 (criticizing the US Supreme Court decision in *Hartford Fire Insurance Co.*
v. California, 1509 U.S. 764 (1993) and EU judgments on extraterritoriality); *see also* Giorgio Monti, *The Global Reach of EU Competition Law*, in: EU LAW BEYOND EU BORDERS: THE EXTRATERRITORIAL REACH OF EU LAW (Marise Cremona & Joanne Scott eds., 2019).

[303] *F. Hoffman-La Roche Ltd. v. Empagran S.A.*, 542 U.S. 155, 165 (2004). But *see* Ralf Michaels, *Supplanting Foreign Antitrust*, 79 LAW AND CONTEMPORARY PROBLEMS 223-247 (2016) (arguing that developed country with effective antitrust enforcement should lend their antitrust enforcement capacity to developing countries).

[304] Tim Mitchell, *Origins and Limits of the Modern Idea of the Economy*, Advanced Study Center, University of Michigan, Working Papers Series, no. 12 (1995) and subsequent writings on economics and the invention of the economy; *see also* Hugo Radice, *The National Economy: A Keynesian Myth?*, 8 CAPITAL AND CLASS 1 (1984) 111-140.

[305] On international law as inter-public law, *see* Benedict Kingsbury, *International Law as Inter-Public Law* 49 NOMOS (MORAL UNIVERSALISM AND PLURALISM) 167 (2009). On varieties, possibilities, and limitations of "global law" *see* Neil Walker, INTIMATIONS OF GLOBAL LAW (2014). On Global Data Law *see* www.guariniglobal.org/global-data-law.

businesses during investigations or push back against governmental interference in markets, while ignoring the possibility that globally distributed but centrally controlled market power may be more than the sum of its parts. Antitrust and competition law remain focused on domestic market power and the competition chapters in trade agreements are mainly concerned with procedural rights of businesses or governmental interference in markets, not with the global concentration of market power. Thus, they are structurally ill-equipped to confront data inequality arising from global control over data infrastructures nurtured by such market power.

III. Confronting Data Inequality for Digital Development

The dominant narrative in the development discourse tends to emphasize the welfare gains stemming from digitalization, data sharing, and data-driven technologies (particularly artificial intelligence/machine learning). International organizations devoted to the promotion of economic growth and human flourishing urge expanded use of digital data for economic gain and social benefit. Although drawbacks of digitalization and risks associated with data use are increasingly being considered, the focus is often on adverse effects on individuals' privacy (and occasionally other) rights, and on security implications caused by increasing reliance on inter-connected systems. Efforts to promote good "data governance" practices often seek to maximize the value of data-as-an-asset, target predominantly public sector actors, and draw inspiration from the extant legal frameworks that we discussed in Part II.[306]

While we do not discount the importance of these efforts, we want to draw attention to a complementary need to consider the unequal power to determine *what* becomes data and, conversely, what does not become data. This requires increased consideration of the power dynamics inherent in the social practices through which data is being generated. In this context, we have also highlighted the role of infrastructural control, particularly when centralized in the hands of corporate actors, in constituting and entrenching unequal control over data.[307] We have illustrated the limitations of extant dominant legal approaches in remedying data inequality as well as the risk that they may, in certain circumstances, further entrench data inequality.[308]

In this final Part, we turn to some interventions that might aid in remedying data inequality, with particular attention to development freedoms of individuals and communities. We focus on countries with developing digital economies, but we emphatically do not believe that data inequality can be resolved in a wholesale fashion without due regard to the particular economic, social, cultural, and historical contexts in which different countries and their populations find themselves in.[309] To the contrary, it follows naturally from our discussion of how data infrastructures shape how and what data is collected and used and for what purpose that contextualization and self-determined experimentation are important pathways towards more equitable and democratic digital development.

At present, it appears that countries looking to develop digital economies are faced with a dilemma. On the one hand, lack of prerequisite physical or digital components (and high costs associated with building them solely domestically) may lead countries to rely on infrastructure provided by the world's leading tech companies which are overwhelmingly based in the US and China. This approach is

---

[306] *See e.g.*, OECD, *The Path to Becoming a Data-Driven Public Sector* (2019), Chapter 2 ("Good data governance is imperative for governments that aim to become more data driven as part of their digital strategy. It can help to extract value from data assets, enabling greater data access, sharing and integration at the organisational level and beyond, and increasing overall efficiency and accountability."); Arturo Muente-Kunigami (Inter-American Development Bank), *We need to urgently review our data governance frameworks* (2020), https://perma.cc/S5KJ-WG6X.

[307] *See above* Part I.

[308] *See above* Part II.

[309] *See also* David M. Trubek, *Law and Development: Forty Years After 'Scholars in Self-Estrangement'*, 66 THE UNIVERSITY OF TORONTO LAW JOURNAL 301, 318 (2016): "Three important ideas have helped shape twenty-first- century law and development: the understanding that development does not follow prescribed script but requires constant experimentation; the recognition that capitalism can take many forms and law will vary with the dominant form of market system; and the idea that legal rights are part of what is meant by development, not just a means to an end."

encouraged by the logic of market efficiencies. Most governments and international organizations proceed on the assumption that capitalism produces economically superior outcomes (at least in the aggregate).[310] On the other hand, because data infrastructures *shape* data and consequently the representations of physical, social, or political phenomena that data aims to capture and reflect (albeit imperfectly), the interest of public constituencies may pull towards more local and collectivist control over data infrastructures and the development of the necessary technical, social, and organizational structures and practices. These latter interests are being supported by growing doubts about, and occasional resistance to an unconditional embrace of a data-driven capitalist economic development model, prompted by the excesses of "surveillance capitalism", the dominance of platforms in "informational capitalism", and post-colonial continuities of data extractivism and exploitation.[311] Some even wonder whether some form of "digital socialism" might be economically viable after all – contrary to Friedrich Hayek's assumptions.[312] What if the unprecedented generation of data – though thus far highly concentrated in the hands of few – and the resulting ability to generate information makes a collectively governed economy and society plausible, or even desirable?[313] The success of entities that rely crucially on data-dependent and highly centralized planning for their commercial success without being subjected to meaningful market pressures lends credence to this possibility.[314]

The need for economies and societies around the world to shape their individual and collective digital destinies coincides with and corresponds to renewed calls to question long standing tenets of the development agenda. The traditional critique of the development's agenda emphasis on economic growth, conventionally measured in the aggregate, is now being reenforced by those who ask about the role of information and communication technologies in development and who advocate for a shift towards other paramount values and objectives, for example human dignity or individual and collective freedom.[315] As the world confronts a global climate crisis, the environmental cost of

---

[310] Branco Milanovic, CAPITALISM ALONE: THE FUTURE OF THE SYSTEM THAT RULES THE WORLD (2019) observes that there is no longer a contest between different economic systems, but rather a question of which variety of capitalism to embrace, and how much state involvement to allow or require. This does not mean, however, that alternatives to capitalism are not imaginable or achievable. *See* Erik Olin Wright, ENVISIONING REAL UTOPIAS (2010).

[311] Shoshana Zuboff, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR THE FUTURE AT THE NEW FRONTIER OF POWER (Profile Books 2019); Julie E. Cohen, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM (2020); Nick Couldry and Ulises A. Mejias, THE COSTS OF CONNECTION: HOW DATA IS COLONIZING HUMAN LIFE AND APPROPRIATING IT FOR CAPITALISM (2019).

[312] Evgeny Morozov, *Digital Socialism? The Calculation Debate in the Age of Big Data*, 116/117 NEW LEFT REVIEW 33 (2019) in response to Viktor Mayer Schönberger and Thomas Ramge, REINVINTING CAPITALISM IN THE AGE OF BIG DATA (2018).

[313] Przemysław Pałka, *Algorithmic Central Planning: Between Efficiency and Freedom*, 83 LAW AND CONTEMPORARY PROBLEMS 145 (2020).

[314] Leigh Phillips and Michael Rozworski, THE PEOPLE'S REPUBLIC OF WALMART: HOW THE WORLD'S BIGGEST CORPORATIONS ARE LAYING THE FOUNDATION FOR SOCIALISM (2019).

[315] *See e.g.*, Manuel Castells & Pekka Himanen (eds.), RECONCEPTUALIZING DEVELOPMENT IN THE GLOBAL INFORMATION AGE (2014) (arguing for "reconceptualizing human development as the fulfilment of human wellbeing in the multidimensionality of the human experience, ultimately affirming dignity as the supreme value of development"). Amartya Sen's capabilities approach is often at the heart of such endeavors. *See* Devinder Thapa & Oystein Saebo, *Exploring the Link between ICT and Development in the Context of Developing Countries: A Literature Review*, 64 THE ELECTRONIC JOURNAL OF INFORMATION SYSTEMS IN DEVELOPING COUNTRIES 1 (2014); Richard Heeks & Jaco Renken, *Data justice for development: What would it mean?*, 34 INFORMATION DEVELOPMENT 90 (2016).

increased digitalization ought to be (re)considered.[316] The intensifying debate about economic inequality must now also confront the impact of digitalization, in particular data inequality-induced information asymmetries and "winner takes all" dynamics common in Western digital economies and even more pronounced transnationally.[317] The COVID-19 pandemic accelerated the embrace of certain affordances provided by digital technologies, most notably Internet enabled video calls. The pandemic also called into question established development priorities, with considerably more emphasis being placed on wellbeing, income equality, and environmental sustainability.[318]

We cannot do justice to these broader questions and debates in this paper, but we want to acknowledge their existence in order to situate the interventions we suggest and to acknowledge their limitations. Considering the lack of robust empirical evidence about optimal approaches to digital development and given the salience, longevity, and path dependencies of infrastructures, developing economies may be wise to retain the freedom to experiment with different developmental approaches.

Addressing data inequality inevitably requires confronting competing values and interests. As we have emphasized throughout this paper, control over data and the power to datafy arises from control over data infrastructures, and these infrastructures are neither agentless nor static; rather, they are deeply political and dynamic. Redistributing existing data (e.g., by making data "open" or by encouraging data philanthropy) or moving away from concentrated infrastructural control to more distributed configurations does not necessarily equalize the power to determine what should and should not become datafied.[319] Focusing solely on questions about how (and by whom) should data be generated without attention to how concentrations of data are accumulated via infrastructural control and how such control is established and maintained, including by exploiting legal ambiguities, risks missing relevant sites for intervention.

Data inequality needs to be addressed with consideration of contemporary capitalist logics and dynamics within which it is situated.[320] An effective reshaping of these logics and dynamics may require

---

[316] For a discussion of environmental harms caused by cloud computing and processing of very large data sets for machine learning, *see* Elettra Bietti & Roxana Vatanparast, *Data Waste*, 61 HARVARD INTERNATIONAL LAW JOURNAL FRONTIERS (2020); *see also* Emma Strubell, Ananya Ganesh, Andrew McCallum, *Energy and Policy Considerations for Deep Learning in NLP*, arXiv:1906.02243v1 (2019).

[317] As Dan Ciuriak, *Rethinking Industrial Policy for the Data-driven Economy*, CIGI Papers No. 192 (Oct. 2018) at 6 puts it: "… the business model of the data-driven economy is based on exploitation of information asymmetry. By further extension, there are fundamental information asymmetries between countries that can build companies on data assets and those that cannot. Information asymmetry is, in some sense, the 'original sin' of the data-driven economy. *See also* Erik Brynjolfsson & Andrew McAfee, THE SECOND MACHINE AGE ch. 10 (2014).

[318] *See also* Kathleen R. McNamara & Abraham L. Newman, *The Big Reveal: COVID-19 and Globalization's Great Transformations*, 74 INTERNATIONAL ORGANIZATIONS 1, 13-15 (2020) (assessing the impact of COVID-19 on globalization and asserting that the pandemic has underscored the importance of digital technologies).

[319] Indeed, in some instances, concentration of infrastructural control might be necessary to empower a previously disempowered constituency. Illustrative of this is the case of indigenous peoples' struggles to reclaim the power to create knowledge about themselves, leading to claims of indigenous data sovereignty. *See* Tahu Kukutai & John Taylor (eds.), INDIGENOUS DATA SOVEREIGNTY TOWARD AN AGENDA, CENTRE FOR ABORIGINAL ECONOMIC POLICY RESEARCH (CAEPR) (2016).

[320] *See* Julie E. Cohen. BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM (2019) (analyzing how law enables informational capitalism); Katharina Pistor, THE CODE OF CAPITAL (2019) (emphasizing the role of lawyers in coding "capital" and shielding it from the democratic process); Amy Kapczynski, *The Law of Informational Capitalism*, 129 Yale Law Journal 5 (2020) (showing how law protects certain data related interests but

a reshaping of law as well. For such an endeavor to succeed, lawyers ought to scrutinize critically the many ways in which law, including international law, has contributed to inequality past and present.[321] Neither legal doctrine nor regulatory dogma nor technical specification are going to remedy data inequality if they are not also accompanied by continuous, iterative, inclusive, public debate, deliberation, contestation, decision-making, and implementation. We thus caution development aid agencies and other similarly positioned actors not to close off but instead encourage expansion of spaces that would allow for such activities to materialize in developing as well as developed economies.[322]

With that preface, in the following, we advance several ideas that might chart pathways for confronting or mitigating data inequality. Some of these ideas may, at first blush, seem experimental and radical, while others may appear to be more iterative and pedestrian. Given the complex interaction of technical, social, and organizational dynamics in data infrastructures, interventions in any one of these dimensions will inevitably produce ripples in others. To that end, we propose both interventions that have direct regulatory effects and others that create enabling environments for more meaningful political consideration and contestation of datafication and digital development. We progress from large-scale and foundational recommendations to more local and targeted suggestions. We consider different actors and scales in the ensuing analysis. It cannot, nor should it, be assumed that nation states, and their peoples and territories, are necessarily the only or most suited actors and scales to address data inequality, which is, fundamentally, both a global and local phenomenon.[323]

Data infrastructures and the entities that control them, transgress territorial borders with relative ease, while the interests of affected publics might be transnationally aligned or in tension with one another.[324] At the same time, nation states remain the dominant form of organized political power and maybe even more so in a world with stark anti-globalist currents. Moreover, they remain the main subjects and objects of international law, as traditionally conceived, and are tasked with steering economic development within the framework of the global economic order as currently constituted. This explains why our recommendations are mainly addressed towards governments and international organizations.

---

not others). *See also* Salome Viljoen, *Democratic Data: A Relational Theory for Data Governance*, YALE LAW JOURNAL (forthcoming), https://ssrn.com/abstract=3727562 (Nov. 23, 2020).

[321] There is extensive scholarship in critical legal studies on these questions. For seminal work on intersectionalist perspectives on law, *see, e.g.*, Kimberlé Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine*, U. Chi. Legal F. 139 (1989). For critical analysis of international law, *see, e.g.*, Ntina Tzouvala, CAPITALISM AS CIVILISATION: A HISTORY OF INTERNATIONAL LAW (2020); Antony Anghie, Bhupinder Chimni, Karin Mickelson & Obiora Okafor (eds.), THE THIRD WORLD AND INTERNATIONAL ORDER: LAW, POLITICS AND GLOBALIZATION (2004); James Gathii, *The Promise of International Law: A Third World View*, GROTIUS LECTURE AT THE 2020 VIRTUAL ANNUAL MEETING OF THE AMERICAN SOCIETY OF INTERNATIONAL LAW (June 25, 2020), https://ssrn.com/abstract=3635509. *See also below* Section III.A. discussing international economic law.

[322] *See also* Amba Kak, *"The Global South is everywhere, but also always somewhere": National Policy Narratives and AI Justice*, AIES '20: PROCEEDINGS OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY (Feb. 2020), https://doi.org/10.1145/3375627.3375859.

[323] *See* Yanni A. Loukissas, ALL DATA ARE LOCAL: THINKING CRITICALLY IN A DATA-DRIVEN SOCIETY (2019), ch 1. ("…data may be shaped by local conditions, yet they serve a combination of needs, near and far….[T]here is no global experience of data, only an expanding variety of local encounters.").

[324] On the interplay between infrastructures, publics and law, *see* Benedict Kingsbury & Nahuel Maisley, *Infrastructures and Laws: Publics and Publicness*, 17 ANNUAL REVIEW OF LAW AND SOCIAL SCIENCES (forthcoming 2021), draft on file with authors.

Our large-scale recommendations can be broadly summarized as: (i) encouraging retention of development freedom (ii) reclaiming infrastructural control, (iii) demanding transparency, (iv) pooling and differentiating access to data, and (v) developing collective data governance mechanisms. In the following, we consider each of these ideas, highlighting their respective legal and infrastructural elements and their relevance for confronting data inequality. Their realization and success will, of course, depend on support by relevant political actors and social movements. We cannot offer an account of how the necessary level of such support could be generated but we believe and hope that laying out ideas about what could be done will at least affect the discourse around data inequality and digital development more generally and might spur relevant actors into action.

A. Retaining Developmental Freedom

As data is growing in importance as a medium for economic, social, and political ordering and as a resource for economic development, governments are tasked with managing the transition towards increasingly digitally mediated economies and societies. As we have seen, their choices may be constrained by existing and emerging commitments under international economic law which may contribute to data inequality.[325] New agreements in the mold of the new NAFTA between the US, Mexico, and Canada (USMCA) and the US-Japan Digital Trade Agreement favor global access to governmental data by encouraging open data policies while limiting states' ability to impose restrictions on cross-border data transfers or to mandate the use of domestic computing facilities.[326] These agreements seek to carry forward core tenets of the world's trading system as constituted under the WTO by applying and extending policy prescriptions and economic development theories to an increasingly data-driven economy.[327] Despite withdrawing from TPP itself, the US successfully inserted its favored rules into instruments of international economic law that are now being advanced and amplified by other countries. The reason for this is likely the belief in a certain model of digital development that can be termed "Silicon Valley Consensus" which emphasizes unrestricted data flows, pushes back against "data localization", lets a mere minimum of data protection regulation suffice, and is generally predisposed against state intervention in the digital economy.[328] States that sign on to this model commit towards "free" data flows as a mode for digital development not just vis-à-vis each other but more broadly, as multi-national corporations are able to avail themselves of

[325] *See above* Section II.A-D.

[326] Thomas Streinz, *International Economic Law's Regulation of Data as a Resource for the Artificial Intelligence Economy*, *in* ARTIFICIAL INTELLIGENCE AND INTERNATIONAL ECONOMIC LAW: DISRUPTION, REGULATION, AND RECONFIGURATION ch. 9 (Shin-yi Peng, Ching-Fu Lin & Thomas Streinz eds., forthcoming 2021). Open data is discussed in more detail below in D.

[327] While potentially effective rhetorically and ideologically, it is conceptually unconvincing to analogize conventional trade in goods and services to "digital trade" just because data "flows" across borders. The most recent agreements – such as the Digital Economy Partnership Agreement pioneered by Singapore, Chile, and New Zealand and the Digital Economy Agreement between Australia and Singapore – drop the "trade" moniker and only speak of "digital economy agreements", while retaining the rules and substantive concepts that the US pioneered through TPP, USMCA, and USJDTA.

[328] Thomas Streinz, *Digital Megaregulation Uncontested? TPP's Model for the Global Digital Economy*, *in* MEGAREGULATION CONTESTED: GLOBAL ECONOMIC ORDERING AFTER TPP ch 14 (Benedict Kingsbury et al eds., 2019). *See also* Anupam Chander*, How Law Made Silicon Valley*, 63 EMORY LAW JOURNAL 639 (2013).

the prerequisite corporate nationality to invoke the treaty protections against regulatory measures that limit cross-border data transfers or require the use of domestic computing facilities.[329]

The EU has realized that this model is at odds with its data protection regime that limits cross-border data transfers according to the (perceived) level of data protection afforded in other jurisdictions. Accordingly, the EU is advancing commitments under international economic law that echo the EU's conceptualization of data protection as a fundamental right and that shield the GDPR from scrutiny and anti-regulatory pressure under international economic law. In other respects, however, the EU's stance is aligned with the Silicon Valley Consensus in pushing back against those modes of data localization in which the EU itself is not engaged in.[330] China has largely refrained from advancing its policy preferences for data governance through instruments of international economic law.[331] In November 2020, China was one of fifteen countries in the Asia-Pacific that signed the Regional Comprehensive Economic Partnership (RCEP) agreement, which contains an e-commerce chapter modeled after TPP but with considerably more leeway for governments to retain restrictive policies.[332] Even with these significant carve outs in place, India, which had been part of the negotiations earlier, refrained from singing RCEP.

The menu of options that presents itself to developing economies in terms of new commitments under international economic law consists of the Silicon Valley Consensus as instantiated in TPP and carried forward in further agreements, including the Digital Economy Partnership Agreement (DEPA);[333] the EU's pro-regulatory position centered on data protection values; and RCEP's new model, which grants broad and self-judging exceptions. The alternative is to refrain from entering into new commitments altogether, as exemplified by India's stance in the WTO and elsewhere.[334] In considering these options, it is worth noting that the digital economy developed without much support

---

[329] Thomas Streinz, *Data Governance in International Economic Law: Non-Territoriality of Data and Multi-Nationality of Corporations* (draft paper, on file with author).

[330] Horizontal provisions for cross-border data flows and for personal data protection in EU trade and investment agreements, Article A.1, https://trade.ec.europa.eu/doclib/html/156884.htm.

[331] China is, however, affecting data governance beyond its borders in other ways, including through digital infrastructure investments. *See* Matthew S. Erie & Thomas Streinz, *The Beijing Effect: China's "Digital Silk Road" as Transnational Data Governance*, NYU JOURNAL OF INTERNATIONAL LAW & POLITICS (forthcoming 2021), https://ssrn.com/abstract=3810256 (Mar. 22, 2021).

[332] Contrast RCEP, articles 12.14 and 12.15 with TPP, articles 14.11 and 14.13. Footnotes 12 and 14, respectively, make clear that it is for the implementing party to decide – and not for other parties or a dispute settlement body to second guess – whether a measure is "necessary". *See* Thomas Streinz, *RCEP's Contribution to Global Data Governance*, AFROMONICSLAW (Feb. 19, 2021), https://perma.cc/ACD5-P2Y9. RCEP is designed to intensify economic ties between the ten ASEAN members and the non-ASEAN countries China, South Korea, Japan, Australia, and New Zealand. *See* Pasha L. Hsieh, *The RCEP: New Asian regionalism and the Global South*, IILJ WORKING PAPER 2017/4 (MEGAREG SERIES), www.iilj.org/megareg.

[333] DEPA was signed (electronically) by Chile, New Zealand, and Singapore in June 2020 and has been in force between Singapore and New Zealand since December 2020. DEPA builds on TPP, which originated initially in an agreement between the same countries. But DEPA goes significantly further by creating new provisions hitherto not seen in international economic agreements, openly announcing itself as an "digital economy agreement" rather than an agreement merely on "electronic commerce" or "digital trade". *See* https://perma.cc/8PLN-9E9P for details.

[334] The economic policy calculation is complicated by a pervasive and somewhat paradoxical lack of data about the digital economy, with neither conventional economic nor trade statistics accounting sufficiently for the value of data and the significance of data flows. It will ultimately depend on countries' particular economic context and trajectory whether or not signing on to any of the currently available models is advisable or whether refraining from entering into such, potentially long lasting, commitments is the more prudent course of action.

from international economic law. The attempts to create new rules in instruments of international economic law, as instantiated by the Silicon Valley Consensus, are more about *entrenching* a vision of the digital economy without forceful regulatory intervention at a moment at which such interventions are on the rise. The data inequality dimensions that we have identified in this paper have been largely ignored in the conventional discourse around "digital trade" and "electronic commerce". This is, in part, a continued legacy of the embedded liberalism that carried the world trading system after WWII and relegated distributional questions to states' domestic social welfare systems.[335] It also reflects, however, an economistic conceptualization of data as a rent generating asset which tends to conceal other dimensions of data inequality.[336] Imagining and designing instruments of international economic law that are more attuned to dimensions of data inequality strikes us as a promising yet uncharted path forward.[337]

Addressing inequalities of the digital economy might thus require novel interventions that may conflict with the theories and concepts under which the world trading system has operated since WWII. This has institutional implications for the WTO, which sits at the heart of this system, but also for development organizations more broadly as they advise countries on which policies to pursue. It is in this context that we caution against premature commitments and suggest that retaining the ability to experiment with different development strategies, digital industrial policies, and attendant social policies might be the more prudent course of action to confront data inequality going forward.[338]

For the moment, some developing economies have pushed back against new commitments on data governance in the WTO's work programme on electronic commerce – which is now proceeding as a plurilateral initiative.[339] The members of the African Continental Free Trade Agreement (ACFTA) were careful not to include provisions on data governance.[340] At the same time, states' existing

---

[335] *See* Andrew T. F. Lang, *Reconstructing Embedded Liberalism: John Gerard Ruggie and Constructivist Approaches to the Study of the International Trade Regime*, 9 JOURNAL OF INTERNATIONAL ECONOMIC LAW 81 (2006). Joseph Stiglitz's GLOBALIZATION AND ITS DISCONTENTS (2003) and Dani Rodrik's THE GLOBAIZATION PARADOX: DEMOCRACY AND THE FUTURE OF THE WORLD ECONOMY (2010) are prominent articulations of critiques of the world trading system and its impact on social welfare. *See also* Sonia E. Roland & David Trubek, *Embedded Neoliberalism and Its Discontents: The Uncertain Future of Trade and Investment Law*, in: WORLD TRADE AND INVESTMENT LAW REIMAGINED 87 (Alvaro Santos, Chantal Thomas & David M. Trubek eds., 2019).

[336] *See above* Sections I.A–B.

[337] *See also* Benedict Kingsbury, Paul Mertenskötter, Richard B. Stewart &Thomas Streinz, *TPP as Megaregulation*, *in* MEGAREGULATION CONTESTED: GLOBAL ECONOMIC ORDERING AFTER TPP ch 2 (Benedict Kingsbury et al eds., 2019) at 60 (imagining truly 21st century agreements, potentially based on the welfare conception championed by Amartya Sen).

[338] *See also* Dan Ciuriak, *Digital Trade: Is Data Treaty-Ready?*, CIGI PAPER NO. 162 (Feb. 21, 2018), https://www.cigionline.org/publications/digital-trade-data-treaty-ready.

[339] 71 WTO members signed a joint statement on electronic commerce during the WTO's 11th ministerial conference in Buenos Aires (*see* WT/MIN(17)/60, Dec. 13, 2017). Negotiations between 76 WTO members commenced in January 2019 (*see* WT/L/1056, Jan. 25, 2019). As of January 2020, 83 WTO members were participating in the negotiations, including all developed countries but only five WTO members from Africa (Benin, Nigeria, Cote d'Ivoire, Kenya, and Cameroon), and no Caribbean or developing Pacific Island countries. *See* Yasmin Ismail, *E-Commerce in the World Trade Organization: History and latest developments in the negotiations under the Joint Statement*, INTERNATIONAL INSTITUTE FOR SUSTAINABLE DEVELOPMENT (Jan. 30, 2020), https://www.iisd.org/publications/e-commerce-world-trade-organization-history-and-latest-developments-negotiations-under. In February 2021, India and South Africa formally criticized these initiatives as in tension with WTO principles of consensus-based multilateralism (*see* WT/GC/W/819, Feb. 19, 2021).

[340] *See* Chijioke Chijioke-Oforji, *The Untapped Potential of the African Continental Free Trade Agreement in the African E-Ecommerce Agenda*, INT. T.L.R. 141 (forthcoming 2021, on file with author).

commitments under international investment law, enshrined in investment chapters of trade agreements and bilateral investment treaties, may be the more consequential constraints in the short-term, especially if states resort to mandatory data sharing requirements to redistribute data.[341] International investment law does not just have anti-regulatory effects on public law measures but may also shape and reshape core concepts of private law.[342] International investment law is likely to be mobilized to contest governmental regulation in the digital domain and may also shape the evolving debate around legal rights to data and questions of legal data ownership.[343] Countries entered into these commitments when digitalization was not yet on the horizon or as pressing as now. Resisting the mobilization of international investment law to protect existing highly asymmetric and concentrated control over data and data infrastructures will be an important challenge for development lawyers going forward and will require prudent judgment by the arbitrators that will be tasked with resolving these conflicts.[344]

If states are relatively unconstrained from existing commitments under international economic law in terms of cross-border data transfers and protections of data as an asset under international investment law, they have more leeway in challenging unequal control over data and data infrastructures. They could, for example, adopt regulatory frameworks that specifically target infrastructural control by platform companies and, where appropriate, mandate access to data for public and commercial actors. Additionally, and especially where regulatory power cannot be effectively asserted, development of independent data collection capacity and commensurate data infrastructures could be pursued. Indeed, if financially and technologically feasible, the latter route might be superior to outright data sharing as the relevant publics could determine for themselves which data ought to be collected and how instead of those choices being dictated by other data collectors in pursuit of their own interests.[345] These options are discussed in the following section.

B. Reclaiming Infrastructural Control

Much of this paper has focused on concentration of infrastructural control in the hands of corporate actors.[346] Although predominantly based in the US and China, and to a far lesser extent in Europe, many large tech companies have been entering previously untapped markets, notably developing

---

[341] *See* Thomas Streinz, *International Economic Law's Regulation of Data as a Resource for the Artificial Intelligence Economy*, *in* ARTIFICIAL INTELLIGENCE AND INTERNATIONAL ECONOMIC LAW: DISRUPTION, REGULATION, AND RECONFIGURATION ch. 9 (Shin-yi Peng, Ching-Fu Lin & Thomas Streinz eds., forthcoming 2021).

[342] *See* Julian Arato, THE PRIVATE LAW CRITIQUE OF INTERNATIONAL INVESTMENT LAW, 113 AMERICAN JOURNAL OF INTERNATIONAL LAW 1 (2019).

[343] *See* Julie E. Cohen, BETWEEN TRUTH AND Power (2019) 257-260. *See also* Section II.B. above.

[344] Where confidence in their ability to resolve such disputes in an equitable and just ways is lacking, withdrawal from the international investment system as currently constituted might be worth exploring. Digital development in a data-driven economy might follow a different logic than in the knowledge-based economy which relied on cheap manufacturing and far-flung global value chains. *See* Dan Ciuriak, *Rethinking Industrial Policy for the Data-Driven Economy* (July 30, 2018), https://ssrn.com/abstract=3223072.

[345] We stress that the normative desirability of these interventions is contingent on the democratic credentials and public values of the political systems that bring them about. If a democratic polity decides to condition or otherwise restrict cross-border data transfers to prevent data extraction, the normative evaluation ought to be different compared to an autocratic imposition of data transfer limitations for reasons of authoritarian self-preservation.

[346] *See above* Section I.C.

economies in Africa and Asia. Often this kind of corporate expansion is being encouraged and welcomed in the hope that increasing supply of resources, expertise, and access to digital technologies might enable developing economies to "leapfrog" in their economic development. Without passing judgment on the merits of this proposition, too often sincere desire for efficiency and quick returns fails not only to take full account of the associated financial, social, and political costs but also of the long-term sustainability of infrastructural dependencies. To be clear, we are not advocating for digital development in isolation; nor are we suggesting that developing economies should forgo any or all services provided by platform companies, cloud, or other data infrastructure providers. Rather, we caution against schematic efficiency and necessity narratives as default positions to justify privatization and corporatization of public services. Concentrated corporate control over critical data infrastructures should not be a quasi-automatic default position. Instead, we advocate for considerate and creative data infrastructure planning that engages the relevant publics who ought to decide for themselves how their environments and lives are being datafied (or, indeed, not).[347]

In this Section we consider regulatory options that depend on considerable state power. For this reason, we emphasize again that concentrated infrastructural control over data in the hands of governments can also be cause for concern.. The regulatory frameworks that we discuss position individual states against foreign data infrastructure controlling corporations, but it is conceivable and perhaps even quite likely that new data jurisdictions will emerge connecting different publics, creating transnational alliances, and presenting different regulatory options that are not aligned with jurisdictional control of a single state.[348] We allude to such possibilities in more detail in our discussion of collective data governance in Section E and, in this part, highlight the potential role that international organizations might play in creating and mediating such new data jurisdictions.

Reclaiming infrastructural control can take different forms. Recent EU initiatives, discussed in Part II above, have opened the door to a regulatory approach that specifically targets online platform companies but does so in a differentiated manner, imposing additional obligations on platforms of particularly large reach. This may be a promising regulatory pathway for countries seeking to gain benefit from resources and infrastructures offered by dominant actors while preserving (or fostering) public control over data infrastructures. Although a direct "transplant" of EU law to other countries would be counterproductive, given the EU's unique political and economic specificities, useful lessons nonetheless can be gleaned from the EU's regulatory agenda. The proposed DSA and DMA seek to reassert public authority over corporate actors with concentrated power over infrastructures and the power to datafy. This might indicate a shift away from the prevalent tendency to treat data as a regulatory object towards regulation of infrastructural control that transcends established concepts under competition law.[349]

It may be difficult for smaller states, without market or political power comparable to the EU, to assert regulatory control over large US- and China-dominated tech companies, or to negotiate their own

---

[347] On reinvigorating planning and foresight in the context of infrastructures, *see* Benedict Kingsbury, *The Wizards of 'Is'*, 8 CAMBRIDGE INTERNATIONAL LAW JOURNAL 171(2019).
[348] *See* Marietje Schaake & Tyson Barker, *Democratic Source Code for a New U.S.-EU Tech Alliance*, LAWFARE (Nov. 24, 2020), https://perma.cc/4P72-WP3Z (calling for a transatlantic alliance against big tech).
[349] *See above* Section II.D.

terms.[350] In addition to traditional lobbying activities, multinational corporations may invoke commitments under international economic law to thwart regulatory initiatives.[351] Indeed, infrastructural control itself may be used to create governmental or corporate pressure against regulatory efforts by smaller states. Henry Farrell and Abe Newman have drawn attention to the dependencies that inter-networked technologies create and how states with control over the companies that build, operate, and maintain these infrastructures gain widespread access to data (through surveillance measures) and may mobilize their control over chokepoints (e.g., by threatening cut offs) to advance their geopolitical objectives.[352] This is not only a question of geopolitical confrontation and alignment but also has implications for economic development, particularly from the perspective of developing economies.[353] China is promoting digital infrastructure investments through its Digital Silk Road, which forms part of the larger Belt and Road initiative, and promises "data sovereignty", yet such promises are tenuous as Chinese technology companies acquire central roles within the relevant data infrastructures transnationally and may attain control over thus generated data.[354] Over the course of 2020, India repeatedly took the extraordinary step of banning certain Chinese apps outright, citing concerns over the "sovereignty and integrity of India".[355] Western platform companies have not yet faced such prohibitions but when subjected to increased regulatory scrutiny or taxation demands, they have repeatedly threatened to withdraw their services.[356] While such threats may seem like a mere negotiation tactic to influence law makers and the public, they may also constitute a leveraging of infrastructural control whose effectiveness will depend on the interdependencies and scale that the relevant platform commands.

Another approach to resisting the dominance of US and Chinese based companies is to invest in the development of alternative public (or public-private) data infrastructures. Even prior to its recent

---

[350] Indeed, it is not yet known whether EU's regulatory efforts will succeed in curbing the infrastructural power of platform companies.

[351] *See* Tim Dorlach & Paul Mertenskoetter, *Interpreters of International Economic Law: Corporations and Bureaucrats in Contest over Chile's Nutrition Label*, 54 LAW & SOCIETY REVIEW 3 (2020) (showing how the transnational food industry challenged food labelling regulation in Chile).

[352] Henry Farrell & Abraham L. Newman, *Weaponized Interdependence*: *How Global Economic Networks Shape State Coercion*, 44 INTERNATIONAL SECURITY 42 (2019); s*ee also* Madison Cartwright, *Internationalising state power through the internet: Google, Huawei and geopolitical struggle*, 9 INTERNET POLICY REVIEW 1 (2020).

[353] *See* Amrita Narlikar, *Must the Weak Suffer What They Must? The Global South in a World of Weaponized Interdependence*, in Daniel W. Drezner, Henry Farrell, and Abraham L. Newman (eds.), THE USES AND ABUSES OF WEAPONIZED INTERDEPENDENCE (2021) ch 16.

[354] *See* Matthew S. Erie & Thomas Streinz, *The Beijing Effect: China's "Digital Silk Road" as Transnational Data Governance*, NYU JOURNAL OF INTERNATIONAL LAW & POLITICS (forthcoming 2021), https://ssrn.com/abstract=3810256 (Mar. 22, 2021).

[355] Reuters staff, *India bans 43 more mobile apps as it takes on China*, REUTERS (Nov. 25, 2020), https://perma.cc/K249-MTD8.

[356] Some illustrative examples include the following: In light of the EU Court of Justice's decision in *Schrems II* (*see above* Section II.A, text accompanying footnote 124), Facebook declared vis a vis the Irish High Court that in the event of a complete prohibition on the transfer of user data to the US it was not clear how Facebook could continue to provide Facebook and Instagram services within the EU (*see* the sworn affidavit by Facebook's head of data protection and privacy for Facebook Ireland Limited, dated Sep. 10, 2020). In response to a proposed Australian media legislation, Facebook announced that it would stop users in Australia (as well as abroad) from sharing local and international news. *See* Will Easton, *Changes to Sharing and Viewing News on Facebook in Australia*, FACEBOOK (Feb 2021), https://perma.cc/7HY7-5AS9. Ride hailing companies Lyft and Uber threatened to pull out of California, if they were forced to classify drivers as employees. *See* Andrew J. Hawkins, *Lyft joins Uber in threatening to pull out of California over driver status*, THE VERGE (Aug. 12, 2020), https://perma.cc/8QHV-CXK3. They eventually defeated the measure via a ballot measure (proposition 22). *See* Kate Conger, *Uber and Lyft Drivers in California Will Remain Contractors*, NY TIMES (Nov. 7, 2020), https://perma.cc/S5BU-KW5A.

regulatory initiatives, the EU supported the development of a European, independent cloud infrastructure (GAIA-X) and promised to make this infrastructure available to others, conditional on adherence to the EU's regulations, in particular GDPR.[357] While the technological and economic success of this initiative remains to be seen, and its potential appeal to entities outside the EU remains uncertain, GAIA-X signals increasing awareness of infrastructural dependencies amidst raising geopolitical contestation between the US and China, which are home to the world's leading cloud providers. Developing countries may benefit if rising competition between different cloud providers not only brings down costs but also increases flexibility about the terms under which this infrastructure is being provided. Similar arguments can be made about reducing dependencies on communication and e-commerce platforms controlled by a small group of US- and China-based technology companies. With proper support and funding alternative platforms for public communication and e-commerce may emerge.[358]

International development organizations could support initiatives to lessen infrastructural dependencies and related digital inequalities and to encourage local digital development and experimentation with public good oriented collective governance frameworks. The Universal Postal Union (UPU) chose to become a cloud services provider itself in lieu of relying on established commercial cloud services. UPU decided to locate the infrastructure and data in the same country that hosts the UPU headquarters, under a jurisdiction that fully respects United Nations privileges and immunities, and to work with a local communications provider. [359] The UPU case opens the possibility that individual IOs or consortia of IOs could provide or support cloud service infrastructure that could link up to create an interconnected cloud federation.[360] For other types of support for data infrastructures one might look to the Federated Information System for the Sustainable Development Goals (FIS4SDGs), an initiative led by the Statistics Division of Nations Department of Economic and Social Affairs (DESA) in partnership with Esri, a company that supplies geographic information system software and geodatabase management applications.[361] The FIS4SDGs initiative is based on the principle of "national ownership", with National Statistical System implementing "internationally agreed standards" for production and dissemination of data and statistics.[362] The National Statistical Offices are envisioned to have a leadership role, "coordinating the [national statistical systems] and improving cooperation between data producers, supporting statistical work of line ministries and other

[357] GAIA-X, *Project GAIA-X: A Federated Data Infrastructure as the Cradle of a Vibrant European Ecosystem* (Oct. 29, 2019), https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Publications/project-gaia-x.html.

[358] France reportedly has plans to launch government versions of Airbnb and Booking.com. Adam Forrest, *France plans government version of Airbnb and Booking.com*, THE INDEPENDENT (May 21, 2020), https://perma.cc/PW4F-NJW8. On public service digital media infrastructure more generally, *see* Ethan Zuckerman, *The Case for Digital Public Infrastructure,* THE KNIGHT FIRST AMENDMENT INSTITUTE (Jan. 17, 2020), https://perma.cc/27TW-KVPM. *See also* Sebastian Benthall & Jake Goldenfein, *Essential Infrastructures*, PHENOMENAL WORLD (July 27, 2020), https://perma.cc/3HWK-Z5KA.

[359] For a review of cloud computing services in the UN system, *see* Jorge T. Flores Callejas & Petru Dumitriu, *Managing cloud computing services in the United Nations system: Report of the Joint Inspection Unit*, (2019) JIU/REP/2019/5, https://www.unjiu.org/content/managing-cloud-computing-services-united-nations-system.

[360] This might enable IOs, as cloud providers, to link up to GAIA-X or to other regional data sharing infrastructures. The European Commission's European Strategy for Data contemplated Memoranda of Understanding with EU Member States, starting with those having existing cloud federation and data-sharing initiatives. *See* European Commission, *A European Strategy for Data*, COM(2020) 66 final, p. 18.

[361] *Federated Information System for the SDGs: A platform for the sharing of national and global statistical and geospatial data for the 2030 Agenda,* UN DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, https://perma.cc/ER8K-9MLZ.

[362] Presumably, these standards would be agreed upon under the auspices of the UN Statistical Commission.

entities, and validating data from different sources for consistency, accuracy and reliability…."[363] The Country Data Hubs will contain geospatially enabled datasets pertaining to specific SDG indicators, as well as interactive analytical visualization and communication applications, such as Story Maps.[364] Through the federated architecture, Country Data Hubs can share with each other SDG-relevant data "enabling users to not only access the data they need when they need it, but also ensure the traceability and accountability of the data, which is maintained at its source." [365] Data hubs of international agencies will aggregate data from national data hubs and allow users to access harmonized data.[366] The UN has already introduced the global Open SDG Data Hub[367] and a number of countries have also launched their own SDG data hubs. [368] This type of data infrastructure, if executed well, can provide a balance between local control over data production processes, collective and deliberative decisions over data use, and transnational data sharing for collectively agreed-upon purposes.

Private commercial actors are not necessarily excluded from data infrastructures created and supported by either local constituencies or international organizations. It is imperative, however, that their engagement is attuned to the dynamics of infrastructural control over data that we have highlighted throughout this paper. The contractually agreed terms ought to be mindful of dependencies (ensuring interoperability to enable switching as needed) and be careful about questions of data generation and control. Negotiating such terms often needs to confront a reality of severe power imbalances. Fostering public data infrastructures may not only produce competitive pressures on commercial actors but may also set standards for best practices and terms under which data is generated and used. Furthermore, collective pooling of resources might rebalance the negotiating power of public entities. In this respect, it is worth noting that international organizations have vast amounts of diverse data that might be usefully pooled together and, with appropriate safeguards, be deployed to create global public-private data sharing platforms that would enable differentiated access to data to different actors.[369] The collective political power of international organizations might also be leveraged to set fair terms for participation of commercial actors.

Last, it is worth noting that it is not just international institutions but international law itself that could be mobilized in the quest of reclaiming infrastructural control over data. Estonia is often hailed as a role model for successful digital transformation as the government decided to introduce novel digital infrastructures for identification and governmental services and encouraged wide-spread adoption of digital technologies by citizens and businesses. When faced with the choice of where to store governmental data to shield it from cyberattacks, Estonia initially considered partnering with a US-

---

[363] *Federated Information System for the SDGs: A platform for the sharing of national and global statistical and geospatial data for the 2030 Agenda*, UN DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS (2019), https://www.unescap.org/sites/default/files/Session_4_Intro_to_federated_information_system_for_the_SDGs_WS_National_SDG_10-13Sep2019.pdf. It is emphasized that the data published by data hubs will be "authoritative data".

[364] *The Federated Information System for the SDGs From Vision to Scale*, THE UNITED NATIONS STATISTICS DIVISION (2019), https://unstats.un.org/unsd/statcom/50th-session/side-events/20190307-1L-Federated-Information-System-for-the-SDGs.pdf.

[365] *Id.*

[366] *Interconnected data hubs and public participation: the data revolution is underway*, UN DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS (Mar. 9, 2018), https://perma.cc/PF26-CGRH.

[367] Sustainable Development Goals, *Welcome to the Open SDG Data Hub*, http://www.sdg.org.

[368] For a short case study on the UAE's SDG Data Hub, see *UAE Data Hub Drives Sustainability Goals*, ESRI, https://www.esri.com/en-us/industries/government/departments/lighthouse-case-study.

[369] For discussion of differentiated data access, *see below* Section III.D.

based commercial cloud provider but eventually settled on a "data embassy", a data center located physically in Luxembourg and protected by an agreement between the two governments, with the necessary technology provided by various private sector entities. Estonia's journey towards its "data embassy" is instructive not only in highlighting the salience of the bond between infrastructures and attendant legal structures under public international law (government to government) and transnational private law (between the government and businesses), but also as an instantiation of creative thinking in light of its particular geopolitical, economic, and regulatory context.

C. Demanding Transparency

One of the key preconditions for regulating infrastructural control over data and for planning digital policy more generally is knowledge about how relevant actors exercise such control, what data they generate and accumulate and through what means, and how they use infrastructures to entrench their market positions and cement control over data infrastructures and data. These questions are related to but also different from the dominant discourse about the opacity and inscrutability of algorithms that is compounded by the rise of artificial intelligence/machine learning algorithms and humans' inability to comprehend the inferences on which these algorithms rely.[370] Antitrust and competition law investigations, such as those carried out in the US, EU, and other jurisdictions may reveal the extent of control over data and related data generating and transacting processes to the authorities, albeit only to the extent to which such disclosures are necessary for the assessments.[371]This is different from and falls short of demanding transparency about corporate data generation to reveal infrastructural power and resulting data control and datafication power asymmetries for purposes of public oversight, contestation, and deliberation. Transparency over data inequality entails asking about how much and what kind of data corporations control, as well as demanding the accompanying metadata outlining the context within which the data was collected.[372] Such demands will not only facilitate ascertaining how economically valuable the accumulated data are but also will provide insight into the process of datafication itself.

Despite the widespread assumption that platform companies control vast amounts of data, surprisingly little is known about *how much* data they actually control, how the data they control is being generated and what economic value "their" data holds. The inability to account for data as an economic asset is, at least in part, a function of contemporary accounting standards which do not account for much data despite the widespread belief that data is becoming companies' most important asset. [373] Jonathan Haskel and Stian Westlake have described how the "knowledge economy" increasingly relies on investment in research and developments which leads to ideas that may or may not be protected under intellectual property law and are only imperfectly accounted for under existing accounting standards.[374] The increasing salience of data for successful businesses compounds the

---

[370] *See, e.g.*, Frank Pasquale, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Andrew D. Selbst and Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM LAW REVIEW 1085 (2018).

[371] On potential and limitations of such inquiries and their remedies *see above* Section II.D.

[372] *See* Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford, *Datasheets for Datasets*, arXiv:1803.09010v7 (2020).

[373] *See above* Section I.A.

[374] Jonathan Haskel and Stian Westlake, CAPITALISM WITHOUT CAPITAL: THE RISE OF THE INTANGIBLE ECONOMY (2017).

problem as the creation of data is not necessarily commensurate with the investment undertaken to create the data. No one knows how much data the world's leading technology companies control. As long as this is the case, scholars and policymakers alike are left with theoretical arguments and mere guessing about the extent of contemporary data control asymmetries and the extent to which these stem, as we posit, from control over data infrastructures.

The accounting for data problem extends to conventional economic and trade statistics which largely do not account for data.[375] Instead, they tend to measure the excesses of the digital economy, where control over data gets commercialized, especially through advertising.[376] This is not only a problem for those who are tasked with advising governments on the state of the domestic and global economy or those who rely on this information for their financial and commercial strategies but also severely undermines the political discourse around digital development and which strategies to pursue. Critics of contemporary digital development strategies struggle to substantiate their arguments in the absence of reliable data. Conversely, proponents of digital development that adhere to the Silicon Valley Consensus likewise struggle to make their case for unrestricted data flows and against data localization measures as they have to rely on questionable proxies (such as bandwidth expansion) or general assumptions (about regulatory uncertainty) in the absence of more precise and differentiated data about who has what kind of data (data control) and between whom data flows.[377]

Some platform companies have begun to make certain kinds of data available to researchers.[378] Such efforts at "voluntary" data sharing, sometimes branded as "data philanthropy",[379] are inherently one-sided as data demanders often do not even know which data exists and data holders often hide behind data protection laws to claim that sharing is impossible (without detailing why).[380] In addition, corporate actors deploy a variety of legal tools – trade secrets protection, non-disclosure agreements, and property claims – to prevent disclosures.[381]

This corresponds with related obfuscating strategies under which companies issue "transparency reports" about certain data-related activities, which are often transparent in name only as they hide important information in "aggregated data" or are short on explanations.

---

375 Dan Ciuriak, *Unpacking the Valuation of Data in the Data-Driven Economy*, https://ssrn.com/abstract=3379133 (Apr. 27, 2019).

376 Mariana Mazzucato, THE VALUE OF EVERYTHING: MAKING AND TAKING IN THE GLOBAL ECONOMY (2018).

377 The EU launched a project on data flow monitoring to map current data stocks and flows within the EU territory. The effort is, however, so far entirely based on voluntary surveys, and hence unlikely to produce an accurate picture of data flow reality. *See* European Commission, *The European Data Flow Monitoring* (Mar. 9, 2021), https://perma.cc/NC59-7L62.

378 *See, e.g.*, Google Research's release of the Objectron Dataset, a machine-learning dataset for 3D object recognition, GOOGLE AI BLOG (Nov. 9, 2020), https://perma.cc/4STP-SMNC. Alibaba released datasets abouts the servers and running tasks in its production clusters; *see* Alibaba Tech, *Open Season for Research: Alibaba Releases Cluster Data from 4000 Servers*, HACKERNOON (Jan. 22, 2019), https://perma.cc/7KKT-V88N. In 2021, Twitter announced it will offer a full history of its full-archive search endpoint to any researcher or developer who applies as part of the launch of a new academic research track; *see* Nick Statt, *Twitter is opening up its full tweet archive to academic researchers for free*, THE VERGE (Jan. 26, 2021), https://perma.cc/5MUT-REZS.

379 Yafit Lev Aretz, *Data Philanthropy*, 70 HASTING LAW JOURNAL 1491 (2018-2019).

380 *See* Mathias Vermeulen, *The keys to the kingdom. Overcoming GDPR-concerns to unlock access to platform data for independent researchers*, OSF PREPRINTS (Nov. 27, 2020), doi:10.31219/osf.io/vnswz.

381 *See above* Sections II.A-C about how these legal technologies may contribute to data inequality.

Certain data protection laws give individuals the (often costly and resource-intensive) right to inquire about the personal data that companies hold about an individual.[382] Data protection and cybersecurity laws require disclosure of certain cyber incidents and data breaches.[383] But there is currently no law that systematically addresses the lack of transparency when it comes to control over data in the digital economy. Regulatory bodies have been reluctant to demand such transparency. Statistical units of governments and international organizations have begun to realize that private actors might have superior data but have refrained from demanding disclosure and sharing of such data.

While this may seem radical, we suggest that more forceful governmental intervention may be needed to remedy data inequality. The transparency requirements we envision would depart from the individual rights-based or incident-based approach under which individuals can demand access to 'their' data or companies need to disclose information about a data breach or cybersecurity incident. Instead, controllers of data infrastructures above a certain threshold (to be determined, for example, by market capitalization, market share, number of users, or type of data) would be required to disclose how much data of what kind they control and through what means (in other words, which data infrastructures).[384] Such a requirement would be akin to the financial disclosure requirements imposed on financial institutions of systemic stature.[385] While this would certainly impose considerable compliance costs and may even necessitate investment in infrastructures necessary to make such determinations (including to protect individuals' privacy as far as personal data is concerned), we do not believe that imposing such requirements on the largest providers of data infrastructures and controllers of data would render their operations unprofitable. Only the world's most powerful regulators with commensurate market power (and possibly multilateral standards setting bodies) will be able to demand and effectively enforce this level of commitment towards transparency. Ideally, such measures would create positive spill-over effects for others, if companies are forced to or decide to implement heightened transparency requirements globally.[386] Increased transparency could also create opportunities for activists to challenge exploitative datafication on a population-level scale. .[387] In the meantime, and concordantly, other actors can make steps in the same direction by demanding transparency whenever they negotiate contracts with data infrastructure providers and by banding together, if necessary, to increase their collective bargaining power.

---

[382] *See* Jef Ausloos & Michael Veale, *Researching with Data Rights*, TECHNOLOGY AND REGULATION 136 (2021), available at https://techreg.org/index.php/techreg/article/view/61 (showing how researchers can leverage data rights to gain access to enclosed datasets).

[383] Mark Verstraete & Tal Zarsky, *Optimizing Breach Notifications*, UNIVERSITY OF ILLINOIS LAW REVIEW (forthcoming, 2021), https://ssrn.com/abstract=3650724 (July 14, 2020).

[384] For a similar proposal based on GDPR access to data rights *see* René L. P. Mahieu & Jef Ausloos, *Recognising and Enabling the Collective Dimension of the GDPR and the Right of Access*, LawArXiv (July 2, 2020) https://doi.org/10.31228/osf.io/b5dwm. As other data protection-based approaches, their proposal is limited to personal data. *See above* on this limitation Section II.C.

[385] *See* for similar ideas drawing on financial market regulation Salome Viljoen & Sebastian Benthall, *Data Market Discipline: From Financial Regulation to Data Governance*, JOURNAL FOR INTERNATIONAL AND COMPARATIVE LAW (forthcoming 2021), https://ssrn.com/abstract=3774418 (Jan. 27, 2021).

[386] This is the dynamic that Anu Bradford has theorized as the "Brussels Effect": Anu Bradford, THE BRUSSELS EFFECT: HOW THE EUROPEAN UNION RULES THE WORLD ch. 2 (2020).

[387] For an example of an effective challenge to the exercise of data-based administrative power by a state, *see NJCM v. the Netherlands* (SyRI), ECLI:NL:RBDHA:2020:1878. *See also* Christiaan van Veen, *Landmark judgment from the Netherlands on digital welfare states and human rights*, OPENGLOBALRIGHTS (Mar. 19, 2020), https://perma.cc/9VWN-MFD3.

The proposed DMA and DSA may mark a turning point as they impose significant transparency obligations, mandating data access not only for supervisory authorities but also for vetted researchers (under DSA), noting that "[i]nvestigations by researchers on the evolution and severity of online systemic risks are particularly important for bridging information asymmetries and establishing a resilient system of risk mitigation…".[388] Unlike much of the extant law discussed in Part II, these recent European initiatives focus not just on data as an object but on data infrastructures themselves. For example, the proposed DMA requires transparency about certain processes of datafication.[389] Yet, these initiatives also illustrate the limits of transparency in regulating infrastructural control given that data infrastructures are notoriously complex and contextual.[390] Understanding how the power to datafy is exercised (or how a platform can be deployed to generate more data) requires access not just to the existing data but also to the processes – technical, organizational and social – through which decisions about data generation are made. Corporations frequently commission ethnographers to understand how their products and services are used.[391] Regulators might similarly consider deploying ethnographies as part of the auditing processes.

D. Pooling and Differentiating Access to Data

In the preceding sections, we focused mainly on regulatory state-level interventions and the role of intergovernmental organizations. In this section, we consider the extent to which different local practices and governance arrangements may have potential in fostering sustainable digital development from the bottom and in reallocating the power to decide, in a more participatory fashion, what data is generated, for what purpose, and on what terms. We consider development of data infrastructures more attuned to local contexts and governed by local communities. The nascent models we consider in this section illustrate that locality is not delineated by geographical location; nor do they represent forms of alternative or "artisanal" practices.[392] Instead, the locality represented here is relational and relative. In some instances, "local" practices are positioned in relation to national or global contexts; in others they emphasize a particular community and are "localized" in terms of

---

[388] DMA proposal, recital 65 provides that "the Commission should have access to any relevant documents, data, database, algorithm and information necessary to open and conduct investigations and to monitor the compliance with the obligations laid down in this Regulation, irrespective of who possesses the documents, data or information in question, and regardless of their form or format, their storage medium, or the place where they are stored." The DSA requires very large online platform companies to provide the supervising authority access to data that "necessary to assess the risks and possible harms brought about by the platform's systems, data on the accuracy, functioning and testing of algorithmic systems for content moderation, recommender systems or advertising systems, or data on processes and outputs of content moderation or of internal complaint-handling systems…" DSA, recital 65.

[389] For example, the proposed DMA requires gatekeepers to provide *at least* a description of the basis on which gatekeepers create user profiles, including, whether personal data and data derived from user activity is relied on, the processing applied, the purpose for which the profile is prepared and eventually used, the impact of such profiling on the gatekeeper's services, and the steps taken to enable end users to be aware of the relevant use of such profiling, as well as to seek their consent. (emphasis added).

[390] *See above* Section I.C.

[391] Leslie Brockow, *Ethnography in Action at Wells Fargo*, MIT SLOAN MANAGEMENT REVIEW (March 30, 2014), ;https://sloanreview.mit.edu/article/ethnography-in-action-at-wells-fargo/; Michael Fitzgerald, *Corporate Ethnography*, MIT TECHNOLOGY REVIEW, (November 17, 2005), https://www.technologyreview.com/2005/11/17/230047/corporate-ethnography/.

[392] On the different meanings off "local" in the context of knowledge and data, *see* Yanni A. Loukissas, ALL DATA ARE LOCAL: THINKING CRITICALLY IN A DATA-DRIVEN SOCIETY ch. 1 (2019).

common interests or goals. Local ownership, sourcing, or practices are not an end in themselves as local practices can be exclusionary or even oppressive as well.

We draw inspiration from different emerging models of collective governance over data and data infrastructures, often billed as "data cooperatives", "data collectives", "data commons" or "data trusts".[393] The emergence of such initiatives has been sporadic and their success is difficult to assess, in part because some are hyper-local, in part because they are very recent, and in-part because long-term sustainability and outcomes of such initiatives is difficult to predict. The legal environment in which these initiatives operate may be changing as well. The proposed European Data Governance Act tries to create more favorable conditions for data cooperatives and other data sharing intermediaries.[394]

Although emerging initiatives often share similar labels, they vary in form, scope and governance type.[395] Some aim to incentivize pooling of data for public purposes, while giving data contributors various degrees of control and choices about how their data is used. For example, Salus, a non-profit citizen data cooperative for health research founded by members of the public in Barcelona in 2018, designed a license that allows data to be donated for research purposes under set conditions.[396] Other initiatives offer individuals opportunities to monetize "their" data.[397] For example, PolyPoly – an open source EU-based shareholder-owned data cooperative – promises to its owners to be able to track and store their online data in a polyPod, which automatically collects and sorts data it finds about the owner on the internet, which can then be monetized.[398] Similarly, Driver's Seat Cooperative aims to give ride-sharing drivers an opportunity to monetize their driving data through sales of insights to city agencies so they can make better transportation planning decisions. Proceeds from sales are shared among the driver-owners via dividends.[399]

Another set of cooperatives and other community-based arrangements have arisen specifically to counter infrastructural control of large commercial cloud providers.[400] For some initiatives,

---

[393] *See also* Ada Lovelace Institute and UK AI Council, EXPLORING LEGAL MECHANISMS FOR DATA STEWARDSHIP (March 2021), https://www.adalovelaceinstitute.org/project/legal-mechanisms-for-data-stewardship-working-group/; Bianca Wiley and Sean McDonald, What Is a Data Trust?, CIGI ONLINE (Oct. 9, 2018), https://www.cigionline.org/articles/what-data-trust.

[394] Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act) COM/2020/767 final (2020). *See* Sean McDonald, *A Novel, European Act of Data Governance*, CIGI ONLINE (Dec. 15, 2020), https://www.cigionline.org/articles/novel-european-act-data-governance.

[395] On the idea of data cooperatives, *see* Alex Pentland, Alexander Lipton & Thomas Hardjono eds., BUILDING THE NEW ECONOMY (2020), Part I: The Human Perspective: New Types of Engagement.

[396] Data must be for use in health research, to be used by non-profit institutions that openly share the results of their research while anonymizing the data at the highest possible level. Use is allowed until data donors withdraw their permission. SALUS COOP, https://perma.cc/7LWL-ZMBN. In 2020, Salus created the Cooperative COVID Cohort project (CO3) to create a cohort of citizen data donors for research on COVID-19.

[397] For normative concerns about the desirability of monetization of personal data, *see above* Section II.B.

[398] Any EU citizen over 18 years of age can buy a share of PolyPoly and become a co-owner. Profits appear to be generated through the for-profit company PolyPoly, which provides data services to enterprises. https://perma.cc/R99J-BHRY.

[399] *See* the self-description at https://perma.cc/4F5B-QA5M.

[400] For example, CoBox offers distributed, encrypted, offline-enabled data hosting cloud platform. Its stated aim is "to facilitate a transition away from giant data centers, huge storm clouds, towards a vision of cloud infrastructure that is light, distributed, and importantly, is offline-first. … CoBox is the beginning of a sovereign commons-based data

collective governance is a core feature. For example, users of CommonsCloud can become consumer partners of femProcomuns and participate in the governance of the cooperative and the CommonsCloud.[401] Similarly, Framasoft, a nonprofit network of projects headquartered in France, which hosts a project to "de-googlify" the internet by offering free alternative services, has an elaborate governance framework that resembles those used in open-source software communities.[402]

One of the first and the best known collective data sharing arrangements that combines data pooling with collective governance is Project DECODE, which is connected to broader efforts to reimagine democratic governance in digitally mediated cities.[403] Project DECODE aims to give individuals control over how and on what terms data that is generated and gathered by apps, interconnected devices, and sensor networks in cities can be made available for broader communal use, with appropriate privacy protections.[404] A pilot in Barcelona, for example, addressed community concerns that data from environmental sensors placed throughout the neighborhood (which recorded noise levels, pollution, temperature, humidity, etc.) might reveal sensitive information or be misused otherwise. In response, residents used DECODE technology (a combination of blockchain and attribute-based cryptography) to share encrypted data anonymously within their community. DECODE's pilot project provided training and support to the individuals on how data can be gathered, analyzed and used to improve city services. Project DECODE envisions that a common data infrastructure will be open to local companies, coops, and other organizations to build data-driven services and products, thus creating public value. In Barcelona, for example, the data shared by citizens in the DECODE pilots "integrates with the Barcelona City Hall digital infrastructures: the data lake CityOS, the IoT open sensor network Sentilo, Barcelona open data portal and the digital democracy platform Decidim."[405] Such integration was possible only after the Barcelona City Council released a new Digital City Plan, with an ethical data strategy. Barcelona also revised procurement deals between city hall and its private sector providers and included 'data sovereignty' clauses in public procurement contracts, requiring suppliers that works for the city of Barcelona to provide the data they gather to deliver services to the City in machine readable format, thus enabling the release of such data as open data so as to allow communities to benefit from it as well.[406]

---

infrastructure and a co-operative distributed cloud architecture." COBOX, https://perma.cc/MZ75-M6P8. Commonscloud.coop, project of a Catalan community, is specifically billed as an alternative to corporate storage clouds such as Google Drive and Dropbox. Commonscloud provides free software, as well as a platform that facilitates community conversation around services provided by the project.

[401] FemProcomuns is a "cooperative of work and consumption, non-profit and social initiative created in Catalonia in 2017, with the aim of consolidating a common ecosystem, based on the principles of open cooperativism, community self-management, the ecological, economic and human sustainability, shared knowledge and replicability." It provides organizational, operational and governance functions to cooperatives. FEMPROCOMUNS, https://perma.cc/YUP2-PASL.

[402] A list of the offered services is available here: FrameSoft, *De-google-ify Internet*, https://perma.cc/GR3Y-4FRE. The governance structure is explained here: https://perma.cc/5B9Y-PX3Y.

[403] *See* Evgeny Morozov and Francesca Bria, RETHINKING THE SMART CITY: DEMOCRATIZING URBAN TECHNOLOGY (2018); Bianca Wiley, *Searching for the Smart City's Democratic Future*, CIGI ONLINE (Aug. 13, 2018), https://www.cigionline.org/articles/searching-smart-citys-democratic-future.

[404] Citizens can set the anonymity level via the DECODE app, so that they cannot be identified without explicit consent. In this way they can keep control over data once they share it for the communal purposes.

[405] *Common Knowledge: Citizen-led data governance for better cities*, DECODE (Jan. 2020), https://decodeproject.eu/deliverables.

[406] For a discussion of distributional effects of open data, *see above* Section II.B.

Whether Project DECODE and other similar initiatives succeed or not in generating enough uptake and participation to achieve their aims remains to be seen.[407] Nonetheless, these examples offer an alternative pathway to development of data infrastructures. One can imagine, for example, that a series of data collectives might link up along sectoral lines (e.g., to expand sharing of health or transportation data beyond immediate community), common values (e.g., promotion of open, free and decentralized infrastructure), geographic proximities (e.g., cities in Europe) and/or along other dimensions. Communities could band together with other similarly predisposed publics to develop the necessary technologies, to exchange experiences and ideas, and to develop commensurate public governance mechanisms attuned to their specific conditions.[408]

Many of the initiatives aim to create new data infrastructures not only to enable pooling of and control over data generation but also to release pooled data to the public as "open data".[409] Making certain data sets publicly available can have substantial public benefit. For example, having access to different datasets on COVID-19 infections, hospitalizations and deaths enabled different platforms to compile real-time COVID-19 profiles for countries (e.g., dashboard created by the John Hopkins University[410]) and unveiled opportunities for scientific examination and collaboration.[411] At the same time, open data, at least as practiced so far, may not be an effective tool to address data inequality.[412] Critically, thus far, calls to "open" data have been mainly directed at governments. Rarely have privately held data been subjected to similar calls for openness. Calls for open public data – including in "digital trade" agreements – are often being propelled by narratives according to which they provide opportunities for small and medium businesses to engage in the digital economy. Yet, the extent to which having access to open governmental data increases the opportunity of smaller enterprises to compete with corporations that control expansive data infrastructures is unclear and empirically unproven. Those who control the means of data generation on large scales gain access to the same "open" data, once made available by governments, as everyone else. Given that one value proposition of datafication lies in producing insights from data, which often requires aggregation with other data sets, open data tends to privilege tech savvy users, who not only have the capacity and resources to integrate and analyze different datasets but who also have access to other datasets (open and closed), which enhances the re-usability value of open data to them. Thus, perhaps counterintuitively, open data can further empower those who already have enhanced access to data, including large data companies like Google and Alibaba, thus failing to correct the asymmetry in data distribution.

Open data also has the potential to exacerbate the inequality of the power to datafy. Decisions about which data is made open and under what conditions are usually made by the data holders and the

---

[407] Public support of the kind illustrated by the Barcelona case is critical to uptake and sustainability of local infrastructures. Without it, many worthwhile initiatives will likely be abandoned due to limited acceptance. *See, e.g.*, THEGOODDATA, https://perma.cc/V532-U8L.

[408] Indeed, one could envisage data being the catalyst for new types of transnational movements that coalesce around common interests (e.g., labor exploitation).

[409] Although there are multiple meanings of "open", as used in this context, "open data" refers to data made publicly available under an open data license or data that is released into public domain. This is legally complicated, because data – unlike software – does not acquire copyright protection as easily. *See above* Section II.B.

[410] *COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)*, JOHN HOPKINS UNIVERSITY, https://perma.cc/C6LP-86HS.

[411] Junaid Shuja et al., *COVID-19 open source data sets: a comprehensive survey*, APPLIED INTELLIGENCE, 1 (Sep. 2020).

[412] *See also above* Section II.B.

processes according to which these determinations are being made (and with what considerations and motivations in mind) are often not disclosed. By the time data is opened up, decisions about how the phenomenon it purports to represent was defined, measured, collected, processed, etc. have already been made, as have choices about what features of the phenomenon *not* to datafy or represent. These choices – and all the biases they contain – become reproduced and entrenched as data becomes released for public use.

These effects could, of course, be mitigated through regulatory design and governance arrangements, for example by fostering participatory multi-stakeholder processes for data collection. The Open Government Partnership purports to take this approach.[413] If open data is not confined to output data and metadata but encompasses a full disclosure of methods and sources through which data was selected and processed, crucial context is being provided, which may then enable productive contestation.[414] Similarly, data can be made more useable and legible to different audiences through choices of format in which it is made available (e.g., machine readable, narrative-style or other formats for low-tech engagement, etc.) and by simplifying the means through which it can be accessed (e.g., download, APIs, etc.), thereby reducing the need to rely on platforms as gateways to open data. Open data is often first consumed and subsequently made legible by intermediaries, which might include librarians, journalists, and nonprofits. [415] However, efforts (and funds) directed at making data open rarely also are directed at supporting and sustaining these intermediaries who lend crucial support to enable wider use of open data by less resourced constituencies.

Still, even with these mitigation strategies, from the perspective of data as an economic resource, there is little evidence that unconditional opening of public data in and of itself can address unequal control over data or redistribute opportunities for monetization. We highlight these challenges not to disparage the value of open data but to encourage a more nuanced and differentiated consideration of how, to what, and under what conditions access to data is being provided. If one seeks to confront this form of data inequality, one cannot assume that interests of all stakeholders are necessarily aligned (though they sometimes are) and that all stakeholders ought to be treated in the same way. For this reason, developing differentiated and conditional "data sharing" infrastructures that are attuned to data inequality might be a worthwhile pursuit, for smaller markets and economies. In this context, we suggest building on Lisa Austin's and Davie Lie's idea of "Safe Sharing Sites" as infrastructural and legal assemblages that make data more regulatable.[416] While their focus in on facilitating data sharing while protecting the data privacy and security of individual data subjects, their approach can be expanded to design data sharing infrastructures, including for non-personal or aggregated data, that cater to additional and different concerns, including those around data inequality. This is not to say that this solution is ready-made or straight-forward. The key point is that addressing data inequality through "data sharing" requires careful consideration of infrastructural and legal elements and that the "safe sharing site" idea encourages thinking along these dimensions. In this regard, we depart from much of the "open data" discourse which tends to be focused on the licensing terms under which

---

[413] OPEN GOVERNMENT PARTNERSHIP, https://www.opengovpartnership.org/process/.
[414] *See* for an example of this approach, the WomanStats project: https://perma.cc/SD45-NYGE.
[415] Ricardo Ramírez, Balaji Parthasarathy & Andrew Gordon, *From Infomediaries to Infomediation at Public Access Venues: Lessons from a 3-Country Study*, PROCEEDINGS OF THE SIXTH INTERNATIONAL CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGIES AND DEVELOPMENT: FULL PAPERS, VOL. 1 (2013), 124.
[416] Lisa Austin & David Lie, *Safe Sharing Sites*, 94 NEW YORK UNIVERSITY LAW REVIEW 4 (2019)

data is being made available as well as "data collaboratives" ideas which are often centered on contractual solutions.[417] While these interventions are valuable in themselves, they are unlikely to address data inequality dimensions because they tend to disregard the severe gradients of disproportionate power that stem from asymmetric control over data and data infrastructures. Those who already control data and data infrastructures stand to gain most from "open data" initiatives and can dictate also the terms in "data collaboratives".[418]

In terms of differentiation, the designers of data sharing infrastructures can choose and decide who gets access to data and on what terms. This design power needs to be checked through appropriate governance mechanisms but can be mobilized for public benefit: those who commit to using data for non-commercial purposes, for example, could be granted access to data for free, while those with commercial use-cases could pay a fee (which could be dependent on business size or commensurate with eventual profit). Logically, onward data sharing outside the "safe sharing site" would need to be constrained and policed to retain regulatory control over data sharing. Importantly, the safe sharing site would not make data available wholesale but attuned to the relevant use case. Comparable solutions are already in place for differential access to data in the context of machine learning competitions: contestants can test the quality of their models without gaining access to the actual testing data. Ideas to share the insights from machine learning (rather than the underlying data) point into a similar direction.[419]

In terms of conditionalities, access to the public data sharing infrastructure could be conditioned on a range of regulatory demands, including adherence to data protection and privacy policies, commitment to respect and foster spaces for deliberation over and contestation of proposed data uses, payment of fees to fund the infrastructure, or even requiring tax residency.[420] Naturally, such conditionalities will only be effective if the expected benefit that can be derived from accessing the data outweighs the costs that these conditionalities impose. [421]

Public data sharing infrastructures can involve and benefit a range of public and private actors but control over the legal and infrastructural design of such platforms is crucial. For this reason, it is of paramount importance to develop governance mechanisms that are resistant to capture and adhere to fundamental administrative principles of transparency, participation, reason-giving, and review to ensure accountability. Developing publicly supported data sharing infrastructures with commensurate governance mechanisms strikes us as the most promising short-term intervention to address data inequality.

---

[417] *See, e.g.*, the Contracts for Data Collaboration (C4DC) project, www.contractsfordatacollaboration.org.

[418] *See* Jonathan Gray, *Towards a Genealogy of Open Data* (General Conference of the European Consortium for Political Research in Glasgow, Working Paper, Sept. 3-6, 2014), https://dx.doi.org/10.2139/ssrn.2605828.

[419] *See* for such ideas Michal Gal & Nicolas Petit, *Radical Restorative Remedies for Digital Markets*, 37 BERKELEY TECHNOLOGY LAW JOURNAL (forthcoming 2021), https://ssrn.com/abstract=3687604 (Sep. 6, 2020).

[420] Denmark blocked firms registered in tax-havens from receiving state aid during the COVID pandemic. *See* Nikolaj Skydsgaard, *Denmark blocks firms registered in tax-heavens from state aid*, REUTERS, April 20, 2020.

[421] The EU is currently navigating these trade-offs as it seeks to construct several sectoral data pools to facilitate data sharing within Europe. *See* European Commission, *A European Strategy for Data*, COM(2020) 66 final.

E. Developing Collective Data Governance

Throughout this Part we have alluded to collective governance over data as illustrations of initiatives to pool data for common benefit, to empower individuals supplying data, and as possible venues for leveraging collective bargaining power. Although often building off Elinor Ostrom's work on knowledge commons or drawing on the governance models of open source software communities, collective governance over data, and in particular over data infrastructures, remains understudied and undertheorized. An in-depth examination of such arrangements is beyond the scope of this paper. However, a few observations are worth making.

Governance over data and data infrastructures is complicated by the fact that both data and infrastructures are relational concepts.[422] As Salomé Viljoen demonstrates, data's capacity to transmit social and relational meaning is not only central to production of economic value from data, but also "renders data production especially capable of benefitting and harming others beyond the data subject from whom data is collected."[423] This explains the important limitations of extant law, which, with some exceptions, views governance of data through the lens of individual rights (e.g., property or fundamental rights).[424] Viljoen argues that, instead, the aim of data governance should be "to develop the institutional responses necessary to represent the relevant population-level interests at stake in data production …securing recognition and standing to shape the purposes and conditions of data production for those with interests at stake in such choices, and thus establish the terms of legitimate mutual obligation."[425] This approach, she posits, would also provide foundation for mandatory data collection, "as long as the purposes and the conditions of such collection are derived from legitimate forms of collective self-willing and further legitimate public ends."[426]

Although we are sympathetic to Viljoen's argument, its implementation is complicated by the relational nature of infrastructures. Like data, infrastructures both shape and are shaped by relations. Data infrastructures are products of organizational dynamics of corporate forms, their components are built and linked up through technical and legal (often contractual) relationships between designers and manufacturers of devices, software engineers, logistics personnel, and users.[427] Data infrastructures thus implicate different publics, whose interests will not be always aligned.[428] The transnationality of data infrastructures also means that the relevant publics are dispersed, often unaware of, and not exposed to, each other's existence. This might explain why the emergence of data

---

[422] Salome Viljoen, *Democratic Data: A Relational Theory for Data Governance*, YALE LAW JOURNAL (forthcoming 2021), https://ssrn.com/abstract=3727562 (Nov. 23, 2020); Benedict Kingsbury and Nahuel Maisley, *Infrastructures and Laws: Publics and Publicness*, 17 ANNUAL REVIEW OF LAW AND SOCIAL SCIENCES (forthcoming 2021), draft on file with authors. *See also above* Section I.A.

[423] Salome Viljoen, *Democratic Data: A Relational Theory for Data Governance*, YALE LAW JOURNAL (forthcoming 2021), https://ssrn.com/abstract=3727562 (Nov. 23, 2020).

[424] Salomé Viljoen, *Data as Property?* PHENOMENAL WORLD (Oct. 16, 2020), https://perma.cc/PQN3-NF6E. *See also above* Sections II.B–C.

[425] *See* Salome Viljoen, *Democratic Data: A Relational Theory for Data Governance*, YALE LAW JOURNAL (forthcoming 2021), https://ssrn.com/abstract=3727562 (Nov. 23, 2020).

[426] *See* Salome Viljoen, *Democratic Data: A Relational Theory for Data Governance*, YALE LAW JOURNAL (forthcoming 2021), https://ssrn.com/abstract=3727562 (Nov. 23, 2020).

[427] *See above* Section I.C.

[428] Benedict Kingsbury and Nahuel Maisley, *Infrastructures and Laws: Publics and Publicness*, 17 ANNUAL REVIEW OF LAW AND SOCIAL SCIENCES (forthcoming 2021), draft on file with authors.

collectives thus far has been very local, with relatively clearly defined publics. Layered on top are gendered, racial, cultural, and other socio-economic dimensions that can silence participation and exclude certain publics.[429] Legal regimes and institutions can both facilitate and foreclose connections between and among publics.[430]

Despite the challenges, exploring ways in which individuals, communities and other groupings of constituencies can be empowered to participate in datafication decisions as well as in design, management and oversight of data infrastructures is a worthwhile endeavor. Deploying technological means to create spaces for public deliberation, creating legal spaces for political and social organizing and engaging the public in auditing process of companies exercising significant infrastructural control over data are just some of the steps that might be taken towards remedying data inequality.

[429] This is an increasingly recognized problem in open-source software communities. *See, e.g.*, Caroline Sinders, *Designing for Community Health and Safe Spaces: The History of the JS Confs and Fighting Harassment to Maintain Healthy Open Source Communities*, CONVOCATION DESIGN + RESEARCH, FORD FOUNDATION (Nov. 23, 2020), https://www.fordfoundation.org/media/5805/ford_report_final-1.pdf.

[430] Ride sharing companies Uber and Ola are opposing drivers' requests for access to data companies have about them for purposes of developing a trade union data trust on the grounds that doing so would violate data protection rights of customers. The case is being currently litigated in the Netherlands. App Drivers and Couriers Union, *Uber and Ola Cabs in legal bid to curtail worker digital rights and suppress union organised data trusts*, ACDU (Dec. 16, 2020), https://perma.cc/R8NM-G747.

Concluding Observations

Data is more than an economic resource. Data is a medium through which economic, social, and political life is increasingly being ordered and re-ordered. Sufficient and sustainable access to data is unattainable for developing economies that lack the prerequisite data infrastructures necessary to generate, store, and process data on their own terms. Individuals, communities, and societies are being deprived of capabilities and possibilities to chart their own digital destinies when those that control the means of data production also control the ability to define, classify, shape, make visible or render unseen identities and environments, labor and leisure, conflicts and solidarities, freedoms and oppression.

Faced with this scenario, the intuition of lawyers and regulators is often to search for an appropriate legal intervention, usually guided by existing legal and regulatory frameworks and institutions. But law is not exogenous; it does not simply act (or not) *on* data. Law co-constitutes, shapes, enables, and is symbiotically intertwined with data infrastructures. Extant legal paradigms and institutions may target discrete issues more or less effectively, but broader impacts of datafication, including on individual welfare, development freedom, and democratic governance are too rarely being considered. When law and lawyers overlook where control over data and its constituting infrastructures is being exercised and where and why it is being entrenched, successful contestation of outsized power to datafy becomes elusive, thereby exacerbating data inequality. As we have argued throughout this paper, effective interventions need to be attuned not only to legal but also to infrastructural dimensions, including the politics of data infrastructures, and may need to creatively explore new regimes and institutions.

There is neither a single nor an easy "fix" to the problem of data inequality. Indeed, revealing and unraveling the root causes of data inequality may take time. It may be worth pausing the pace of "leapfrogging" to (re)evaluate the degree to which contemporary patterns of datafication and "free flow" of data resemble patterns of colonial extractivism.[431] Such (re)evaluation may lead to re-imaging of existing legal domains and re-aligning of legal frameworks to better account for economics and politics of data infrastructures. Still, law – even if reimagined – will not be a silver bullet for remedying data inequality. Data, from the moment of its conception to its exploitation, is deeply political. Interventions to remedy data inequality thus must be situated within continuous, iterative, inclusive, and public debate. Development organizations should encourage and create spaces that foster opportunities for communities to reclaim collective governance over data. Here, a reinvigorated conception of human rights might be brought to bear by focusing not predominantly on invocation of individual rights but by building transnational movements across different publics and territorial boundaries to enable and support decisive and forceful action that can confront and overcome countervailing interests.[432]

---

[431] Nick Couldry and Ulises A. Mejias, THE COSTS OF CONNECTION: HOW DATA IS COLONIZING HUMAN LIFE AND APPROPRIATING IT FOR CAPITALISM (2019).

[432] On the processes through which human rights ideas and practices developed in cosmopolitan centers are being translated into terms for local contexts *see* Sally Engle Merry & Peggy Levitt, *The Vernacularization of Women's Human Rights*, in HUMAN RIGHTS FUTURES 213–236 (Stephen Hopgood, Jack Snyder & Leslie Vinjamuri eds., 2017). On the transformational impacts of human rights movements, *see* Gráinne de Búrca, REFRAMING HUMAN RIGHTS IN A TURBULENT ERA (2021).

States – and their publics – must be able to experiment with digital development policies without being overly constrained by international economic law. Naturally, such experiments will not always succeed. But the moment to chart new and alternative pathways is now as data infrastructures are being built at fast pace and on massive scales.