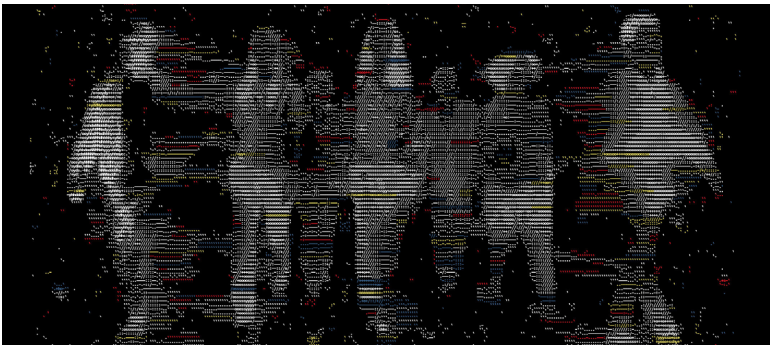
DATA AND DEMOCRACY



Data and Democracy

An Introduction

By Amy Kapczynski

KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY

In October 2020, the Knight First Amendment Institute at Columbia University convened a virtual symposium, titled "Data and Democracy," to investigate how technological advances relating to the collection, analysis, and manipulation of data are affecting democratic processes, and how the law must adapt to ensure the conditions for self-government. This symposium was organized by the Institute's 2019-2020 Senior Visiting Research Scholar, Yale Law Professor Amy Kapczynski, and co-sponsored by the Law and Political Economy Project at Yale Law School.

The essays in this series were originally presented and discussed at this two-day event. Written by scholars and experts in law, computer science, information studies, political science, and other disciplines, the essays focus on three areas that are both central to democratic governance and directly affected by advancing technologies and ever-increasing data collection: 1) public opinion formation and access to information; 2) the formation and exercise of public power; and 3) the political economy of data.

The symposium was conceptualized by Knight Institute staff, including Jameel Jaffer, Executive Director; Katy Glenn Bass, Research Director; Amy Kapczynski, Senior Visiting Research Scholar; Alex Abdo, Litigation Director; and Larry Siems, Chief of Staff. The essay series was edited by Glenn Bass with additional support from Lorraine Kenny, Communications Director; A. Adam Glenn, Writer/Editor; Madeline Wood, Research Coordinator; Chand Rajendra-Nicolucci, Research Fellow; Kushal Dev, Research Fellow; and Ashwin Pillai, Intern.

The full series is available at knightcolumbia.org/research/

INTRODUCTION

We live in the age of big data and algorithmic governance. Moore's law, operating over decades, has today enabled the collection, storage, and continuous updating of vast troves of data. Sensors and digital networks are ubiquitous: Our watches and search engines gather data that can reveal our health risks and moods.[1] Our pacemakers leave digital traces that insurers can access to investigate fraud.[2] Our phones capture and continuously transmit an almost unfathomable sea of data about our locations in space and the social world.[3]

Companies today, not only in the social media sector but in finance, health care, education, and beyond, increasingly understand that control over data, data processing, and the platforms through which data flow create critical opportunities for power and profit. The accumulation and processing of data have become "a core component of political economy in the 21st century," and data a critically important form of capital.[4] Some of the most powerful companies in the world today, for example, make their money not from selling consumer products, but by providing their core products for free, so that they can gather and monetize data, largely by deploying ever-richer profiles of individuals to advertisers who are willing to pay for it.[5] Digital

networks and sensors also enable new ways of acting on consumers still more directly, but at a new distance, promising not only more insight into customers but more control over them too. Auto lenders today can and do respond, for example, to missed payments by virtually repossessing cars, leaving drivers stranded in intersections.[6]

We have also recently entered what some call the era of big data and artificial intelligence, as new techniques have emerged to analyze large data sets, along with ubiquitous data-gathering technologies.[7] Machine-learning techniques developed in recent years, such as neural networks, enable the extraction of patterns or correlations in data sets to predict relationships or outcomes that are not otherwise discernable to humans.[8] Using these techniques, for example, programmers can now search databases of mammograms to detect relationships between images and health outcomes in order to predict, in ways that a human radiologist could not, which patients are most at risk.[9] The same techniques lie behind developments like self-driving cars and facial recognition software, and are used to tune our social media feeds, and so to shape our experiences, relationships, and political order. Yet while these evolving practices of data gathering and processing are spoken about everywhere, they are also in some ways fundamentally opaque. We know much less about those who gather the data—from the national security state to our health apps and social media companies—than they know about us. And most people have little sense of the vast array of digital traces and tracking practices that surround them.

Machine-learning techniques pose distinctive, deep secrecy challenges, because they create powerful new claims to knowledge that cannot be explained or readily audited. Yet it is increasingly clear that they are also prone to profound errors and are perceived as far more powerful and reliable than in fact they are. While big data and AI do enable dramatic new forms of power and knowledge, we are also in the grip of a certain breathlessness about what big data and AI will enable.[10] Machine-learning techniques are not, after all, in any real sense of the word, "intelligent." Rather, they are an artifact of both the data that underlie them and the choices of the humans who program them.[11] They are replete with human errors and biases, and introduce new and startling kinds of mistakes—ones that humans would never make. For example, an algorthim designed to identify COVID risk

predicted worse outcomes for patients whose scans used a particular font, because that font was used on charts at hospitals that happened to have sicker patients.[12] Machine learning and AI also inevitably reproduce artifacts and constructions of our social world from which they draw their insights. Algorithms designed, for example, to discern relevance by looking at how commonly someone's posts are "liked," or to discern risk of violence by mapping people to their zip codes and the criminal history of their parents, can only reproduce and entrench social realities of disadvantage and racism that are nothing if not constructed.

Our position within digital networks today profoundly shapes our life chances, in ways that we only dimly understand and that raise significant concerns for all of us. Scoring and sorting practices not only constitute our identities and access to social media but also shape our ability to access credit, employment, housing, and medical care. The implications are also structural. A new "big data divide" has emerged: "Those with access to data, expertise, and processing power are positioned to engage in increasingly sophisticated forms of sorting that can be 'powerful means of creating and reinforcing long-term [or newly generated] social differences.'"[13] Whether you are Uber or an Uber driver, a creditor or debtor, a prosecutor or defendant matters tremendously to how data are gathered and deployed, and how the benefits and risks are allocated. The key challenges that new technologies and forms of technopower pose today are, therefore, not merely challenges to privacy and autonomy, but challenges to a broadly conceptualized understanding of *democracy*, a form of government rooted in equally shared power.

An earlier period of optimism about the possibilities of these new forms of our data- and information-intensive age has thus given way to a mood of alarm and pessimism, even impending doom.[14] What is needed, however, is careful attention to the dynamics and implications of these developments. This can help us consider not only their risks but possible opportunities for greater and more effective democratic power over our data-intensive age.

In October 2020, the Knight First Amendment Institute and the Law and Political Economy Project collaborated to convene a conference centered on a distinctive set of questions about data and democracy: What challenges and opportunities, we asked, do processes of datafication and algorithmic governance hold for our ability to self-govern? How do the new forms

of technopower associated with datafication and algorithmic governance affect the institutions and processes of democracy that are the vehicle for the assertion of public interests in, and control over, our information age? How has the law shaped these new practices and forms of power, and how might legal reforms help us subject them to democratic processes and values?

This essay serves to introduce the superb series of articles that emerged from our discussions. They go beyond describing the conditions and risks that attend our data-intensive age, to encourage radical thinking about remedies and offer detailed proposals for reform. The work that follows is rich and varied, but several broad themes and insights emerge from the series, and can help orient us to the challenges to come.

DATA IN DEMOCRACY

THE FIRST THEME running through these articles is that data are not something merely wielded by democratic governments, but also something that *constitutes* them. Data are not simply used by democracies, but play a formative role in shaping and creating the *demos*, the "we the people" who are supposed to rule.

It has long been clear to scholars that modern democracy is data intensive. Modern data processing techniques can be found, for example, in earlier forms in government offices in the 19th century, where modern bureaucracy, as the "step-by-step, distributed and nominally objective procedures for selection and sorting," was born.[15] Data and information processing were central to acts of statecraft in the early 20th century, including for the codification of the birth certificate,[16] the institutionalization of systems of social insurance,[17] and the development of business practices such as the actuarial rating that made early 20th century insurance markets possible.[18]

Two papers in the collection, by Dan Bouk and danah boyd, and by Bertrall Ross and Douglas Spencer, trace the technopolitics of data gathering at what we might consider the beating heart of democracy: the act of counting and constituting the people. Their focus is, respectively, the census count that has, by constitutional mandate, happened at regular intervals since the founding of the United States, and the modern campaigning practices

through which parties cultivate voters, and shape the activities and voice of the voting public. Both show that data are not just deployed by the modern state, but recursively *constitute* the state, meaning that critical questions about data gathering and processing sit at the heart of democracy.

A core puzzle follows: How can data practices embody and embed democratic values, when they tend to appear both opaque and technical, and when they generate infrastructures of democracy that are not centrally thematized in politics and upon which the shape of politics itself must depend?

Modern democracy implies forms of statecraft that can sort and identify people and populations, both to shape how they are represented—what district their votes are counted in, for example—and how programs are designed and resources allocated. This is the technology of the U.S. census, and Bouk and boyd's remarkable article shows that as a technology, the census is, and has long been, anything but simple. Those in charge of the census must decide, for example, who and what to count, how much effort to exert to count the "undercounted," and who is allowed to use the resulting data for what purposes (and so how to trade off the utility of the census against privacy concerns that might implicate who is willing to be counted). As they show in rich detail, each of these questions has at different times been the subject of dramatic and bitter conflict.

While infrastructures that underlie the census are rarely the subject of public debate, Bouk and boyd's work shows that this is less because those infrastructures are not the subject of potentially explosive disagreement, than because these disagreements run so deep that they are suppressed in periods of ordinary politics. In these periods, new claims for more scientific ways to count are mobilized, only to later be brought back to earth by the conflictual nature of the democracy that surrounds it. Political theorist Claude Lefort argues that democracy is a system of government where the seat of power remains an "empty place," so that claims to who "the people" are and who properly represents them are always open to question. Bouk and boyd bring us an account of the way that technologies of data gathering and processing also inscribe an unstable, unlocatable center at the heart of the democratic project.

Does this insight mean that we cannot argue that there are better and worse ways, more and less democratic ways, to organize the census, because

all choices are equally "political"? This would be the wrong lesson to take from their account, which shows vividly how choices about counting can express ideologies of nativism, racism, and political self-dealing, or instead incline toward anti-racism and electoral accountability. There is ample room to distinguish between different modalities of counting, but on the basis of the commitments that they advance, rather than on an elusive technocratic neutrality.

It is not only in the census but also in the processes of regular elections and campaigns that data practices shape the core of our democracy. The implications for elections of new powers of data gathering and processing techniques have been the subject of much recent debate, particularly in the wake of the 2016 U.S. election. The Cambridge Analytica scandal, and the apparent efforts of foreign governments to sow discord on social media, brought home to many for the first time how algorithmic sorting and action might influence election outcomes. But these were just a subset of a much larger set of puzzles and problems that data-intensive techniques today generate for our elections and political campaigns. While politically and racially motivated gerrymandering practices, for example, have long been features of the American political landscape, today they can be far more exquisitely targeted.[19] Election security issues also loom large, and can impact elections in a variety of ways, from targeted attacks on voting infrastructure to attacks on the communications and data security of candidates.[20]

In their article, Ross and Spencer describe how political parties are deploying increasingly sophisticated data-driven strategies to mobilize and persuade voters. Their work parallels recent scholarship by Marion Fourcade and Kieran Healy that highlights the risks of scoring, sorting, and microtargeting in market settings. Fourcade and Healy stress that forms of scoring and rating designed to serve profit-making ends threaten not only our privacy but also reorganize our collective world, creating new categories and experiences.[21] "Companies seek good matches when they want to link a wealthy customer with a quality credit card, or a heavy drinker with a bad insurance plan. … Past and present social positions, social connections and ingrained behavioral habits shape not only people's desires and tastes but also the products and services pitched to them."[22] These processes create forms of social advantage and disadvantage, and "have reactive or

performative effects on individual behavior, on organizational strategy, and on people's life-chances."[23] Because individuals have preferences that are not fixed, but that change in response to their environment, feedback that is shaped around preferences may end up "deepening those differences by reinforcing the behaviors that caused them to be identified in the first place."[24]

Ross and Spencer illuminate some of these very same processes operating in the political domain, as parties and campaigns adopt tools of sorting and microtargeting that allow them to sort voters into high- and low-yield investments. As in markets, we can expect such sorting to have patterned implications, not only reflecting forms of hierarchy and disadvantage in the social world but also amplifying them. We might add to that a question drawn from Fourcade and Healy's work: How might not only voters but political messages and party platforms be influenced by the new ways of sorting that microtargeting enables?

The granular data needed for such sorting and targeting is, as their work describes, an unexpected consequence of legislation passed in the 1990s to facilitate voter registration and to protect against voter fraud. The law hardly solved the problem of access to the polls, and may unintentionally have laid the foundation for new forms of exclusion. As Ross and Spencer show, empirical analysis demonstrates that access to more granular voter data correlates with less party investment in mobilization of low-income voters, perhaps because they have lower rates of voting and are thus seen as poor investments. These dynamics would be self-reinforcing and are, they argue, likely responsible for decreased voter turnout among the poor and a growing inequality in political representation.[25]

Their account, along the way, sharply challenges the notion that data transparency has any natural relationship to democracy or political equality. In fact, it suggests the opposite. In an age of microtargeting and differential access to data, more access to data might entrench exclusion, further aggravating deep failures of democracy. In our data-intensive age, it is becoming easier to see that the ideal of democracy conflates two very different things: majority rule and rule of "the people's will." As Pierre Rosanvallon notes, "[T]he justification for power by the ballot box has always implicitly rested on the idea of a *general* will," a notion reinforced by the evolution of new claims of equality and rights.[26] At the core of democracy is a tension between

these ideas—the practical measures of representation and a set of ideals that enable challenges to those particularities. Microtargeting in political life, not only in voter mobilization but also in campaign speech and advertising, exposes and amplifies this disconnect, creating new questions about how evolving data systems may work against the ability of parties to serve as vehicles for solidaristic political projects.

Both papers show, in different ways, that the mere disclosure of more data, including about the inner workings of democracy itself, will not automatically translate into deepened democracy, or more trusted or trustworthy government. Trust is produced by something else, such as experiences of actual representation, and fair and effective administration, that in turn are contingent upon context and institutions, as well as a commitment to the democratic project as something more than optimizing the "market" for voters.

FROM DATA TRANSPARENCY TO DATA PUBLICITY

A **SECOND THEME** in the collection addresses the extraordinary challenges that secrecy poses for the democratic project in an age of big data. Read together, and in conjunction with the emerging literature on the subject, they make a powerful case that we cannot achieve the insights that we need into data and AI systems through a simple insistence on passive and unmediated "transparency." If access to data is to serve public ends, it will need to be *active*, *sensitive to underlying structures of power*, and in many cases, *conditional*.[27] In contrast to earlier optimism about the liberating potential of "open access" and unfettered transparency—open all the data! for everyone!—these papers set out what we might call the case for mediated and positive forms of data "publicity."[28]

Scholars have argued for some time that transparency or openness alone cannot create accountability and fairness in data and algorithmic systems.[29] We also cannot assume that in our data-intensive age, access to data translates readily into reliable information or knowledge. Instead, it can lead to the familiar experience of being overwhelmed by data, exhausted by accumulating demands for attention and interpretation. Data transparency

can be weaponized, not only against government[30] but also against responsible science itself.[31] The meaning of data sharing is shaped by the context that surrounds it, in other words. The implications of data and its uses are deeply intertwined with structures of power and networked processes dominated by those who have access to data sets and the technologies needed to manipulate and act upon them. Collectively, several of the papers that follow make the case for structured data publicity in different contexts that can, however, bring us closer to a more genuine democratic accountability over the design of data systems.

There is no doubt that secrecy in our data-intensive age poses profound challenges for democratic discourse and accountability. For democracy to be effective, citizens, and the experts and intermediaries upon whom citizens and governments rely, must have access to the data and information that they need to inform their judgments. In an increasingly data-intensive world, this becomes all the more important. One deeply troubling feature of the algorithmic age is its tendency to generate new or more serious obstacles to access to information about both government and the private sector.

Data and information flows today are mediated by processes that are fundamentally more opaque than their historical counterparts. One problem is simply the vast complexity of data-intensive processes. A layperson might have at one time been able to grasp enough about actuarial calculations or sentencing guidelines to have evaluated them, but it takes far more time and expertise to unravel their more data-intensive and algorithmically complex counterparts. The challenge is amplified by the fact that these practices can and do change frequently and recursively. Digital networks, by design, also encode asymmetric access to the information that flows through them: While a rider and driver in a cab might have had roughly similar access to information about their route in earlier periods, today the Uber app gathers a rich stream of data for the company but makes very little of that data available to either drivers or customers. Facebook and Google give users only limited insights into the data that are gathered about them. And government agencies that adopt new data-intensive practices, such as using algorithms to govern allocations to Medicaid recipients or employment evaluations for teachers, do not design them in a manner that allows those affected by these processes to have access to them.

These problems are amplified by the inherently opaque nature of AI processes that draw insights that cannot be readily explained, even if—as is rarely the case—outside experts had access to the algorithms, their outputs, and the underlying data sets on which they were trained.[32] Moreover, individuals may all be displayed different outputs, making reverse-engineering the inputs difficult. A reader once could open a national newspaper and see immediately what everyone else opening that paper would see, enabling shared critiques about what was elevated to the front page and what had been relegated to obscurity. We have only the most rudimentary insights into how the social media feeds of different people are constructed on platforms like Facebook, and the company has sometimes aggressively pursued journalists and researchers who have created new digital tools to help us understand the impact and nature of their targeting practices.[33]

Complexity and secrecy by design are only one aspect of the problem. As technological and governance processes have become more complex, data have become subject to stronger legal protection. Trade secret law has expanded in significant ways in recent decades, and today companies can argue that any valuable commercial information that is sufficiently secret is entitled to strong protection, not only from competitors but also from the public.[34] A watershed moment was the 1984 Supreme Court case *Ruckelshaus v. Monsanto*, which held that trade secrets could be a form of property protected from government taking by the Fifth Amendment.[35] Some lower courts, and many legislatures, have adopted broad interpretations of *Monsanto*'s rule, and struck down or neutered laws designed to inform citizens about everything from the ingredients of cigarettes to the composition of fracking liquids that are injected into the ground.[36]

The Freedom of Information Act regime has, at the same time, proven profoundly inadequate. FOIA's provisions have never extended to the private sector, creating asymmetries between what the public can know about government and about industry, which contributes to the hermeneutic of suspicion about government power.[37] FOIA does enable access to some important information held by governments, but many agencies take years to respond to anything but the simplest requests. Effective use of FOIA all too often also requires access to counsel and the resources to litigate, in part because agencies are deluged with requests from private sector actors who

crowd the queue and even deliberately intend to slow and obstruct agency work.[38]

FOIA also gives agencies the discretion to withhold requested data if it risks disclosure of confidential commercial information or interference with privacy. The former barrier, always formidable where corporate data was concerned, has become more acute after a June 2019 Supreme Court case, *FMI v. Argus Leader*.[39] That decision reversed long-standing doctrine that had authorized lower courts to order the release of information under FOIA unless harms to corporate interests were "substantial." It enables agencies to withhold any information that is merely confidential, if secrecy is the norm in the industry, or perhaps merely if it was promised by the agency.[40] Courts also often bar disclosure of information that would otherwise be made public in litigation, deferring to broad claims of trade secrecy or confidential commercial information in both the civil and criminal contexts. There is a clear tension between prevailing concerns about judicial management, and the vindication of the public's right to information and open courts.

Privacy is also becoming a much more powerful means to defeat requests for public information in the big data age, because computational science in the presence of vast and easily combined data sets makes reidentification of specific individuals much easier. Other, more subtle shifts in law have also made it easier for private and public actors to keep digital information and data secret. Our law has come to accept click-wrap contracts (contracts to which there is no meaningful ability to negotiate terms or register assent) and broad terms of use, both of which are regularly used to bolster the secrecy by design that digital intermediaries enjoy.[41] For example, firms like Facebook impose take-it-or-leave-it conditions on those who access the platform, forbidding them from gathering publicly available information in certain ways, even for research purposes. Compounding the problem, the Department of Justice and Facebook interpret the Computer Fraud and Abuse Act to make it criminal to violate a website's terms of service.[42]

The recent direction that the increasingly conservative Supreme Court has taken First Amendment law presents additional, profound challenges. Although there are clear reasons to think that effective free speech today might require affirmative rights to access at least some kinds of data, the Supreme Court has resisted the idea that the First Amendment encodes

any right to be informed. Courts have also not raised what might seem to be obvious First Amendment concerns triggered when striking down laws, for example, on takings grounds, that are designed to inform the public. Supreme Court cases expanding the domain of protected commercial speech are also giving companies new claims against "compelled" speech that many are seeking to use as a shield against not only laws that require corporate disclosure but also laws that regulate their products. In 2011, a company called IMS Health, which buys and sells prescription data, primarily to enable pharmaceutical companies to market more effectively, won a watershed case suggesting that not only are data "speech" protected under the First Amendment but also that restrictions that allowed doctors to opt out of the sale of their personal prescribing information—justified on both privacy and public health grounds—violated the Constitution.[43] Emboldened by cases such as these, Google and Apple, for example, have called upon the First Amendment to claim that their corporate speech rights are violated when they are required to unlock phones or rank search returns fairly.[44]

Action by platforms themselves could provide an important remedy, particularly as regards access to information and data needed to understand how the platforms work. There are ample reasons to be concerned that algorithmic processes designed to maximize engagement with these platforms may also amplify disinformation and heighten polarization. In response to these concerns, platforms like Facebook and Twitter have promised reforms, but it is difficult to hold platforms accountable without more granular access to information about the content that they are demoting and promoting. Because free speech protections make direct public regulation of these filtering decisions impossible or difficult, the stakes of publicity in this context are particularly important. As John Bowers, Elaine Sedenberg, and Jonathan Zittrain show in their paper, platforms do not routinely share this data with researchers, citing privacy concerns and concerns about enabling access to the very content that they believe is so toxic that it must be downranked or removed. Bowers, Sedenberg, and Zittrain draw on a model from post-Nazi Germany, the "poison cabinet" or Giftschrank, to show that firms can provide access to such material to researchers in a way that accommodates such concerns. For example, they might commit to creating an archive encompassing all information about COVID-19 that is downranked or removed from a site

as misleading, and enable access for trustworthy researchers. How could such a system address the structural conflicts of interest that platforms have as overseers of this archive, and would legislative action be needed to push platforms to enable access in this way? The authors suggest that there will be an important role for the public and for nonprofits, as well as for a commitment to resources to enable data access and use, in any successful Giftschrank model. How public mandates for this kind of archiving and sharing of information might run into corporate claims to trade secrets or free speech rights remain to be seen.

Mathias Vermeulen investigates a similar terrain from the European perspective. He, too, is focused on researchers' access to the data they need to analyze, identify, and redress misinformation online, and offers a cautionary tale about self-regulation. In Europe, voluntary measures have proven far too limited to provide the information and data that researchers most need. EU regulators have recognized this and included provisions in the General Data Protection Regulation (GDPR) and the proposed Digital Services Act (DSA) that Vermeulen argues can and should be operationalized to create a system for mandatory data sharing with trusted researchers, while remaining consistent with data privacy protection requirements. Mandatory measures are necessary, he argues, and his account offers important insights about how they might be designed consistent with Europe's data protection law. The GDPR is influential around the world, making the lessons drawn here of consequence for jurisdictions around the world.

Hannah Bloch-Wehba brings our focus back to government activity, with attention to the private vendors who supply AI tools to government, and on how to confront the opacity that these relationships engender. Such agreements are commonplace today, reflecting the expansion of public-private contracting in recent decades. The pressures that lead to "fissuring" in the private sector have led to similar "fissuring" and outsourcing in government, along with pressures to lower labor costs and contract in, rather than keep on staff the many diverse kinds of techical experts that might today be needed to realize data-intensive projects.[45] This interacts with open-government rules in problematic ways, as Bloch-Wehba describes: FOIA applies only to information in the control of the government, and in practice, vendors often have exclusive control over critical aspects of data-intensive and algorithmic

tools that they are supplying to government. In addition, they may enjoy contract terms that further insulate their activities and data from private view. Bloch-Wehba argues that data publicity can be achieved only through ex ante oversight and control over agreements with private sector companies, and urges reforms in procurement practices that would prioritize or even require publicity from private contractors. She also describes recent examples, including from the criminal justice context, where important experiments with similar approaches are underway.

Rebecca Wexler—who has done pathbreaking work to call attention to the implications of the uses of data-intensive technologies in criminal prosecutions for criminal defendants and their procedural rights to access evidence[46] —extends the discussion about data-disclosure in the criminal justice context. Her contribution illuminates asymmetries in how criminal defendants and the prosecution access data when those data are held abroad. In her description, as digital networks and technologies advanced, an asymmetry was built into the law of transnational data sharing, giving those governments acting in a prosecutorial role a far greater ability to access such data than those defending the accused. The article reminds us that growing private power over data does not *always* come at the cost of state power. Recent scholarship on the broad deregulatory trend of the 1980s and 1990s associated with "neoliberalism" has recently emphasized the state not as deregulatory but as re-regulatory.[47] Consider the dramatic rise, particularly in the United States, of a more punitive and extensive carceral state, expressed in both mass incarceration and via a militarized border, at the same time that ideologies of "small government" took hold.[48] Consistent with this insight, Wexler's account shows that different aspects of the state have been empowered, rather than sidelined or challenged, by rising private technopower.

Wendy Wagner and Martin Murillo take on the challenges of data publicity from a different perspective. Their focus is on the potential of new algorithmic systems developed and deployed by administrative agencies to enable more effective regulatory oversight in areas that have long been data intensive. Identifying toxic chemicals, assessing pollutant loads, and evaluating product safety all could be made easier by AI and other algorithmic techniques developed in recent years, they argue. Such tools must also

encode accountability by design, and Wagner and Murillo offer a structured series of best practices that should guide agencies in this work at each step in the process: algorithmic design, development, evaluation, and output. More troublingly, they show that existing administrative law systems are not well configured to enable these best practices, and in fact may discourage them.

STRUCTURAL APPROACHES TO DATA GOVERNANCE

WHAT STRATEGIES of governance might be capable of rendering highly networked forms of technopower accountable to public priorities, especially given the mismatch between growing private power over data and algorithmic systems, and the powers and resources of the public administrative state? These systems crisscross old industrial categories, and thus also the regulatory structures that grew up to facilitate and discipline them. (What is a health care company today, when an app on your phone or a social media platform may know before you do that you are pregnant or at risk of Alzheimer's disease?) Networked data systems also operate at a scale and a speed that seems to exceed what a reactive administrative state can hope to handle.

The prevailing architecture of administrative power over data-intensive firms is also ill-suited to the task. We have no generalized law of privacy in the United States, and no general regulatory requirements that might apply to the development of data-intensive products, beyond a few highly regulated sectors like pharmaceuticals. Data have been legally constructed as part of a "public domain," open for capture by those who can absorb and record them.[49] And structures like the Communications Decency Act, enforcement of click-wrap contracts and terms of use, and background patent, copyright, and trade secrecy law also have been woven together by firms to create layered private power over the networks through which data flow, as well as over data and algorithmic systems themselves.[50]

Early concerns about privacy online attracted only weak regulatory responses, formulated around a consumer deception model that could be remedied if consumers were better informed and "consented." The results

were broad terms of use and privacy policies that did not protect privacy, but effectively laundered legal consent despite persistent consumer concerns about privacy. Government surveillance too has been subject to a notice-and-consent model, because Fourth Amendment law centered consent as a value, authorizing both the giving over of data by the data subjects themselves, and government collection from private parties who had voluntarily given data to private sector actors like banks and telephone companies.

Julie Cohen, Frank Pasquale, and Aziz Huq and Mariano-Florentino Cuéllar all make powerful arguments for the inability of individualized, consent-based approaches to realize public values in this context. There is no effective way, for example, for apps and browsers to genuinely enable users, before they share their medical information or browsing history, to understand what might be done with their data. And tomorrow's possibilities cannot be predicted today. The logic of digital networked processes, as Cohen stresses, also operate at a scale that requires standardization, making them a poor fit with the logic of individual choice and consent. "Organizing a regulatory regime around individual control rights," she notes, "imports a governance structure that is atomistic and post hoc. Individual users asserting preferences over predefined options on modular dashboards have neither the authority nor the ability to alter the invisible, predesigned webs of technical and economic arrangements under which their data travels among multiple parties."

As Pasquale points out, a single entity or actor that breaks an agreement can also undo consent- and contract-based models of control. (This is merely the obverse of an influential argument for durable property rights like patents and copyrights that give exclusive control over information and are good against the world: Contracts were thought of as too weak, because it is too easy for control over digital information to slip the bounds of the contract.) Echoing arguments from the literature on the constitutive nature of privacy rights,[51] Huq and Cuéllar point out that data and information flows shape our identities and preferences. What does it mean to say that I agree to give up data, that I have chosen to doomscroll on Instagram, if my choices are a reflection in part of conditions that I encounter, which normalize and shape what seems legitimate, what seems compelling?

Structural approaches, not based on consent or organized through

logics of individual choice, and approaches that are capable of asserting and attending to the patterned forms of network power, are essential to democratizing our data-intensive age. Pasquale and Cohen argue for sharp departures from existing regulatory models. But these may take different forms, as each of these articles shows. Pasquale focuses on interventions like bans on particularly dangerous technologies and broad licensure requirements for certain algorithmic tools. Cohen draws regulatory inspiration from processes of stress-testing and auditing used in banking and consumer finance regulation, and argues that we must consider dramatic changes to our enforcement infrastructure, perhaps making intermediaries more responsible for the acts of third parties, and exploring criminal and disgorgement remedies. Huq and Cuéllar argue for a more contextual approach, founded in pragmatic judgments that are "tailored to the specifics of different institutions and environments." They are skeptical of reliance on any particular regulatory agency, and urge the need for grappling with the implications of AI systems at all levels and institutions of government. They also make an affirmative case for harnessing individuals as a source of power and private ordering, offering as examples the recent organizing of tech workers at Google, and tech-forward mutual aid efforts in the COVID-19 pandemic.

Kiel Brennan-Marquez and Daniel Susser argue that the emergence of the "platform" phase of capitalism calls into question the very existence of markets as we know them, and their relationship to both freedom and efficiency. Drawing upon the work of scholars like Shoshana Zuboff, they contend that the kinds of surveillance and behavioral influence made possible by new technologies offer new reasons to doubt that markets will enhance freedom. These same technologies also offer new ways to determine what people want and need. Since Hayek's formative work, markets have been defended as a means to harness distributed information—better to have Ford, responding to market signals, predicting the cars that we want, than the government, so the theory goes. But if retina scans and your Google search history can forecast your consumer needs, can planning operate on a new, information-rich footing, in vastly more efficient fashion? They suggest that AI and big data might enable new ways to guide production to better meet essential needs, though not necessarily entirely without markets. They urge us to abstract away from some of the more immediate problem-solving we

are doing, to think broadly about possible futures, from platform feudalism, to digital socialism, to things in between.

The interventions in this series, in different ways, all help to illuminate the fundamental choices that we have in front of us. New technologies and markets do not follow their own laws, but respond to the decisions that we make. So what kind of markets, what kind of regulatory state, what kind of data publicity, what kind of democracy will we make with these new technologies? That is up to us, but the debates and analysis in this collection offer critical insights and arguments to illuminate our way.

NOTES

1 *See* Lori Andrews, *A New Privacy Paradigm in the Age of Apps*, 53 Wake Forest L. Rev. 421 (2018).

2 *See, e.g.*, Beth Mole, *Ohio man's pacemaker data may betray him in arson, insurance fraud case*, Ars-Technica (Feb. 8, 2017, 10:01 AM), https://arstechnica.com/science/2017/02/ohio-mans-pacemaker-data-betrays-him-in-arson-insurance-fraud-case/ [https://perma.cc/3JAM-ZATR].

3 The New York Times has developed an interactive series that helps to visualize and understand the extent of these powers. *See, e.g.*, Charlie Warzel & Stuart A. Thompson, *How Your Phone Betrays Democracy*, N. Y. Times (Dec. 21, 2019), https://www.nytimes.com/interactive/2019/12/21/opinion/location-data-democracy-protests.html [https://perma.cc/Z44L-ENBX].

4 Jathan Sadowski, *When data is capital: Datafication, accumulation, and extraction*, 6 Big Data & Soc'y 1, 1 (2019).

5 Shoshana Zuboff, The Age of Surveillance Capitalism (2019) (describing this, largely though a study of the evolution of Google and Facebook).

6 Rebecca Crootof, *The Internet of Torts: Expanding Civil Liability Standards to Address Corporate Remote Interference*, 69 Duke L. J. 583, 585 (2019).

7 *See, e.g.*, danah boyd & Kate Crawford, *Critical Questions for Big Data*, 15 Info., Commc'n & Soc'y 662 (2012); M. C. Elish & danah boyd, *Situating methods in the magic of Big Data and AI*, 85 Commc'n Monographs 57 (2018).

8 David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. Davis L. Rev. 653 (2017).

9 *See, e.g.*, Ziad Obermeyer & Ezekiel J. Emanuel, *Predicting the Future—Big Data, Machine Learning, and Clinical Medicine*, 375 New Eng. J. Med. 1216, 1218 (2016); Alejandro Rodriguez-Ruiz et al., *Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison with 101 Radiologists*, 111 J. Nat'l Cancer Inst. 916 (2019).

10 *See* Elish & boyd, *supra* note 7.

11 *Id.* at 58 ("[T]he logics, techniques, and uses of these technologies can never be separated from their specific social perceptions and contexts of development and use.").

12 Will Douglas Heaven, *Hundreds of AI tools have been built to catch covid. None of them helped*, MIT Tech. Rev. (July 30, 2021), https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/ [https://perma.cc/GRL8-H98B].

13 Mark Andrejevic, *Big Data, Big Questions|The Big Data Divide*, 8 Int'l J. Commc'n 1673, 1676-77 (2014).

14 *See, e.g.*, Zuboff, *supra* note 5.

15 Marion Fourcade & Kieran Healy, *Seeing Like a Market*, 15 Socio-Economic Rev. 9, 10 (2017).

16 Colin Koopman, How We Became Our Data (2019).

17 François Ewald, The Birth of Solidarity (2020).

18 Daniel B. Bouk, How Our Days Became Numbered (2016).

19 The same techniques can also be used, some argue, to challenge gerrymandering and develop new processes for defining fair districting practices. *See* Louise Matsakis, *Big Data Supercharged Gerrymandering. It Could Help Stop It Too*, Wired (Jun. 28, 2019), https://www.wired.com/story/big-data-supercharged-gerrymandering-supreme-court/ [https://perma.cc/PQ8C-YNUB].

20 *See, e.g.*, Karl Manheim & Lyric Kaplan, *Artificial Intelligence: Risks to Privacy and Democracy*, 21 Yale J.L. & Tech. 106 (2019).

21 Fourcade & Healy, *supra* note 15, at 10.

22 *Id.* at 17.

23 *Id.* at 22.

24 *Id.* at 23.

25 *See also* Bertrall L. Ross II & Douglas M. Spencer, *Passive Voter Suppression: Campaign Mobilization and the Effective Disenfranchisement of the Poor*, 114 Nw. U. L. Rev. 633 (2019).

26 Pierre Rosanvallon, Democratic Legitimacy (2011).

27 These points resonate with the arguments in Christopher J. Morten & Amy Kapczynski, *The Big Data Regulator, Rebooted: Why and How the FDA Can and Should Disclose Confidential Data on Prescription Drugs and Vaccines*, 109 Calif. L. Rev. 493 (2021).

28 *Id.*

29 *See, e.g.*, Mike Ananny & Kate Crawford, *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability*, 20 New Media & Soc'y 973 (2018).

30 David E. Pozen, *Transparency's Ideological Drift*, 128 Yale L. J. 100 (2018).

31 For a discussion of the example of the April 2018 "Transparency Rule" proposed by Trump's EPA, see Thomas O. McGarity & Wendy E. Wagner, *Deregulation Using Stealth "Science" Strategies*, 68 Duke L.J. 1719, 1767-68 (2019).

32 *See, e.g.*, Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. Pa. L. Rev. 633, 638 (2017).

33 *See, e.g.*, Surya Mattu et al., *How We Built a Facebook Inspector*, The Markup (Jan. 5, 2021), https://themarkup.org/citizen-browser/2021/01/05/how-we-built-a-facebook-inspector [https://perma.cc/ZP3B-MFXM]; Issie Lapowsky, *Platforms vs. PhDs: How tech giants court and crush the people who study them*, Protocol (Mar. 19, 2021), https://www.protocol.com/nyu-facebook-researchers-scraping [https://perma.cc/RR2V-C4VP]; Cory Doctorow, *Facebook Is Going After Its Critics in the Name of Privacy*, Wired (Nov. 20, 2020, 8:00 AM), https://www.wired.com/story/facebook-is-going-after-its-critics-in-the-name-of-privacy/ [https://perma.cc/6M7B-UCPS].

34 Amy Kapczynski, *The Public History of Trade Secrets*, U.C. Davis L. Rev. (forthcoming 2022) (on file with author).

35 Ruckelshaus v. Monsanto, 467 U.S. 986 (1984).

36 Kapczynski, *supra* note 34.

37 David E. Pozen, *Freedom of Information Beyond the Freedom of Information Act*, 165 U. Pa. L. Rev. 1097 (2017); Pozen, *supra* note 30.

38 Margaret B. Kwoka, *FOIA, Inc.*, 65 Duke L. J. 1361 (2016).

39 Food Mktg. Inst. v. Argus Leader, 139 S. Ct. 2356 (2019).

40 *Id.*

41 Julie Cohen, Between Truth and Power (2019).

42 It remains to be seen whether they will continue to do so in light of the Supreme Court decision in Van Buren v. U.S., 593 U.S. __ (2021). *See also* Knight First Amendment Institute at Columbia University, *Knight Institute Calls on Facebook to Lift Restrictions on Digital Journalism and Research*, Knight First Amendment Inst. Colum. U. (Aug. 7, 2018), https://knightcolumbia.org/content/knight-institute-calls-facebook-lift-restrictions-digital-journalism-and-research [https://perma.cc/3BJP-GT2D].

43 Sorrell et al. v. IMS Health Inc. et al., 564 U.S. 552 (2011).

44 *See, e.g.*, Search King, Inc. v. Google Tech., Inc., No. CIV-02-1457-M, 2003 WL 21464568, at *4 (W.D. Okla. May 27, 2003); Apple Inc's Motion to Vacate Order Compelling Apple Inc. to Assist Agents in Search, and Opposition to Government's Motion to Compel Assistance, *in the Matter of the Search of an Apple iPhone Seized During the Execution of a Search Warrant on a Black Lexus Is300, California License Plate 35kgd203*, 2016 WL 2771267 (C.D.Cal. Feb. 25, 2016).

45 Ronald J. Gilson et al., *Contracting for Innovation: Vertical Disintegration and Interfirm Collaboration*, 109 Colum. L. Rev. 431 (2009); Cohen, *supra* note 41.

46 Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 Stan. L. Rev. 1343 (2018).

47 *See, e.g.*, Jamie Peck, Constructions of Neoliberal Reason (2010); Quinn Slobodian, Globalists (2018).

48 *See, e.g.*, Bernard E. Harcourt, The Illusion of Free Markets (2012).

49 Cohen, *supra* note 41.

50 *Id.*; Amy Kapczynski, *The Law of Informational Capitalism*, 129 Yale L. J. 1460 (2020).

51 Robert C. Post, *The Social Foundations of Property: Community and Self in the Common Law Tort*, 77 Calif. L. Rev. 957, 959 (1989); Paul M. Schwartz, *Privacy and Democracy in Cyberspace*, 52 Vand. L. Rev. 1609, 1664–66 (1999).

About the Author

Amy Kapczynski is professor of law at Yale Law School. She is the author of numerous journal articles and essays, including "The Lochnerized First Amendment and the FDA: Toward a More Democratic Political Economy," which was published online by the Columbia Law Review in connection with the Knight Institute's 2018 symposium, "A First Amendment for All? Free Expression in an Age of Inequality." She co-directs Yale University's Global Health Justice Partnership, the Collaboration for Research Integrity and Transparency, and the Law and Political Economy Project. In 2017, she co-founded the Law and Political Economy blog, which grew out of her seminar of the same name. Prior to joining the Yale faculty in 2012, Kapczynski taught at the University of California, Berkeley, School of Law. She also served as a law clerk to Justices Sandra Day O'Connor and Stephen G. Breyer at the U.S. Supreme Court, and to Judge Guido Calabresi on the U.S. Court of Appeals for the Second Circuit. She was a senior visiting research scholar at the Knight First Amendment Institute in 2019-2020.

© 2021, Amy Kapczynski.

About the Knight First Amendment Institute

The Knight First Amendment Institute at Columbia University defends the freedoms of speech and the press in the digital age through strategic litigation, research, and public education. It promotes a system of free expression that is open and inclusive, that broadens and elevates public discourse, and that fosters creativity, accountability, and effective self-government.

knightcolumbia.org

Design: Point Five
Illustration: ©Erik Carter



KNIGHT
FIRST AMENDMENT
INSTITUTE at
COLUMBIA UNIVERSITY