THE POLITICAL ECONOMY OF DATA IN A NETWORKED,平台-BASED ENVIRONMENT

REPORT FROM A DECEMBER 2018 WORKSHOP HOSTED BY GEORGETOWN LAW’S INSTITUTE FOR TECH LAW & POLICY AND CORNELL TECH’S DIGITAL LIFE INITIATIVE

Aaron Fluit, Fellow
Alexandra Reeve Givens, Executive Director
Georgetown Institute for Tech Law & Policy
# TABLE OF CONTENTS

I. INTRODUCTION

II. MAPPING THE TERRAIN
   A. Prior Approaches to Mapping the Data Economy .......................................................... 3
   B. Data Metaphors .................................................................................................................. 5
   C. Data Taxonomies .............................................................................................................. 9
   D. Data Tensions ................................................................................................................... 12
   E. Mapping The Problem Space: Access & Control Goals For The Data Economy ..................... 18

III. EXPLORING THE SOLUTION SPACE
   A. Regulatory Challenges .................................................................................................... 20
   B. Transparency & Accountability ......................................................................................... 21
   C. Restrictions on Data Collection, Processing, & Storage .................................................... 23
   D. Competition Law ............................................................................................................. 24
   E. Labor Law: Personal Data Markets .................................................................................... 27
   F. Intellectual Property Law .................................................................................................. 28
   G. Contract Law ................................................................................................................... 31
   H. Tax Law .......................................................................................................................... 32

IV. QUESTIONS & PATHS FOR FURTHER EXPLORATION
   A. Is Data Privacy a Lost Cause? ........................................................................................ 33
   B. What Can Transparency Actually Accomplish, and is it Worth it? ..................................... 33
   C. Data Portability: What Does it Mean and What Should it Mean? ....................................... 33
   D. Data Philanthropy: What Does it Mean and is it worth working toward? .......................... 34

V. CONCLUSION

VI. RECOMMENDED READINGS

VII. LIST OF PARTICIPANTS

VIII. ABOUT THE CONVENORS
This report summarizes the presentations and reflections of a group convened by Professors Julie Cohen and Helen Nissenbaum in December 2018 to consider regulatory and governance issues related to data holdings and access to data. Convened in partnership with Georgetown Law’s Institute for Technology Law & Policy and Cornell Tech’s Digital Life Initiative, the workshop brought together scholars and practitioners with wide-ranging expertise in data use and governance, data privacy and protection, competition, and intellectual property law. The workshop was sponsored by Microsoft.

Conversations about the promises and perils of data permeate every facet of society. From industry to academia, from the halls of government to kitchen tables, attention is focused on data as never before. Recent advances in machine learning and other forms of artificial intelligence have spurred a great data rush. Researchers, retailers, advertisers and political campaigns all seek massive, rich, datasets. Data has manifested as an invaluable resource for private industry, for public interest organizations, for governments at every level, and for individuals.

It is increasingly clear that these diverse groups often have competing interests regarding access to, and control over, the multitude of increasingly rich (high-dimensional) datasets. Individuals are often disturbed to discover the extent to which data about them is collected, processed, and distributed. In many cases, a few large corporations assert control over proprietary datasets, and researchers often argue that the concentration of data within a few large firms, and restricted access to these datasets, hinder scientific and social progress. Governments and businesses increasingly rely upon big data for automated decisionmaking, the results of which have been criticized for perpetuating unfairness and social inequality. Individuals, governments, public interest groups, and businesses speak in broad terms about the need for greater accountability over data use, but there remains considerable uncertainty and disagreement about how to achieve these goals, and even about what accountability and control mean in different contexts.

Complicating these issues further, we lack a unified legal vocabulary to describe the kinds of interests that individuals and entities can assert in data -- and how data is collected, used and held. We refer to “the user’s data”, or “a company’s datasets”, but extant regimes of intellectual property protections were not designed with data and/or algorithms in mind, and do not effectively map to the interests being asserted. Access to data by various third parties seeking competitive entry, data portability, and/or the ability to provide various public goods is an important public goal, yet we lack the legal vocabulary to describe the interests asserted by those parties as well.

---

1 See, e.g., Report from the Georgetown Law Round Table on the Ethical Reuse of Data in a Machine Learning World, Institute for Technology Law & Policy, Georgetown University Law Center at 15-16 (Oct. 27, 2017).

The workshop and this report aim to spark a substantive conversation about the appropriate legal frameworks used to describe core relationships and interests in today’s data economy. The discussion proceeded in four sessions:

1. Mapping the Terrain: Understanding and Mapping the Data Economy
2. Competition and Data Labor: Issues and Possible Elements of the Solution Space
3. Intellectual Property: Issues and Possible Elements of the Solution Space
4. Testability and Transparency: Issues and Possible Elements of the Solution Space
II. MAPPING THE TERRAIN

Data sit at the core of the emerging information economy. However, compared with what is known about other major components of the market economy (e.g., land, industrial production, securities markets, patent and copyright records), the data economy is still coming into view—and certain parts of it are relatively invisible.

Many of the challenges in mapping the data economy arise from its distinct nature: the “market” encompasses every person on earth, in almost every transaction they engage in, with almost every entity they engage with, commercial or otherwise. Data can be created and transferred without the data subject even knowing that data about them has been generated. Data is, of course, intangible and inexhaustible as a resource, capable of transfer an infinite number of times.

Opening the workshop session, one discussant noted that the challenges of mapping the data economy arise also because of the choices society makes (or has not made) in how we account for data assets. Intellectual property rights and securities are also intangible, but they are bound by public registration requirements and regulatory reporting obligations that make them easily inventoried, with the market visible to all. In contrast, our legal structures do not conceive of data in such a boundaried way. Fuzziness in the legal constructs that are applied to data thus exacerbates already deeply challenging obstacles in understanding how and where data is used, transacted and held.

These threshold challenges of even seeing the data economy—understanding its scope, its participants, and the nature of its transactions--greatly complicate efforts to develop appropriate analytical or legal frameworks. Importantly, lack of visibility within the data economy also stymies future efforts at accountability. For instance, data visibility is a prerequisite for understanding what data companies hold when evaluating a merger in the context of competition law. Visibility is critical for any effort to assert personal data rights: individuals can hardly challenge the collection of data about them if they do not even know their data has been collected. Visibility is also important for thinking about the value transferred in “data philanthropy”, when companies allow datasets they hold to be analyzed for public good. Visibility is, of course, essential for broader transparency objectives, for unlocking insights about what data sets exist and how they can be used. Broadly speaking, data visibility is needed to assess and regulate balances of power, between individuals, businesses, and governments.

Given this backdrop, the workshop began by questioning how policymakers and academics can approach locating relevant data assets, and what exactly is needed to see and define the data economy.

A. PRIOR APPROACHES TO MAPPING THE DATA ECONOMY

Workshop participants began by describing prior efforts at mapping the data economy that had fallen short on this challenging goal. One participant discussed how, several years ago, the European
Commission tasked a research team with developing a mathematical model of the data economy. The team struggled for three years without success, and ultimately concluded that developing such a precise model would prove impossible. Opting instead for a descriptive approach, the researchers visualized the network of relationships involved in the data economy, identifying four broad categories: (1) agents and agent populations; (2) data and technologies artefacts; (3) goal-oriented strategies; and (4) environmental factors or framework conditions.

The final report identified the current set of “agents and agent populations” as including data holders, solution providers, data distributors, data users, and enablers, and included spatial and conceptual locations, capabilities, and memory as relevant properties of agents. The Report described artefacts as “the objects used by agents,” noting that artefacts “are used by agents to support or facilitate responses to or interactions with other agents or the environment.” It distinguished six dimensions of data artefacts: the source, content, format, terms of use, purpose, and position in the data value chain. Strategies “describe how agents react to their surroundings and pursue their goals;” and “include deliberate choice but also patterns of response that pursue goals with little or no deliberation.” Examples of strategies included those of regulatory bodies and policymakers, and strategies of data acquisition, manipulation, exploitation, and consultation, as well as changes in strategy resulting from changes in agent populations. With respect to environmental factors or framework conditions, the Report put “particular emphasis on legal, socio-economic, and technological factors.”

The workshop participant noted two important lessons from this effort: first, the importance of managing expectations with regard to mapping the data economy, given its challenges. Second, she noted that it may not be necessary to always see the “whole forest”, but that being able to see and analyze a few trees can be helpful in understanding and discussing the data economy. In particular, the European team found “data reuse” to be a useful area of focus, and trained much of its analysis on the reuse of data for purposes other than for which it was originally collected. She noted that the European Union’s General Data Protection Regulation (GDPR)’s provisions on data reuse now expressly acknowledge, and indeed are based on, the recognized challenges in identifying the purpose for which data will be used at the time of its collection.

---

4 Id. The Report noted that artefacts “are used by agents to support or facilitate responses to or interactions with other agents or the environment.”
5 Id. at 14.
6 E.g., Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation), 2016 O.J. (L 119/6) Recital 33 (“It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed
Another participant also described an effort to map the information economy within the United Kingdom. That study focused its analysis exclusively on the production side of the digital economy, ignoring “[data as] inputs, for the simple reason that it is now hard to think of any economic activity where digital inputs do not feature, and given the pace of change in internet tools and platforms, definition and measurement problems for digital inputs are severe.” The U.K. paper cited a “clear” statement from the U.K. Government about the challenges of mapping the data economy: “Addressing the lack of clear and universally-agreed metrics will be an early priority for Government and industry. There will be a need for continual reassessment of the scope and definition of the information economy as it evolves.”

B. DATA METAPHORS

Society uses a mix of metaphors to help conceptualize the data economy. One of the session’s discussants presented several of these options, noting that while none fits perfectly, each metaphor points to a different suite of potential policy interventions, and can be useful as a framing tool.

1. Data as the New Oil

Data is often referred to as “the new oil.” Like oil, it can have tremendous value, drives much of the modern economy, and is the subject of a rush. The speaker argued that, like businesses built on oil, businesses built on big data tend toward monopolies, due in part to the fact that the resources required to harvest meaningful amounts of data and oil create significant barriers to entry. Data firms are replacing energy firms as the most valuable in the world. Finally, like oil, big data carries with it a number of negative externalities; in particular, there are concerns about privacy and the discriminatory effects of decision-making based on big data.

On the other hand, unlike oil, data is not always valuable as a raw commodity or raw material: a distinction must be drawn between structured data that is useful for analytics, and unstructured data that may be much less useful. Neither is data always or necessarily rivalrous: in fact, data can be considered a non-rivalrous public good, at least in certain contexts. Another difference between

to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research.”).


9 Although one could argue that raw crude oil and unstructured data share similarities in that both need to be refined in order to turn them into valuable commodities.
oil and data is that the producers of oil are extinct, whereas the producers of data are very much alive and with us today, and they have agency.

Exploring the value of this metaphor, the speaker noted that analogies to oil evoke at least three potential policy interventions. Because the power of big data is concentrated within a few large firms, one answer might be to break up those companies. Another is to consider an idea like “renewables” — publicly owned and curated data, or to create data cooperatives like we have energy cooperatives, to increase the power of data subjects and smaller consumers.

2. Data as Pollution

Similar to the recognition of the negative externalities of data under the metaphor of data as “the new oil,” the speaker noted that another metaphor gaining traction is data as pollution. In a world where companies are storing information they do not even need on a long term basis (such as Marriott’s long-term storage of customers’ passport data, revealed during a massive data breach), companies’ large data pools can be viewed as a liability, and a public hazard to be contained.

If amassing unnecessary pools of sensitive data were likened to environmental pollution, the range of potential policy interventions could include interventions similar to those found in environmental protection laws, laws regulating superfund sites, and also tort laws of public and private nuisance. The speaker noted that reframing our perception of data pools from a source of value into a source of liability might have helpful incentives for corporate behavior.

3. Data as the New Tobacco or Data as Food

Data has also been characterized as “the new tobacco,” in an attempt to capture some of its more harmful effects. Like tobacco sales and marketing, personal data combined with knowledge of human psychologies can be used to exploit human frailties, especially in the context of data access and control. Unlike tobacco, data is not always unhealthy. This distinction has led some to propose thinking of data not as the new tobacco, but as food. Like food, data can be addictive and unhealthy, but it is also necessary, useful, and sustaining.

---

The speaker noted that characterizations of data as tobacco or as food point to several potential policy interventions: mandatory warning labels, public education and awareness campaigns, and the banning of certain unsafe or unhealthy practices concerning the handling and use of data.

4. Data as a Tsunami

Data has also been likened to a tsunami: data as water that can escape control, overpower, and have large-scale effects. The speaker noted the common vocabulary we use for both water and data: we talk of data leaks, spills, runoff, exhaust, and so on, all capturing a sense that data is challenging to contain. Data sharing agreements or data protection laws that stipulate destruction or return of the data are difficult to enforce once data gets baked into the systems that make use of it: in many ways, requiring the return or destruction of original data is like drawing water back out of a cake.

On the other hand, unlike tsunamis, data is a human construction. “Raw data” is in a sense an oxymoron, because data does not occur naturally, absent human intervention. Additionally, data is not entirely uncontrollable, although sometimes it may appear to be.

Notwithstanding these differences, the metaphor of data as a tsunami points toward some interesting potential policy interventions. Data “seawalls” can be constructed and mandated through cyber security regulations, for instance to prevent data breaches. And because data is similar to a tsunami in that once the data exists and has been published it becomes difficult or impossible to control, it may make much more sense to focus on rules around data collection rather than rules on data usage or data breach.

5. Data as Infrastructure

The metaphor of data as infrastructure is particularly common in the government and public interest contexts, where data is often treated as a common good. Information provided by the U.S. Census Bureau and other government and corporate entities is used by a range of public interest organizations and federal, state, and local government agencies to power significant resource decisions—for example, to determine the most effective and efficient allocations of funding for social welfare programs.

In *Amazon’s Antitrust Paradox*, Lina Khan conveys early accounts from Amazon employees that CEO Jeff Bezos’ “underlying goals were not to build an online bookstore or an online retailer, but rather a ‘utility’ that would become essential to commerce.”¹¹ Today it would be hard to argue that Amazon has not “emerged as central infrastructure for the internet economy.”

Beyond data as an infrastructure-like resource, we can also think of the infrastructure of data: the frameworks and mechanisms for data access and control. In *Who Owns Big Data?*, Michael Nielsen challenges the status quo, which often takes “for granted that a powerful data infrastructure is available only at a few big for-profit companies, and to secretive intelligence agencies such as the NSA and GCHQ.” Nielsen’s paper explores “the possibility of creating a similarly powerful *public* data infrastructure, an infrastructure which could be used by anyone in the world. It would be Big Data for the masses.”

If data were treated as public infrastructure, the potential policy interventions could look similar to those used in response to concerns over concentrations of power in the hands of a few telecommunications and utilities companies, and over discriminatory treatment in other public infrastructure contexts. For instance, such a concept of data as infrastructure might warrant regulations that ensure common carriage and non-discriminatory access, as well as potentially breaking up the largest data firms.

### 6. Data as Property

Data as property may be the most widely used framework through which data access and control are viewed. Workshop participants noted that, even in the absence of de jure laws regulating data as property, companies often assert claims of ownership and intellectual property rights over data in their possession. Individuals think of personal data as “theirs”, especially when that data is exposed during a breach. To be sure, not all data is personal data, and a great deal of non-personal data consists of what might be characterized as “facts about the world,” which the data-as-property metaphor fails to capture. Nevertheless, to the extent that data appears to resemble property, there are a range of potential policy interventions that can be borrowed from traditional property law, including intellectual property law of copyright, patents, trademarks and trade secrets. These policy interventions are discussed in more detail in Section II.F below.

### 7. Data as Labor

Some commentators have noted that generating data often requires a commitment of time and energy, and especially in the context of data provided by individuals to platform providers, the only compensation is the provision of free services. One workshop participant questioned whether, with all of the rich, high-dimensional data that people post online for free, we are becoming *data serfs*. For the major platforms owned by a few large firms such as Google, Facebook, and Twitter, there is a perception that the corporate value extracted from users’ data makes them akin to data lords--

---

especially when the concentration of such services gives users no comparable alternative to which they can turn for a better deal.

Of course, not all data requires time and energy commitments that are characteristic of labor. But treating data as labor, at least in those contexts where appropriate, points to a suite of potential policy interventions that have developed under labor law. These policy interventions are discussed in further detail in Section II.E below.

C. DATA TAXONOMIES

Workshop participants suggested that, in addition to thinking about the contextual relevance of certain metaphors for big data, it is also helpful to consider a taxonomy of data. In addition to the four general data relationship categories that the European Commission’s research team proposed, consisting of (1) agents and agent populations; (2) data and technologies artefacts; (3) goal-oriented strategies; and (4) environmental factors or framework conditions, workshop participants also considered: taxonomies related to data types and sources; a matrix of data visibility and disclosures; and a binary model of data acquisition. These potential taxonomies, and the potential policy interventions and challenges relevant to them, are discussed briefly below.

1. Open Data vs. Closed Data

One familiar dichotomy when discussing data types and sources is the distinction between open and closed data. This dichotomy sometimes uses the terms public and private data, where private (or closed) data has been defined as “any data that are not generally available to the public.”

One workshop participant suggested that the dichotomy between open and closed data may not accurately capture important but subtle distinctions, as there are some datasets that are more accurately described as “semi-open data.” This inability to capture accurately the full range of data sources through the traditional binary of open vs. closed data is becoming especially relevant in the context of smart cities and the nascent algorithmic accountability movement, where efforts to achieve meaningful transparency through open records laws are often hindered by “aggressive trade secret and confidentiality claims” due to governments’ outsourcing of algorithmic implementations to third parties, often private companies.

---

13 EuDEeco, Modeling the European Data Economy, supra n.3.
15 See, e.g., Robert Brauneis & Ellen P. Goodman, Algorithmic Transparency for the Smart City, 20 Yale J.L. & Tech 103, 153 (2018); see also Tim O’Reilly, Open Data and Algorithmic Regulation; Michael Flowers, Beyond Open Data: The Data-Driven
2. Data Produced from Data

Several workshop participants felt it important to emphasize the extent to which data can be, and is, endogenously created. Endogenously-created data, or data that is produced from the processing of other data, might often be captured under the category of closed or proprietary data, or the “semi-open” structure described above. Such data is especially prevalent in the context of machine learning and big data analytics, where a proprietary machine has created the new datasets. The relationship of endogenously-created data to machine learning is potentially significant to the task of mapping the data economy for at least two reasons.

First, a common refrain in the era of big data and machine learning is that “data have also become highly valuable for uses other than what data producers or collectors initially intended.” The further processing of data for purposes other than those for which it was originally collected has implications for the speculative value of data, which can complicate efforts to quantify the value of data holdings through legal transparency and accountability mechanisms. Second, instances of further processing and the generation of new datasets also have implications on the legal and regulatory frameworks around data ownership and control. It is not always clear whether individuals have rights with regard to data that was produced from their data. Moreover, intellectual property interests in derived data are also unclear. In his paper entitled Data Producer’s Right & the Protection of Machine-Generated Data, Peter Yu describes the European Commission’s effort to create “a new data producer’s right for non-personal, anonymized machine-generated data,” highlighting many practical and policy complications of recognizing such a right.

3. Data About Data

Workshop participants also discussed metadata, or data about data more broadly. For instance, once a researcher or business decides on a goal that will require data analysis, they will need to make a determination about the data they need access to in order to achieve that goal. In order to answer that question intelligently, however, it will be necessary to know what data are available. Because information about what datasets exist in the world is so important to research, one workshop participant suggested the idea of mandatory dataset filings, like disclosures required by the SEC and

---

17 Id.
in other regulatory frameworks that emphasize transparency requirements. This idea will be explored further in Section II.B below.

In the context of contracting over information, data about data is valuable as a means of resolving what is referred to as the “information paradox” in negotiations.\(^1^9\) In his article entitled *Exchanging Information Without Intellectual Property*, Michael Burstein describes the use of contract law mechanisms to allow parties share useful data about underlying valuable information, without compromising the value of that information. Through this approach, he explains that contract law mechanisms can be an alternative to traditional intellectual property protections, and the challenges embedded therein when applied to data. Recognizing the distinction between data and data about data has several potential, and potentially significant, implications on the choice among available policy interventions when designing an information governance regime.

### 4. Invisible vs. Visible Data; Implicit vs. Explicit Disclosures

One workshop participant remarked that, since both data and metadata are being used to train algorithmic systems, it might be helpful to think about mapping data within a matrix. The proposed matrix would consist of invisible vs. visible data, and implicit vs. explicit disclosures. Because aggregations of data reveal underlying patterns that are invisible to individual users, and because companies and governments draw upon people’s data to infer other information that is not explicitly expressed within the data, this matrix could be used to track the degree to which users are actually informed about the uses to which their data is being put.

The participant used Instagram filters as an example of a potential use case for the matrix. Instagram filters are highly visible to all users and the use of the filter itself is an explicit disclosure, but companies can draw hidden inferences from this highly visible data. In the case of Instagram filters, correlations can be drawn about the mental states of people who use certain filters, which allows analysts to infer whether certain users are depressed, for instance. These correlations are invisible, and the inferences drawn can be thought of as implicit disclosures, unknown to the users themselves. Recognizing and tracking the types of data disclosures on this matrix could provide useful information about how companies like Instagram, and Facebook and Twitter, are using our visible, explicit data to inform invisible decisions related to our health, for instance. And perhaps, the participant suggested, classifying the data and the disclosures into different quadrants could help us to think about how informed users are about what their data is being used for; and for thinking about interventions, in the implicit space in particular, because the invisible implicit space might be where the biggest concerns lie regarding potential interventions.

---

5. Models of Data Acquisition: Builders & Buyers

A participant suggested framing the discussion in terms of a binary model of data acquisition, where data holders either build their own data collection infrastructure, or buy data from other entities who previously collected the data.

Under the "build-it" model of data acquisition, the builders consist of companies that are good at building platforms to gather their own data. Facebook, for example, “built” a social network system that provides a stream of useful data for all sorts of machine learning & big data applications.

Under the "buy-it" model, the buyers partner with other companies to access their data. IBM, for example, can partner with Sloan Kettering to use their data to build a predictive learning modeling system.

The two models may lend themselves to different forms of oversight and regulation.

D. DATA TENSIONS

Workshop participants highlighted several tensions that often arise when considering different approaches to data governance and mapping the data economy. Some of these tensions are outlined below.

1. Data Access: Efficiency vs. Friction

In the context of data access, there is a lot of tension between open-data advocates who, on the extreme end of the spectrum, argue for unrestricted access to as much data as possible, and those who warn against the perils of releasing data without understanding or appreciating the negative consequences. This tension was characterized by workshop participants in terms of efficient access vs. friction in data access. The arguments for efficient access frequently invoke innovation and competition. Because researchers are discovering new applications and value in big data, and because it is not always possible to foresee the value of certain data at the time it is collected, efficiency advocates argue that roadblocks to further processing and sharing of data have a chilling effect on innovation. And because closed data often denies data to small businesses, startups, and citizen scientists, while at the same time preserving the concentration of data within a few large

---

20 Although equality of efficient access to data for big and small organizations alike would intuitively promote competition, one potentially powerful counterpoint to the premise that efficiency and competition are symbiotic complements is that, as one participant remarked, the optimally efficient number of platforms is one for everything. Platforms thrive when they serve a critical mass of users, and amassing all data in massive data lakes or oceans would allow the most efficient access to that information, at least for the one who holds that data.
firms, laws and policies that restrict access also stifle competition. On the other hand, the potential perils of open and unrestricted access are well documented, including concerns about privacy, discrimination, chilling free speech, and restricting autonomy.

Reflecting these concerns, one workshop participant suggested that the data economy needs well-placed friction to help protect users while still allowing innovation.21 Going further, the participant noted the need to develop an argument that friction is a worthy goal in and of itself, and not merely a cost to be navigate in pursuit of another goal. The participant made a broader critique about the centrality of economic theory in modern legal theory. Because economists tend to value efficiency in transactions and the optimal reduction of transaction costs as the ultimate goal and virtue, the participant urged that the discipline of economics has had a pernicious effect on legal theory, due in part to the fact that the discipline is deeply normative while masquerading as a purely descriptive model.

Participants also discussed whether concerns about the friction caused by over-regulation and recognition of individual rights are overblown. One speaker analogized that, even in systems like France, where creators and authors have an inalienable, individually-vested right to pull back copies of a work whenever they want, mechanisms have developed to minimize the friction while vindicating those rights. Collective rights societies are the natural consequence of creating or recognizing a multitude of atomistic rights in a space where they need to be adjudicated en masse: eventually another intermediary will be created to clear all the rights collectively.

Because regulations create friction in data access, the tension between efficient access and access friction will resurface throughout the discussion of potential policy interventions in Part II below.

2. Data Compensation

Returning to the conversation of data-as-labor / data compensation, one participant drew a distinction between data consumers and data producers, and the tensions that exist between those poles. The speaker noted that questions of whether users should be compensated for their data arise even when data does not require the time and energy we typically associate with work. In many cases, users are forced to allow companies to access information stored on their devices or in their accounts, such as photos and contacts lists. In other cases, companies track users' movements while they go about their lives, including not only geolocation data but also the movements of their computer cursor as they browse the web.

---

Many people and organizations have voiced concerns about inadequate compensation for their data. According to one participant, these concerns raise two questions: (1) what are the reasons that we are not being adequately compensated?; and (2) how do we solve this issue and ensure adequate compensation? These issues will be discussed in further detail in Section II. A third question involves the foundational ambiguities around the value of data: for instance, what is the true price people pay for “free” apps? Is there any way to value data in light of unforeseeable applications? Data valuation continues to be a source of great uncertainty and tension.

3. Data Asymmetries

Workshop participants highlighted many ways in which the data-driven economy is characterized by significant asymmetries in data access, holdings, and power.

As companies amass ever greater pools of data, and the analytical tools to leverage that data become ever more sophisticated, the power associated with data holdings is greater than ever before. Those who hold and analyze data have incredible sway over the course of human events: for instance, whoever holds and controls climate data can control how that data is used to shape public policy and opinion. One participant noted how these asymmetries argue in favor of data visibility, for policymakers and the public to know who the major holders of data are, what data they hold, and the ways in which they can impact the fate of the world.

Another participant emphasized how companies with large data holdings exercise market power in multiple directions, and to affect potential entrants and competitors. This is especially true for two- and multi-sided platforms, where the platform providers facilitate transactions for the provision of goods or services. For example, at issue in the Supreme Court case Apple v. Pepper are Apple’s constraints on app developers and end-users. Apple requires that developers pay an annual fee to access the App Store platform, and the company imposes other app pricing constraints as well as content restrictions and conditions. On the other side of the platform, Apple prevents end-users from using any other app distribution platform to access services on their Apple iPhone. 22

Powerful, data-driven platforms also have access to valuable insights about potential competitors who, in many cases, must rely on the dominant platform to conduct their business. In Should We Be Concerned About Data-opolies?, Maurice Stucke explains:

22 In re Apple iPhone Antitrust Litigation, Amended Consolidated Class Action Complaint, Joint App. at JA-48 (“By design, Apple programmed the iPhone in a manner that prevented iPhone purchasers from downloading any Third Party Apps offered by software manufacturers who did not share their revenues with Apple or pay a fee to Apple to sell through iTunes.”).
The data-opoly can use its relative advantage in accessing and processing personal data, such as watching for trends in its proprietary data from posts on a social network, search queries, emails, and the like, to quickly identify (and squelch) nascent competitive threats. The dominant firm can acquire entrants before they become significant competitive threats or blunt the entrants’ growth. It can manipulate its search engine results to make it harder to find the entrants or remove them from the app store.23

Several workshop participants also highlighted the impact that data asymmetries are having on perceptions of strategic competition. For instance, when companies are asked whom they see as their biggest strategic competitors, they identify those that have “stickier” customer relationships — and data is often a critical guarantor of customer stickiness. For example, Google feels threatened by Uber, despite the fact that the two companies do very different things. One reason is the relative stickiness of these companies’ relationships with their consumers. Social media platforms are an even stronger example of sticky customer relationships premised on the platform’s data holdings. Because the utility and allure of social media platforms is often directly related to the size of the userbase and the richness of personal information shared, potential competitors face substantial barriers to entry, and users are left with virtually no alternative platform to which they might migrate. One workshop participant suggested that it may be helpful to have a better understanding of these perceived strategic threats, in order to appreciate that a critical component is not just what data an entity has, but the strategic threats of sticky customer relationships.

Workshop participants suggested that another important piece to map in the context of data asymmetries, is the computational resources a company can bring to bear on its data holdings. One participant noted that it may not matter how much data an entity owns if its lacks the resources to analyze and leverage that data. The participants urged the need to think about whether there are potential competitors that rely on the computing or server power of a third party — for instance, the growing reliance on Amazon’s server power and the outsourcing of computations — in assessing relative market power. The participant noted that concerns about computation power to not regularly feature in antitrust analysis, but they should.

4. Competing Interests in Data Access & Control

Some of the most fundamental tensions in data governance are those involving the competing and conflicting stakeholder interests in data access and control. Individuals have privacy and autonomy interests in maintaining control over the flows and uses of data about them, and they have an interest in adequate compensation for their data, including their time, attention, and labor.

Traditionally, these individual privacy interests have been pitted against a litany of societal interests: public safety and national security, scientific and medical research and innovation, distributive justice interests in fairness and equality, and economic efficiency and consumer welfare, to name a few. Workshop participants highlighted the rise in recent scholarship around privacy as an inherent, broad social value, and privacy as a “public good,” but one participant remarked that this framing has largely failed to gain traction because traditionally, the societal interests in privacy have not been highlighted.

At the same time that data can be used to further social and distributive justice interests in fairness and equality, the burgeoning algorithmic accountability movement has also shown the myriad ways in which data has been employed as an instrument of solidifying and exacerbating systemic and institutional racism, sexism, inequality, and injustice. On the other hand, the growing interests in data transparency and algorithmic transparency may frequently stand in tension with individual privacy interests, as well.

Furthermore, one participant noted, even data that is already publicly available can have very serious private valences. For example, OKCupid used the explicit, visible data provided by users to analyze implicit, invisible, characteristics, that most users probably had no expectation would be revealed through what they were sharing. In another example, researchers scraped videos that transitioning transgender individuals had posted online documenting their transition. The researchers tracked how faces changed during the transition process, and used that to train facial recognition software in order to be able to identify individuals even as they underwent significant changes in real-time. Another example is a study that used photos posted by LGBTQ individuals to test whether an algorithm can recognize “gay faces” on a dating app. These are all examples of publicly available data with serious private valences.

Participants also noted that the tensions between data benefits and harms comprise a far broader set of concerns than privacy. For one, data sharing can have both negative and positive impacts on both private and public entities. Open access to what are now proprietary datasets could empower small businesses, startups, and citizen scientists to engage in activities that would have otherwise been impossible. Public infrastructure, such as public streets and transportation, are affected by the data collected and held by companies such as Waze. Finally, open access to data can be a means of democratizing decision-making, for instance through increased transparency -- but businesses have interests in restricting access to information that may be vital to meaningful democratization of the political economy more broadly.

5. Challenging Conventional Data Claims

Workshop participants questioned several conventional claims and assumptions about the nature of data and big data analytics, some of which are highlighted here.

a. Data as Static

One participant sought to avoid talking about data in the abstract; and, in talking about data more concretely, to always be mindful of the temporality of data. The participant suggested that a particular dataset can be thought of as a snapshot of the world, in that it captures a picture of the world at a given time. Thinking about the temporality of data means realizing that the value of any given dataset diminishes over time as it becomes stale, and can even expire, sometimes even at the moment of collection. In this sense, while the value diminishes or expires, liability typically endures with the dataset. So, when talking about mapping the visibility of data and keeping track of what databases exist, it may be better to think in terms of data streams, and data pipelines, rather than simply the fixed datasets as units. The participant noted that data temporality is an important aspect of data as a shared infrastructure, and also can help to explain why a company like Google would perceive a strategic threat from Uber. The ongoing relationships with customers that Uber’s model encourages allow Uber to continue to be involved in the world of their customers, to collect a stream or pipeline of data snapshots, and this can create more value as data about longitudinal relationships.

b. Data as Objective

Another participant challenged what may often be an implicit assumption about data: that it reflects an objective picture of facts. The participant argued that because data is a human construct, given form and value by human operators all the way through its lifecycle, it is unavoidably permeated by human subjectivity and bias.

The inherent subjectivity and bias of data is increasingly important to keep in mind in an era of big data analytics and machine learning, and will be an important consideration when thinking about the virtues and drawbacks of different regulatory structures in the context of data governance. Because regulating data access and control will inject friction into data processing, it is important to consider how that friction might make less restricted datasets more attractive, and to question whether that may create a higher risk that freely available datasets include bias that could pollute and distort analyses.

c. The Promises & Necessities of Big Data

Several participants challenged the ability of big data and AI to deliver on the promise that they will dramatically improve analysis over traditional analytics. At the core of this critique is a rejection of
one of the central premises of big data analytics: that data processors need as much data as possible to achieve the best results (“in the world of big data, more is always better.”)\textsuperscript{25}

One participant suggested this challenge be framed as “big data as laziness,” charging that analysis can be completed with smaller, smarter use of tailored datasets. The participant referenced Jonathan Frankle and Michael Carbin’s \textit{Lottery Ticket Hypothesis},\textsuperscript{26} which articulates that in any deep learning system, you throw out 98% of the data in the first step and hone in on the 2% of the data that you really need. The participant noted that right now, we’re just not good at identifying the 2% of useful data at the outset.

The participant argued that this point is potentially very important in society’s efforts to create a comprehensive data governance regime. If we simply accept that researchers need all the data all the time, and shape law and policy around that presumption by creating entitlements to larger and larger pools of data, it could be extremely difficult to claw back the position in ten years’ time even if the presumption proves incorrect.

\textit{d. The Red Herrings of AI & Predictive Analytics}

One participant pointed out that in many cases, especially in the context of smart cities, AI and predictive analytics are often red herrings which distract from the truly important aspects of data governance. The participant explained that smart city technology is almost exclusively concerned with data-driven decisionmaking, not predictive analytics. This is something that the participant believed important to keep in mind when considering the problem and solution spaces.

E. MAPPING THE PROBLEM SPACE: ACCESS & CONTROL GOALS FOR THE DATA ECONOMY

Overall, workshop participants identified several high-level problems of data access, control, and governance. These can be summarized as follows:

First, the lack of data visibility and transparency have prevented informed discussion and decisionmaking, and the workshop identified the need for increased visibility as a vital first step in continuing the conversation. Ensuring transparency in some contexts will require mechanisms for compelling access to datasets that are currently withheld from scrutiny.


Second, some participants noted the need for mechanisms that ensure individual data subjects and creators have a meaningful choice about what they do and do not share. The ability of individuals to withhold data will need to be balanced against the need for that data to further public goods like scientific advancement and efficient distribution of public funds and services.

Third, there needs to be an affirmative obligation on companies to protect the data of their users, and to the extent that data is shared, the obligation needs to pass with the data.

Fourth, some participants identified the need for mechanisms that prevent and reverse the quiet accumulation of market power through data holdings and monopolization of platforms and computational resources.

Fifth, some participants identified the need for limits on collection and retention, to mitigate the risks of abuse and exposure of sensitive data in cases of breach.

Sixth, the group considered the need for restrictions against certain kinds of processing that are deemed harmful or otherwise undesirable from a public policy perspective. Some examples include processing that contributes to bias, discrimination, and unfairness more broadly. Distinguishing harmful and beneficial data processing is itself a problem due to competing interests.

Finally, some participants noted a need for accountability for harms that occur when data is improperly processed, inadequately protected, or wrongfully shared, and not just in security breach cases. Current punishments and accountability mechanisms have been insufficiently severe to establish effective deterrence of the kind of behaviors that society wants to prohibit. Improper processing includes processing that leads to health and safety risks, for instance in the areas of healthcare and pharmaceutical research.

Subsequent panels in the workshop considered varying regulatory approaches to the data economy. A first session focused on competition and data labor, a second on intellectual property, and a third on testability and transparency. Throughout these discussions, several themes emerged; these themes are summarized below before turning to the specific workshop topics.
III. EXPLORING THE SOLUTION SPACE

A. REGULATORY CHALLENGES

The workshop touched upon a number of regulatory challenges to data governance, a few of which are highlighted here. These regulatory challenges are often due in part to, or exacerbated by, some of the foundational ambiguities swirling around data governance and mapping the data economy.

1. Avoiding Unintended Consequences

Several participants urged caution when advocating for certain regulatory interventions; recognizing that data cuts across so many disciplines, it raises particularly acute challenges to identifying and avoiding unintended consequences. In his paper discussing proposals for sui generis database protection, Peter Yu describes some of the complications that could “arise when the protection of the proposed data producer’s right spills over into other areas of the law, including those within and outside the intellectual property system.”\(^\text{27}\) In particular, Yu explores complications concerning “privacy, trade, and investment,” in an argument that demonstrates the need to consider the full spectrum of institutional complexities around particular policy solutions.\(^\text{28}\) Similarly, Michael Burstein has argued that “policy interventions that seek to promote transactions in information must be made with a more complete understanding of both the social welfare trade-offs involved in different strategies and the specific business and legal environments in which information transactions take place.”\(^\text{29}\)

More generally, participants noted that any time we think about creating new rules to solve a problem, we need to be aware that solving for one issue will inevitably cause new issues. Returning to the example of YouTube videos of transitioning transgender individuals, a participant noted that one control to prevent this scenario would be to recognize a property right in one’s image when posted online. Yet such a right might also have the undesirable effect of granting white supremacists the power to assert a property right over online images of themselves attending hate rallies, thus protecting them from public scrutiny and social and professional consequences.

One participant pushed back against the seriousness of concerns about unintended consequences of regulation, arguing that they have largely proven to be unsubstantiated, and that the internet and digital data technologies and infrastructures are too resilient to be adversely impacted in a significant way.


\(^{28}\) Id.

2. First Amendment Concerns

Participants noted the First Amendment concerns which have been fairly widely discussed, especially in the context of algorithmic accountability regulation. The Supreme Court has held that "the creation and dissemination of information are speech within the meaning of the First Amendment."30 One participant noted that these concerns deserve special attention given the current composition of the Supreme Court, whose decisions may suggest a majority consensus in recognizing the First Amendment rights of companies in their software code and algorithms. In his article Information Fiduciaries & the First Amendment, Jack Balkin notes "a far larger trend in which the First Amendment has gradually been transformed into a bulwark of protection against business regulation.31 This trend may make direct regulation of data and analytics systems difficult, and thus may suggest the need to pursue alternative approaches to data governance. On the other hand, because First Amendment jurisprudence evolves and has shown considerable sensitivity to nuance in other contexts, it is far from certain that the First Amendment will continue to present serious regulatory difficulties in the data access and control space.

B. TRANSPARENCY & ACCOUNTABILITY

The group discussed the need for some mechanism for holding data processors accountable for their actions, including when those actions reflect bias, discrimination, and other notions of unfairness. Participants also explained the need for more accountability regarding collection of data, which too often happens out of the sight of regulators and the public. Some participants suggested that certain automated decisionmaking and machine learning applications, such as making decisions regarding insurance rates based on who our friends are, will be unpalatable from a public policy perspective, even when these applications do not fall under currently recognized prohibitions. To make regulating for accountability possible, transparency regulations will be essential. The group focused their discussion on different approaches and components to meaningful transparency, and considered some of the challenges, and potential solutions to achieving transparency through regulation.

1. Components of Meaningful Transparency & Accountability

The workshop discussion echoed sentiments in a recent editorial by Julia Stoyanovich and Bill Howe, where they argue that "Algorithmic Transparency Starts with Data Transparency."32 In order

to understand, evaluate, and regulate machine learning algorithms, for instance, access to the training data — the learning datasets that those algorithms are trained on — will be an essential first step to any accountability mechanism.

More than just transparency as data access, however, participants stressed the need for the data to be understandable; to be explainable and interpretable. Meaningful explainability for individuals affected by an automated decisionmaking algorithm will require a disclosure and explanation of what data was used to make the decision, but also what data was not used as part of the automated decision. For instance, it will be important to know whether healthcare decisions are being made based on pre-existing conditions, and the basis for inclusion or exclusion of that information.

The workshop discussed two potential approaches to imposing accountability mechanisms, expressing two different policy goals or values. One path involves the simple imposition of mandatory data disclosures. The policy behind this approach is that transparency in itself is a virtue and once meaningful transparency is achieved, the market will ensure appropriate accountability. The other path involves merit-based regulation, where transparency is just the first step. This path requires normative judgments about the merits of certain data collection and processing operations, and may be less palatable to data processors and thus more controversial and challenging for policymakers. However, one participant urged that ensuring transparency will not automatically lead to accountability. The participant suggested the need for a “stick,” particularly when dealing with data monopolies.

2. Challenges for Meaningful Transparency & Accountability

The group discussed a variety of barriers to openness regarding the different types of datasets necessary to ensure meaningful transparency and accountability. Companies might naturally not want to reveal the information used to make certain decisions or train algorithms for a variety of reasons. Companies have concerns about competition and theft of innovations that were expensive to develop. Objections to training data transparency based on trade secrets have become particularly prominent in the case of smart cities. There are additional security concerns that revealing algorithms would expose them to exploitation and circumvention -- for instance if Google revealed its search engine algorithm, people could exploit its operations in order to spam and manipulate the top results.

One participant noted that technology makes it increasingly easy to re-identify individuals within a dataset or to otherwise compromise privacy. This presents problems for transparency, for instance,
when access to training data is an essential component of effective oversight. One participant suggested the possibility that cutting edge privacy preserving technologies, such as differential privacy, might provide the means for allowing meaningful transparency and oversight while also preserving privacy. Differential privacy provides one path to the creation of synthetic datasets that resemble the underlying data without revealing too much about the actual data subjects.

Some participants explained how the Freedom of Information Act (FOIA) and state open records laws have served as a solution to some of these transparency challenges. Christopher Soghoian has used FOIA in creative ways, and Paul Ohm has suggested that FOIA may be used as a "hacking tool." Other participants highlighted the ways in which FOIA and open records laws have fallen short of delivering truly meaningful transparency. Catherine Crump has written about how procurement policies have made transparency a challenge for many at the local level. Others like Ira Rubenstein have written about the need to have procurement policies that ensure local communities are stakeholders in any decisions, and to have transparency requirements baked-in to the policies by design. One participant noted this will also require greater candor and cooperation from state and local governments, especially in the burgeoning smart city context.

One participant expressed general skepticism about the potential for meaningful transparency and accountability in the machine learning context in particular, due to the highly dynamic nature of training data. The participant found it difficult to believe that there will ever be a discrete unit of training for any particular application that the public will be able to access and scrutinize.

C. RESTRICTIONS ON DATA COLLECTION, PROCESSING, & STORAGE

Aside from recognizing meaningful transparency as an essential first step toward any meaningful data regulations, one participant remarked that limits on data accumulation and aggregation could be useful. Regulations could require entities to not only disclose, but also justify, the logic of their data accumulation, whereas currently accumulation is often merely a passing operational expense. For these limits to be effective, the participant suggested that regulations would need to impose a real cost on data collection and storage.

33 Id.
35 https://its.law.nyu.edu/facultyprofiles/index.cfm?fuseaction=profile.publications&personid=30084
Participants frequently referred to provisions in the General Data Protection Regulation (GDPR) to identify components of meaningful data restrictions, including the prohibition on processing of personal data without consent, and the right to be forgotten. One participant remarked that, where the law requires content to be removed, such as in DMCA takedowns, companies are actually very good at responding to and complying with such legal obligations. Newer technologies such as the contentID system for YouTube have allowed companies to become better at complying with the law—as such, the market responds to effective regulation, when such action is legally required.

Several participants, however, expressed profound skepticism about whether many provisions of the GDPR and similar privacy and data protection laws could be implemented in practice. One participant noted that the right to be forgotten and the right to data portability, while worthy of aspiration, are generally technically impossible, at least at the moment. The participant estimated that it may take 20 years before these rights receive meaningful technical support.

One participant noted that meaningful data portability requires a technical shift from systems built to collect and keep data to systems that support sharing and transfer of data. Even though data portability is described as a solution to help small companies, large corporations are at an advantage in satisfying these new requirements.

Discussing the right to be forgotten, one participant noted that it is not currently possible to guarantee that copies of data are actually removed. One participant emphasized the temporality of data, the fact that data changes and undergoes transformations and reincarnations. For instance, with a model doing predictive analytics, the results that are based on some particular data constitute new data that is in a sense comprised of the original data. There is an open question about what it means to truly delete or “forget” the underlying data in such a situation. Moreover, deep deletions of data within a system can have substantial downstream effects due to the ways in which datasets and systems are intertwined.

Takedown demands can also be much more burdensome for smaller companies that cannot develop systems and processes that are as effective or efficient as the ones used by the dominant firms. One participant noted that small companies are also more likely to receive takedown demands, because people are intimidated to go after large firms to de-index, but feel comfortable demanding takedowns from smaller players.

**D. COMPETITION LAW**

In a workshop session focused on competition, participants discussed the potential for antitrust and competition law to address concerns about asymmetries: asymmetries in data holdings, in access to data, in computational resources, and otherwise. Some participants felt that antitrust law is already suited to handle these issues, and indeed has been handling them to some degree, assuming that protecting the competitive process is the appropriate “north star” for antitrust regulation. Some
further reflections on the competitive significance of data, and other challenges, opportunities, and potential limitations of antitrust law as a data governance solution are discussed below.

1. Scrutinizing the Competitive Significance of Data

Participants discussed the important opportunity for antitrust authorities to examine and potentially limit data consolidation when they perform merger reviews. Historically, antitrust authorities have focused less on this question, especially when data collection is not a “core” business of the merging companies. In many instances, the largest data firms are not engaging in horizontal mergers, but are conducting vertical mergers, building out their own internal vertical integration capacities, or branching off into what appear to be entirely unrelated markets.

In this context, one participant noted, the focus cannot rest simply on the amount of data possessed by any company, but on the competitive significance of the data.37 Another participant remarked that, in conducting this analysis, antitrust authorities need to take a broader view of “competitive significance” than they have to date. Given the cross-functionality of data and the value that can be revealed by combined datasets, most data has some competitive significance, and should be viewed as such. The participant raised the example of a toy store merging with a restaurant chain. Although the two businesses are not traditional competitors, the datasets of each entity could possess significant competitive value.

Another participant noted that authorities would likely distinguish between consumer data and proprietary datasets, commenting that some consumer data is so readily ascertainable that it may not be viewed as competitively significant.

2. The Essential Facilities Doctrine

One participant noted that data competitiveness concerns often arise outside the context of traditional horizontal and vertical mergers, and instead arise from, for example, the lack of meaningful alternatives to existing platforms. Because in many cases there are no obvious low-cost opportunities to bolster competition, the participant wondered whether it would be feasible to finance business models that can achieve scale while avoiding the issues of current platforms. One proposal put forward was to fund these sorts of essential services from sources that do not require the sort of compromises that come with venture capital funding.

Another participant alluded to the potential use of the essential facilities doctrine to address this issue of platform and market monopolization. Large data firms and platform providers like Amazon can leverage their dominance in terms of data access and computational resources to gain invaluable insights into market trends, to demand supracompetitive prices from suppliers and distributors, to manipulate the behavior and perceptions of consumers, to monitor the market with an eye out for potential upcoming competitors, and even to exclude potential competitors from their platform altogether. Recognizing that vertically integrated firms with this sort of market power “may deny access to a rival in an adjacent market,” the essential facilities doctrine “imposes sharing requirements” on the dominant firm to ensure that competitors have easy access to the essential facility.38

One more “radical remedy” of the essential facilities doctrine entails “dividing the facility among multiple owners.”39 In his article on the essential facilities doctrine, Sandeep Vaheesan cautions that while this remedy can mitigate “the threat of monopoly leveraging, [it] could sacrifice important efficiencies.”40

3. Limitations of Traditional Antitrust for Data Governance

Several participants raised concerns about potentially intractable limitations of antitrust frameworks to solve problems of data access and control. Participants noted that, for one, it is difficult to make generalized assertions about the effectiveness of antitrust law as a governance tool, as the treatment of each case is so highly fact-specific.

One participant highlighted the risk of thinking of antitrust as a panacea. If breaking up the big firms or requiring shared access to market platforms is not accompanied by a reconfiguration of the incentives driving data accumulation and control, then the result may be a world with a multitude of competitors without having covered any meaningful distance toward addressing the underlying problems. The participant pointed to regulations on the telecommunications industry as an example, where relying on competition has not always made much difference.

Another participant challenged the assumption inherent in antitrust law that competition and pro-competitive behavior are unwaveringly good. Pointing to a problem of toxic competition, the

38 Lina Khan, *Amazon’s Antitrust Paradox*, YALE L.J. 710, 800–03 (2017). Khan explains that the four-factor test for an essential facilities claim was enumerated by the Seventh Circuit in *MCI Communications Corp. v. American Telephone & Telegraph Co.*, 708 F.2d 1081, 1132–33 (7th Cir. 1982). However, the 2004 Supreme Court case of *Verizon v. Law Offices of Curtis V. Trinko, LLP*, 540 U.S. 410 (2004) cast doubt on the future viability the essential facilities claims. See Khan at 801.


40 Id.
participant argued that bringing in more application and platform developers may actually exacerbate many of the harms that a data governance regime ought to address (for example, through a race to the bottom to monetize consumer data).

Finally, one participant questioned whether the fundamental assumptions of antitrust are irrelevant in the platform-based data economy. Specifically, the assumption that competition is the solution may be undermined by the nature of giant platforms, where in many cases it may be most efficient to have a single platform. The participant suggested that what data governance really needs to prevent are some of the practices that entities currently get away with by virtue of their size. In order to prevent identity theft, or to prevent reverse redlining of mortgage rates, then it may be time to start making those rules more directly, though this is not a palatable option for people who believe markets solve everything.

E. LABOR LAW: PERSONAL DATA MARKETS

The group also touched upon the possibility of solving some data access and control issues by treating data as labor, and establishing personal data markets. The notion of data as “labor” to be compensated is gaining increased traction in academic circles, as discussed in the recent piece by Ibarra et al. entitled Should We Treat Data as Labor? Moving Beyond “Free” and the book Radical Markets by Eric Posner and Glen Weyl, in which they “propose creating a data labor movement to force digital monopolies to compensate people for their electronic data.” The group considered potential advantages and disadvantages to treating data as labor.

One participant questioned whether applications and platforms that rely on crowd-sourcing should be required to share their profits with the users who helped generate the value. For instance, when Waze was bought out, investors and executives received a share of the $1.15 billion payout. The users who contributed their geolocation data as well as their time in reporting traffic hazards and incidents are the ones who created the value, but all users received in return was the free service.

The group discussed the idea, recommended by Christina Mulligan, of collective action as a solution to the data compensation problem. One participant suggested that what we need are data unions, which could provide a mechanism for users to engage in collective bargaining over Terms of Service. The participant also referenced the Data Transfer Project, an initiative by some of the dominant data firms to support data portability, arguing that what is truly needed is the ability to automatically transfer datasets from one platform or service to another through an API. The participant remarked

---

that the ability to quickly and effortlessly transfer from one platform to another would make collective bargaining even easier. Users would not need an NLRB for data bargaining if they could turn to a data union representative who could negotiate with the data firms, backed with the threat that they could pull all their members’ data from the platform unless the platform provider agreed to certain demands.

Another participant noted that data unions could focus on more than protecting privacy and ensuring adequate compensation. Data collectives could also promote and facilitate shared control and democratic values, such as enabling democratic decision-making and democratization of the political economy more broadly.

One participant was skeptical about a particular claim found in the data as labor literature, regarding work as providing a sense of dignity, and the ability of data labor to fill the gap predicted to emerge as AI replaces humans in the workforce. Exploring the relationship between data as labor and the rest of auctions-based theory, the participant felt that the idea that posting on social media, for instance, could provide dignity seemed absurd. The participant thought that the only possible way of justifying such a claim would be if the platform were to fade away completely, and the users were allowed to take complete control over how they post. Moreover, the participant pointed out that data as labor is incongruous with our current conceptions and interactions with data. Right now, datasets are de facto treated as property, as discrete units that can be shared and transacted over. For treating data as labor to be an appealing option, the participant remarked, it would have to be the case that no one was really the lasting owner of the data that is being produced.

F. INTELLECTUAL PROPERTY LAW

One session of the workshop considered whether and how intellectual property can address some of the concerns and tensions of the data economy. In particular, the group discussed whether the framework of intellectual property can offer a way of restoring some balance of power between data subjects and creators on one hand, and data collectors and consumers on the other.43

Several participants expressed hope that the benefits of using the intellectual property law framework could be substantial, but those benefits would be conditioned on defining the property rights in a limited way, while also limiting contractual and technological workarounds.

The group acknowledged that the friction created by recognizing data property rights may be at odds with the aim of creating the most value in markets. For example, several participants noted

that a proposal currently under consideration in Europe that would prohibit the processing of information for machine learning purposes without a license would have significant negative effects (this is discussed in more detail below). At the same time, workshop participants acknowledged that companies already treat data as property over which they assert de facto ownership rights. Accordingly, it may be both realistic and possible to adopt an ownership framework, but use complementary IP tools such as fair use, collective licensing, and contractual arrangements to deal with the frictions created by such a regime.

One participant noted the potential broader benefits of recognizing data property rights beyond the matter of compensation and control. Pointing to Justice Gorsuch’s dissent in the Carpenter decision about whether access to an individual’s cell phone geolocation data constituted a search, they noted that if we frame a person’s interest in their geolocation data as an individual property right, it will be much easier for a court to find a search or a seizure of that data, rather than looking to whether someone has a reasonable expectation of privacy in the information.

Some issues related to particular areas of intellectual property law are discussed below.

1. Copyright Friction

Participants noted that intellectual property law, and copyright law in particular, can create friction in accessing data. This can create problems by limiting the datasets that researchers can draw on, skewing results and undermining the inclusivity and quality of the resulting work. For example, if copyright law implicitly pushes engineers toward easily accessible, low-risk data, there is greater risk that bias in that data (such as underrepresentation) propagates programs that leverage the data. One participant highlighted the example of the “Enron emails”, a large publicly-available dataset from the prosecution of Enron officials in the early 2000s that has been used to train various algorithms for classifying emails and natural language processing. The participant noted that the email traffic of white collar employees involved in a massive criminal trial may not be the best source of training data for widely-relied upon computer programs used by the entire population. The participant also flagged concerns with another widely available public dataset, Wikipedia, which contains rich information but often skews towards covering white, male subjects in greater depth than under-represented minorities.

The group also considered how certain copyright doctrines might bear in relation to data mining, machine learning, and AI. In the United States, there is a strong legal argument that data processing is a transformative use covered by the fair use doctrine, but in the European Union this approach is gravely under threat. If Europe succeeds in implementing a copyright directive that prohibits the processing of information for machine learning purposes without a license, the impact would be severe. Not only would the inputs for algorithmic training be restricted, undermining algorithms’ accuracy and representation. Such restrictions would likely also restrict transparency, with companies reluctant to reveal or share their training data lest doing so exposes them to potential liability.
**2. Trade Secrets Friction**

Workshop participants discussed the utility of trade secrets law as applied to data, noting in particular its negative implications for transparency and accountability. One participant highlighted how municipalities are contracting with private enterprise to implement smart city technology, and government officials may not realize that when they contract with these private companies, they often only receive the end product and do not have access to the underlying training data. This undermines governments’ ability to test and validate their own systems, and is also a major impediment to production in response to open records requests.

Participants suggested that some trade secrecy problems can be dealt with by tuning exceptions to disclosure and data production in procurement policies, and by amassing a larger corps of computer scientists who will refashion themselves as public civil servants, to lessen the need for outsourcing of such operations in the first place. One participant suggested that if the IP theme is pursued, it would ultimately require unwinding some of these established frameworks which are harmful to transparency and accountability, and to recognize when trade secrecy claims are meritless in order to prevent them from being used to avoid accountability.


Although the idea of data as intellectual property has many potential benefits for control, compensation, and even privacy, participants noted that there are real concerns about thinking of data in this way, and many externalities to consider.

For one, in the context of personal data as property, a participant drew the analogy to the existence of antecedent rights in individuals that are in some respects inalienable; that we can think of the body as property giving rise to rules protecting our sovereignty and autonomy, but at the same time, we generally do not want to commodify the human body. A participant cautioned that the commodification and commercialization of personal data could have dangerous consequences. Another participant referred to this as the paradox of personal data as property, since people do think of their data, like their bodies, as “theirs” — as belonging to them.

Building on the concern about personal data as property, one participant wondered about the issues that arise when personal data is aggregated into the proprietary dataset of a company. When the ownership rights of individual data subjects and creators morph into the rights of data holders, the implications of that corporate ownership and control over access to information about individuals are not yet clear. Participants stressed the need to be clear about which stages of data creation and processing, and which parties to the process, ought to be included in the intellectual property framework, especially in the context of this dichotomy of data subjects and corporate data holders.

Another participant wondered, with regard to the goal of using property rights as a way of restoring power to the individual data subject, whether there is any reason to expect that granting a property
right will work out differently from granting a privacy right. Notably, people still waive their privacy rights all the time. The participant remarked about the possibility of making the right non-waivable, but was not confident that would provide a realistic and practicable solution. Another participant noted the interest analog of individual artists who assert their intellectual property rights through collective licensing organizations. The participant noted that, once rights are recognized, market solutions develop to help administer and navigate the frictions created by those rights.

Another participant posed an even prior question regarding the propriety of considering intellectual property in this space at all — a normative question about whether these intellectual property concepts ought to map onto data protection and data access governance in the first place. For one, the participant argued that IP law does not appear to be necessary on the grounds upon which it is traditionally founded. In particular, IP law is typically understood as a solution to the particular problem of underproduction of intangible goods, due to a lack of incentive to produce such intangible goods when ideas can be freely appropriated by others. In the modern data economy, there is little reason to fear underproduction. There may be underproduction with regard to specific types of complex datasets that are not readily commercialized but have a significant public use – but other tools can help solve the incentive to create such datasets, such as prizes or government grants. And there are many mechanisms available outside of IP law that will also incentivize production and protect against uncompensated exploitation by rivals; for instance, terms of service, technological protection measures, and other contract rights and provisions, which is where many protections are currently found.

Another participant countered that the current system of enforcing data rights through terms of service has given rise to a system of bespoke terms and conditions, dictated by the dominant platforms. A possible advantage of constructing a new IP paradigm for data is avoiding the public and private harms that result from the current system of enforcement through terms of service. At any rate, the participant urged that when deciding whether or not to use a particular framework in the solution space, it is important to be mindful of what is currently filling that space, and if any other framework can fill the gaps or offer better solutions.

Rounding out the session, one participant noted that property rights are efficient and work well when you have very clear and specific information about the scope and boundaries of those rights. Without those clear metes and bounds, a property right essentially creates more problems than solutions. When you have questions about what the data is exactly and where one person’s ownership begins and another’s ends, such exchanges become very difficult.

G. CONTRACT LAW

As discussed in the previous section, participants noted that rights resembling property and ownership rights, and copyrights, already exist de facto. They are included within terms of service
and contractual provisions, and companies frequently assert ownership rights in data. One issue is that asymmetry in contracting gives control to companies, and very little power to individuals who simply click through the terms of service to access the platform they want to use.

One participant raised the point that terms of service and use do not help individuals whose rights are violated by third parties. If, for example, one of the transgender individuals who uploaded a video to YouTube tracking their transition sought to stop their video being misused, the participant was skeptical that that individual would be able to successfully litigate a claim against the researchers who scraped their data. Another participant noted that contract law and terms of service have been effective in preventing some harmful uses of data. For instance, people and platforms frequently rely on violations of terms of service to prevent the dissemination of “revenge porn.”

The group discussed how terms of service may be amended to limit or restrict uses for certain particular purposes. This could be a path toward introducing additional friction in the system, and terms could also be amended to offer exceptions for scraping for public good research services. In fact, one participant noted, this is already happening for APIs: Google releases an API for machine learning purposes, but withholds the underlying training data.

One participant noted that the critiques of current contractual frameworks (such as asymmetry in contracting) may be fixable by finding mechanisms that give individual users and data subjects greater negotiating power. It may be possible to find these sorts of mechanisms and systems without wading into the thicket of intellectual property law. One potential solution may be found in the data as labor framework, and the formation of data unions or data cooperatives, discussed in Section E above.

H. TAX LAW

One participant posited a novel application of tax law as a path towards transparency, accountability, and limits on data collection and use. They suggested that a “data wealth” tax could require entities to estimate the value of their data holdings to figure out how to tax that data. The tax itself could disincentivize unnecessary and unproductive data hoarding by imposing levies on data accumulation, while the reporting requirements could provide much needed transparency and visibility into the data economy. Participants pointed out the challenges to this approach; in particular, the difficulties associated with assigning value to data. It might be possible for a new tax framework to impose minimum values on certain categories of data holdings however, to at least achieve transparency. The tax framework might also be able to impose progressive costs based on the duration of data retention, as a way to increase the costs of data hoarding, since the current costs of hoarding have proven insufficient to deter such practices.
IV. QUESTIONS & PATHS FOR FURTHER EXPLORATION

A. IS DATA PRIVACY A LOST CAUSE?

One participant queried whether much of the focus of the workshop centered on transparency, accountability, and explainability was an implicit recognition that privacy has failed, or whether those are steps along the path to ensuring privacy. In other words, is it too late to scale back data collection practices and minimize data holdings, such that the focus for transparency has shifted to regulating data processing and use?

Other participants answered that it is not too late for privacy, that privacy has not failed, and it should not be abandoned as a goal. In fact, one participant argued that privacy has not even been given an opportunity to truly get off the ground, at least in the United States. Looking at Europe’s stance over the last half-century, on the other hand, shows the viability of another path forward.

The lingering ambiguity over what privacy actually means, what it requires, and whether anything can be done to reclaim whatever privacy may have been lost in the past, suggest that the privacy debate remains worthy of further consideration and elucidation.

B. WHAT CAN TRANSPARENCY ACTUALLY ACCOMPLISH, AND IS IT WORTH IT?

Some participants urged that transparency is not only essential to understanding and accountability, but that it is valuable and good in itself. Other participants were more skeptical. One participant questioned whether even considerable improvements in transparency would meaningfully contribute to the solution space. The participant suggested that, for instance, the provision of access to underlying training data may not suffice to allow effective understanding and oversight of the operations of machine learning and AI systems.

Another participant pointed out that emphasizing and mandating transparency measures could be extremely costly from a compliance perspective. Further, to the extent a data governance regime prioritizes transparency, it is important to think about accessibility and whether and how that data can actually be used. Further exploration of ways to ensure that the data can actually be used and accessed in meaningful ways might be helpful.

C. DATA PORTABILITY: WHAT DOES IT MEAN AND WHAT SHOULD IT MEAN?

The group discussed the lack of public consensus about what is meant by data portability, especially from a technological perspective. The GDPR requires only that data be obtainable in a machine readable form, but participants remarked that falls far short of what many people intuitively think of when data portability is mentioned. For many, true data portability would require the ability to easily transfer one’s social media profile, including pictures, contacts, and posts, to a competing
platform. Some participants argued such portability would increase users’ bargaining power when negotiating terms of service with platform providers; others noted portability brings its own security risks of consumers unwittingly sharing information with bad actors. This topic of data portability, and the feasibility of true interoperability and cross-platform compatibility, is deserving of further consideration.

D. DATA PHILANTHROPY: WHAT DOES IT MEAN AND IS IT WORTH WORKING TOWARD?

When a recent paper on the notion of data philanthropy44 was mentioned at the workshop, some participants expressed skepticism about the motivations and ultimate value of a nascent data philanthropy movement. The paper defines data philanthropy as “the act of sharing private data assets to serve the public good.”45 One participant cautioned about the use of a “philanthropy” framework, maintaining that it presumes and normalizes corporate ownership of data subjects’ information. Because the notion of data philanthropy and its implications were raised but not fully explored at the workshop, this theme may be worth further exploration.

45 Id. at 1.
V. CONCLUSION

One broad takeaway from the workshop was that no single legal framework is suited to address the myriad problems that arise around the data economy in our networked, platform-based information environment. Participants argued that an information governance framework would need to draw upon multiple elements from those discussed, including competition, intellectual property, labor, and privacy and consumer protection laws.

One participant suggested that regulators need to move away from a purist vision of their role and look at broader norms. Other participants pointed out how controversial and fraught with political risk that can be. Alternatively, it may be best to push for the creation of a new regulatory body, to have a single common regulator preside over the data economy as a whole.

One potential guide on the path toward an information governance framework is to ask the questions: where does regulation work well, and where does it need improvement? Facebook users, for instance, might be in a better position than government to decide appropriate uses of their data, but only if there is some legal mechanism for users to impose their determinations on Facebook. Or, perhaps the regulation approach is infirm if the public’s faith in the regulators has irreparably eroded due to regulatory capture. And perhaps if Facebook is not up to the task of making good ethical choices, then they are not the right entity to be making those choices.

One participant noted that it is important to be mindful of what current protections, regulations, and norms are simply the result of historic accident, or bad framing or analogizing – and whether those frameworks or elements of those frameworks ought to be reconfigured or abandoned in favor of something that makes more sense. For instance, it is merely an accident of legal framing that videos fall under copyright law, while so many other data points that are generated on Facebook and social media, like users’ likes, have not received copyright protection.

One participant noted the substantial difficulty of breaking out of all these traditional and well-established frameworks and moving toward an information governance framework, but that doing so is necessary to win back control over these matters. For instance, the privacy debate could provide some insights on how to proceed. But some participants believe that the information governance debate is what is truly needed, and it remains unclear whether experts, practitioners, regulators, and academics are equipped with the vocabulary and the knowledge necessary to have a more meaningful and productive discussion about information governance yet. There are tools that have yet to be invented, and that need to be invented, to begin to engage in this debate.

Even though the information governance framework is undeveloped and immature, and the equipment needed to have the debate arguably lacking, multiple participants suggested that the time is ripe to act in furtherance of a comprehensive information governance regime. Bills and proposals are before Congress and state legislatures. Regulatory agencies like the FTC and FCC are inviting comments on various proposals related to data collection and processing. There is urgent need for a roadmap to guide policymakers through this complex web of issues around data access and control. But precisely because the framework and the equipment needed to provide such a
roadmap are missing, the ultimate question at the end of the workshop is: what are some of the specific necessary preconditions that must be satisfied to get to a place where it would be possible to provide such a policymakers’ roadmap?

For one, there are enormous questions about how to divide the various impacts of data processing into clear categories of “benefits” and “harms.” Many of these uncertainties and ambiguities have been outlined in the preceding discussion. There may never be consensus on the merit of certain data applications. In light of that, what elements of the benefits vs. harms debate are essential to providing meaningful guidance to policymakers?

One participant suggested that, in light of the multitude of recent data-related scandals, from Cambridge-Analytica to the seemingly endless stream of massive data breaches and exfiltration of personal data from platforms with “bugs” and “glitches,” there is no need to worry about whether policymakers will recognize the harmful implications of the status quo — at least not the most egregious harms. What is needed now is to brush away what the participant described as a fatigue around the word privacy. The debates about benefits and harms, and legal and technical solutions to complex data-related problems have been fought among privacy scholars and advocates for decades, and that field has developed answers to many of the questions raised at the workshop. It may be possible to avoid the need to start from scratch and construct an entirely new framework. Instead, the participant suggested, what is needed is to breathe new life to these older solutions, to make the proposals now more forcefully than ever, given that the time is ripe.
VI. RECOMMENDED READINGS

A. MAPPING THE DATA ECONOMY

7. Autorité de la Concurrence, Opinion no. 18-A-03 of 6 March 2018 on data processing in the online advertising sector (March 6, 2018), http://www.autoritedelaconcurrence.fr/doc/avis18a03_en_.pdf

B. TRANSPARENCY & ACCOUNTABILITY


C. REGULATING COLLECTION, PROCESSING, & STORAGE


D. COMPETITION


E. DATA LABOR


F. INTELLECTUAL PROPERTY


VII. PARTICIPANTS

Elettra Bietti  
Harvard Law

Jennifer Broxmeyer  
Facebook

Mike Burstein  
Cardozo Law

Anupam Chander  
Georgetown Law

Julie Cohen  
Georgetown Law

Cynthia Cole  
Baker Botts

Aaron Cooper  
BSA | The Software Alliance

Aaron Fluitt  
Georgetown Law

Alexandra Givens  
Georgetown Law

Andrew Glickman  
Securities & Exchange Commission

Jake Goldenfein  
Cornell Tech

Ellen Goodman  
Rutgers Law

Chris Hoofnagle  
U.C. Berkeley School of Information

Amy Kapczynski  
Yale Law

Yafit Lev-Aretz  
Zicklin School of Business, City University of New York

Amanda Levendowski  
NYU School of Law

Susan Mann  
Microsoft

Terrell McSweeny  
Covington & Burling

Solomon Messing  
Facebook

Suzanne Munck  
Federal Trade Commission

Chaya Nayak  
Facebook

Helen Nissenbaum  
Cornell Tech

Paul Ohm  
Georgetown Law

Elizabeth O’Neill  
Cornell Tech

Julia Powles  
NYU Information Law Institute

Tanina Rostain  
Georgetown Law

Jule Sigall  
Microsoft

Gigi Sohn  
Georgetown Law

Julia Stoyanovich  
NYU Tandon School of Engineering

Maurice Stucke  
University of Tennessee

Frank Torres  
Microsoft

Helena Ursic  
Yale School of Law

Yvonne Wang  
Cornell Tech

Peter Yu  
Texas A&M University School of Law
A. GEORGETOWN INSTITUTE FOR TECHNOLOGY LAW & POLICY

The Institute for Technology Law & Policy at Georgetown Law is dedicated to training the next generation of lawyers and policymakers with deep, practical expertise in technology law and policy, and providing a forum to address pressing problems and opportunities involving technology and the law.

Among its initiatives, the Institute seeks to bring together policymakers, academics, business leaders and technologists to develop policy solutions informed by a deep understanding of technology. It also works to promote tech competency among policymakers, hosting an annual two-day immersion program for congressional staff to learn about new developments in technology policy, and workshops to support legislative staff and other experts. For more information visit www.georgetowntech.org.

B. CORNELL TECH DIGITAL LIFE INITIATIVE

Embedded within the progressive teaching mission of Cornell Tech on Roosevelt Island, the Digital Life Initiative (DLI) was launched in 2017 to analyze the societal tensions arising from existing and emergent digital technologies. Inspired by the core values of justice, democracy, privacy, responsibility, security, and freedom, we support collaborative research projects that explore ethics, policy, politics and quality of life within prevailing socio-technical systems. For more information visit https://www.dli.tech.cornell.edu/.

The convenors express their sincere thanks to Microsoft for supporting this event.