Emerging Directions in the Data-Society Interface

Nicole P. Marwell and Cameron Day

THE UNIVERSITY OF CHICAGO
CROWN FAMILY SCHOOL OF SOCIAL WORK, POLICY, AND PRACTICE
Advancing a More Just and Humane Society

Robert Wood Johnson Foundation

ACKNOWLEDGMENTS

Support for this report was provided by the Robert Wood Johnson Foundation. The views expressed here do not necessarily reflect the views of the Foundation. We thank the Foundation and the University of Chicago's Crown Family School of Social Work, Policy, and Practice for enabling this work. We especially appreciate the collaboration of Foundation senior program officer George Hobor, who was an early, enthusiastic supporter of our vision and helped bring it to fruition. Additional thanks to Dr. Florencia Torche for her review of key sections of the report. Finally, thanks to the Public Interest Technology University Network for supporting the first author's creation of a course at the University of Chicago in collaboration with Dr. Christopher Berry. Some of the ideas presented in this report were initiated during the development of that course, and we thank Dr. Berry especially for helping to shape the ideas we present in the section on data analysis.

AUTHOR INFORMATION

Nicole P. Marwell is Associate Professor at the Crown Family School of Social Work, Policy, and Practice at the University of Chicago. She is also Associate Faculty in the Department of Sociology, and an Affiliated Scholar at the Data Science Institute. Nicole's research examines questions of urban governance, with a focus on how nonprofit organizations, government bureaucracies, community organizing, and formal politics shape cohesion, inequality, and exclusion in contemporary cities. In one line of inquiry, she asks how legitimate knowledge is constructed, and then deployed for governance purposes by state and non-state actors. Nicole holds a Ph.D. in sociology from the University of Chicago, and previously held faculty positions at Columbia University and the City University of New York.

Cameron Day is a PhD student in the University of Chicago Department of Sociology. Cameron researches histories of institutions of social control in the United States, with a particular focus on the intersection between law, science, politics, and media in efforts to control violence. He has previously worked on projects involving spatial concentration of urban public health crises, the application of clinical science to social problems by nonprofit organizations, and the intellectual history of risk assessment tools in the criminal justice system.

Contents

Introduction

Today's world is experiencing an explosion of new data and analytic methods, and as a society we are only just beginning to understand the consequences for reshaping social life. As new forms of data and analysis affect practice in fields as diverse as criminal justice, agriculture, social services, health, education, disaster management, and city governance, scholars and thought leaders are examining what we refer to in this report as the "data-society interface."

> **"Data-society interface"**
> We use this term to call attention to the ways in which the availability of new data, more data and new analytic methods are changing the social practices of individuals, groups, organizations and societies.

It is urgent that we examine the research that has emerged over the last decade on how data is shaping our society as this data is being used in real-time to inform decisions and allocation of resources. This includes efforts to critically examine and understand the consequences of the "datafication" of society when the data guiding our decisions is produced through the lens of human decision-making. Data is often portrayed as an objective, neutral process of representing reality. It is critically important, however, to recognize the human labor and decision-making that goes into producing data. As we increasingly use data to make claims about how to act in the world, we must not lose sight of the disconnect between data and the reality it represents.

This report aims to illuminate key issues posed by the use of data to inform priorities and decisions in our society, with an introduction to key discourses and findings in the study of the data-society interface. The report offers a guide to help us critically examine the ways we understand and use data and reminds the reader that we need to treat data as a tool in service to human contemplation, assessment, and decision-making, rather than accepting data as "truth" without the necessary human debate and discussion of attendant ethics and politics. There continues to be great promise for using data to solve social problems, if our institutions and decision-makers are thoughtful about how and why we use data, how we make meaning from it, and how we can hold these processes accountable to our values and serve our broader collective goals for improving society.

We begin by interrogating what seems to be the simplest of questions: what are data? We then draw on distinct theoretical approaches to the data-society interface to pose three overarching questions that users of data and data analytics should be asking. A key epistemological point underlies all of these questions: **If we continue viewing data and analytics as value-free and objective, we miss the chance to understand the ways this technology carries social and political choices, costs, and benefits.** The report then summarizes and considers key distinctions among the major analytical approaches to data: descriptive, causal, and predictive. In an extended example regarding the social problem of mass incarceration, we show how each of these approaches has offered insights to inform action. We conclude the report with a discussion of data ethics, including questions of bias and fairness in the use of algorithmic decision-making.

What are Data?

The last decade has seen the rapid rise of what is often called "datafication"—the rendering of nearly all transactions, images, and activities into digital representations that can be stored, manipulated, and analyzed through computational processes.

- Datafied *transactions* include individuals' purchases at retailers, restaurant inspections carried out by local government employees, or use of food stamps and other human services. While shopping, restaurant inspecting and accessing human services are not new activities, in our time of datafication they now leave behind "digital exhaust" (the digital traces of human activities such as computerized record-keeping or interactions with websites) that can be easily collected, retrieved, examined, and combined with other data sources in ways that analog records (records kept on paper or other physical materials) cannot.

- Datafied *images* run the gamut from satellite imagery of city blocks to security camera footage of building entrances to geotagged photographs uploaded to photo-sharing websites. With new forms of computational analysis, these images can be repurposed in unanticipated ways, such as to train models of computer vision, for use by law enforcement seeking criminal suspects, or to pick up signals of gentrification by tracking geotagged Instagram images of where people eat out at restaurants.

- Examples of *activities* that can be datafied are myriad, including daily travel patterns traceable through cell phone location data; food preferences, which can be revealed through online restaurant reviews; or levels of physical fitness and health, which can be implied by data gleaned from workout tracking apps. These datafications may be used by individuals to pursue self-understanding, painting a picture of their "quantified selves."[1] They also may be used by corporations, governments, and other organizations for marketing, social planning, or surveillance, either at the individual level or aggregated across collectivities, such as place or population group. For example, the COVID-19 pandemic prompted ideas for using activity data for public health, such as attempting contact tracing through the use of cellphone location data.

[1] Wolf, Gary. 2009. "Know Thyself: Tracking Every Facet of Life, from Sleep to Mood to Pain, 24/7/365." Wired; Swan, Melanie. 2013. "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery." Big Data 1(2):85–99; Lupton, Deborah. 2016. "The Diverse Domains of Quantified Selves: Self-Tracking Modes and Dataveillance." Economy and Society 45(1):101–22.

FEATURED SOLUTION

Community Noise Lab in Mississippi

Community Noise Lab, located at the Brown University School of Public Health, is conducting a comprehensive and large-scale environmental exposure assessment in the Jackson Metropolitan area of Mississippi, democratizing the collection and analysis of environmental exposure data. Cardiovascular disease is the leading cause of death in the state and cardiovascular mortality is the highest rate in the nation, with African Americans in Mississippi at higher risk for cardiovascular disease and death than White Americans. This is the largest single-site community-based epidemiologic investigation of environmental and genetic factors associated with cardiovascular disease among African Americans ever undertaken. The research will create equitable and community-centered exposure assessments for each community, measuring noise, air, and water pollution. It will also examine the potentially far-reaching exposure to misclassification and equity issues in traditional environmental health studies—due to reliance on environmental data from national models—as opposed to models developed from community-scale environmental assessments. The lab's primary aim is to holistically explore the relationship between environmental exposures and health by working directly with communities to support their issues using real-time monitoring and exposure modeling; a smartphone app that allows users to objectively measure and subjectively describe exposure events in their community to better understand and address inequities.

Implications of the "Datafication" of Society

Datafication exerts increasing influence over how we conduct business, develop public policy, exercise political voice, form social relationships, and engage in many other aspects of contemporary living. Indeed, many of today's efforts to shape the social world begin with a data enterprise: by collecting and analyzing data, we draw others' attention to particular opportunities or problems.

> **Datafication** is the rendering of nearly all transactions, images, and activities into digital representations that can be stored, manipulated, and analyzed through computational processes.

Data then become part and parcel of how we decide to act in the world. As our ability to compile data has become ever more enhanced by technological developments, we have been increasingly willing to take data as holistic and unproblematic representations of the world. Wide-ranging data production processes have led us to feel more and more that we can only apprehend reality through data representations.

This increasingly taken-for-granted assumption conflates datafied representation with the truth of the world itself. This assumption that data represents reality is the conceptual underpinning of the promise that big data enthusiasts see in datafication: an ability to clearly view, and thus to assess and intervene in, the state of the world. However, this assumption is rooted in the belief that the data produced through datafication are "raw"—unadulterated representations of reality.[2] In this view, the products of datafication are free from the bias and error that data based on human *reporting* of action (such as in surveys or narratives) may contain.

Data, however, are never "raw" or unbiased. As critical data studies scholars remind us, data are always produced through the lens of human decision-making.[3] A common analogy used to explain this idea is to think about the relationship between a map of a territory and the territory itself. While the map may represent many key aspects of the territory, such as its forests, bodies of water, settlements, and so on, the map is limited in its representation. The map does not capture everything about the territory, but only those aspects of the territory that the human mapmaker saw fit to record. And so it is with data: **data is a representation of the world, but that representation is only**

[2] Gitelman, Lisa. 2013. "Raw Data" is an Oxymoron. Cambridge: MIT Press; Kitchin, Rob. 2014. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. London: SAGE.

[3] Gitelman 2013; Gillespie, T. 2014. "The Relevance of Algorithms." In: Gillespie, T., Boczkowski, T.J., & Foot, K.A. (eds.) Media Technologies: Essays on Communication, Materiality, and Society. MIT Press; Ananny, M. & Crawford, K. 2018. "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability." New Media & Society, 20: 973-989; Brunton, F. & Nissenbaum, H. 2015. Obfuscation: A User's Guide for Privacy and Protest. MIT Press; Ziewitz, M. 2017. "A not quite random walk: Experimenting with the ethnomethods of the algorithm." Big Data & Society 1.

partial, and has been assembled through human choices about what (and what not) to record.

A real-world example can be found in patients using patient ratings and reviews to guide their healthcare decisions. Studies have found, however, that the reviews featured on healthcare organizations' websites tend to be biased toward positive reviews, while patient reviews as a whole tend to be biased against physicians of color. This example illustrates that while patient reviews are important data, it is critical to examine how they are a partial representation of reality and thus subject to human choices and potential bias about what to record. This recognition should inform decisions about, for example, how patient ratings should (or should not) be incorporated into healthcare organizations' government reimbursement rates or decisions about salary and raises for individual physicians.

The Friction Behind Data Production

Every tool that we have to record data has been built by humans. This means that even for the most passive-seeming data collection, we already have built certain assumptions into our data collection strategy. For example, if we are seeking to measure air quality, we already have determined that carbon dioxide is the right variable to capture, that we know the threshold level that matters, that the location of our air quality sensor is appropriate for measuring carbon dioxide in a particular place, and so on. To recognize that all these decisions have been made, by humans, before data collection can occur is to challenge the notion that datafication automatically captures the world in an unmediated, frictionless way. Because data are not "captured." They are not simply a representation of action that occurs in the world. Instead, data are *produced*. They are a human creation. Data represent human choices about which elements of the world we want to look at, and which ones we don't.[4]

We are not always aware of these choices in the process of coming to agreement about which pieces of our world are important enough to represent and record. These choices are made by social groups that share common culture, norms, or ways of doing things such as organizations and communities. All data perform the task of reducing complexity, allowing the world's activities to be represented, manipulated, and analyzed, and for people to draw conclusions for action. How are these decisions made? How does the existence of particular forms and sets of data reveal the norms and values of the social groups that produced them, as well as their judgments about which parts of the world are important enough—or simply feasible—to represent and track? By examining these questions about the social groups that produce data, we can learn about which situations and issues are judged as important enough for and amenable to human intervention—and which are not.

FEATURED SOLUTION

AISP's Toolkit for Centering Racial Equity Throughout Data Integration

The Actionable Intelligence for Social Policy program at the University of Pennsylvania recognizes cross-sector data sharing and integration can help create actionable intelligence that can be used to understand community needs, improve services, and build stronger communities. Yet, cross-sector data can also reinforce legacies of racist policies and produce inequitable resource allocation, access, and outcomes. This raises fundamental concerns, as administrative data increasingly are used as inputs for evaluation, research, and risk modeling that inform policy, resource allocation, and programmatic decisions. Since 2019, AISP has led a diverse workgroup of civic data stakeholders to co-create strategies and identify best practices to center racial equity in data integration efforts, resulting in the development of a *Toolkit* "to create a new kind of data infrastructure—one that dismantles 'feedback loops of injustice' and instead shares power and knowledge with those who need systems change the most." They are currently working with organizations across the country to transform state and local approaches to data infrastructure and create tools for driving community-led research agendas that lead to more equitable health and social outcomes.

[4] Bowker, Geoffrey and Susan Leigh Star. 1999. *Sorting Things Out: Classification and its Consequences*. Cambridge: MIT Press. Kitchin 2014; Gitelman 2013.

FEATURED SOLUTION

Intersectionality as a Lens for Revising OMB Standards on Race

The Institute for Study of Race & Social Justice at the University of New Mexico, led by Nancy López, director and co-founder, is working to propose revisions to including intersectionality in federal administrative data collection and reporting through the U.S. Office of Management and Budget (OMB) standards, as well as the Census and other administrative data. They propose including separate questions on self-identified race and *street race*: "how people are perceived (and treated) in our society, regardless of who they are, how they feel and how they self-identify." The authors write, "We can all play a part, large and small, in changing the research questions, the national and local conversations and the national narrative about race in the Latina/o/x community in the U.S. and beyond. The future of race and social justice for Latina/o/x and other disadvantaged communities in the years to come depends on our willingness to transform the status quo of data collection on race and ethnicity, to practice solidarity and action for social justice and advance liberation. As the current administration considers revising the Office of Management and Budget guidelines, we invite you to join us in asking if a street race question can be included as an additional "gold standard" that will help us document and rectify racial inequity." See the study for *sample question formats* for intersectional data collection.

The Work of Data Standards

A key component of the process of data production involves the creation of data standards: unified conventions for representing aspects of the world as data. The benefit of such standards is that they enable communication about particular objects or attributes across different social worlds, allowing people to act on those objects from their different positions.[5] For example, in the United States, data standards for race and ethnicity have been established by the Census Bureau, and are used widely by researchers, judges, policymakers, and others in the regular performance of their work. At the same time, the meaning of these standards is far from self-evident, and their use requires tacit acceptance of the historical meanings sedimented within them.

For the first 60 years of the U.S. Census, population categories functioned primarily to distinguish free people from enslaved people; freedom equated to whiteness, slavery assumed blackness, and a category of "other free people" accounted for the few who fit neither of these linked norms. Starting in the mid-19th century, scientific interest in establishing claims of biologically-derived racial hierarchy led to more fine-grained categories for "color" or "race," feeding into the nation's evolving racial order. Throughout the 20th century, these categories continued to shift, although the move to self-identification of race and ethnicity in 1960 recognized the fundamentally social—rather than biological—character of race. Nevertheless, while the idea of race as a biological reality generally has been rejected by scientists, the residue of this older way of thinking remains entrenched in present-day data standards for race. As we continue to gather data using this standard, it carries forward its pseudoscientific origins, with potentially harmful results.

For example, when algorithmic decision-making is used in clinical medicine, a patient's race often is included in a set of diagnostic predictors that determine whether or not a particular treatment is recommended. Recent studies have shown, however, that such algorithms can require black patients to be sicker than white patients before treatment is recommended.[6] This phenomenon results from a failure to adequately specify *how* race might be contributing to the presentation of disease in a patient. The heritage of our racial data standards assumes that these categories contain reliable biological or genetic distinctions. Even though few contemporary scientists would agree with this assumption, moving past it is difficult, especially given that race is known to be consistently correlated with a range of medical conditions and outcomes.

This example illustrates the process by which data standards absorb and reconstruct aspects of the social collectivities that produced them. In the case of the Census, the racial classifications we

[5] Star, Susan Leigh and James R. Griesemer. 1989. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." Social Studies of Science 19: 387-420. Bowker and Star 1999.

[6] Obermeyer, Ziad, Brian Powers, Christine Vogeli and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." Science 366: 447-453. Vyas, Darshali A., Leo G. Eisenstein and David S. Jones. 2020. "Hidden in Plain Sight -- Reconsidering the Use of Race Correction in Clinical Algorithms." *New England Journal of Medicine.*

consider standard today passed through a prolonged period of social struggle, encoding different aspects of our racial order over time, such as enslavement, citizenship, the biological theory of race, and so on. We have increasingly discounted the original thinking that inspired these categories such that today, many users of Census racial categories agree that they have no biological basis. Nevertheless, we continue to find them useful for understanding group-based social trends, such as racial disparities in employment, income, health, or educational attainment.

While this usage can help develop understanding of current social conditions, the appropriation of this data standard for predictive analysis about individuals reveals its limitations. Absent any real biological basis, why should race be used as a factor in determining an individual's medical treatment at all? In fact, frequently observed associations of medical conditions with race can be traced to other, mostly social factors, for which race is a proxy in the U.S.: poverty, lack of medical care, low-quality housing, and others. Data standards obscure these complexities, and knowing this should remind us of the importance of interrogating standard practices. The relative legibility of race as a problematic data standard has made it one of the first such practices to be subjected to critical review, but we should be mindful of seeking out similar shortcomings within other data standards as well. For example, data standards regarding the classification of cause of death have been strained recently by both the opioid crisis and the COVID-19 pandemic. Opioid-related deaths may be undercounted because of reporting standards that allow use of the broad category "unspecified drugs," rather than the narrower category of "opioids," as a contributing factor in these deaths.

Similarly, the ability to classify deaths as attributable to causes other than COVID-19 has led to a significant undercount of deaths from the pandemic; in this case, analysts offer social explanations of this phenomenon that range from staffing levels of coroner's offices that are too low to fully investigate every death, to political goals of minimizing public perceptions of the severity of COVID-19.[7,8]

SUMMING UP

Datafication often is portrayed as an objective, neutral process of representing reality. It is critically important, however, to recognize the human labor and decision-making that precedes the arrival of data in the world. The social groups, such as organizations and communities, that produce data and data standards are helping to produce the reality that data aspires to represent. As we increasingly use data to make claims about how to act in the world, we must not lose sight of the friction that underlies our representations.

FEATURED SOLUTION

Fairness in Precision Medicine

Data & Society, an independent nonprofit research institute, is exploring *Fairness in Precision Medicine* – a growing field that uses multiple data inputs from genetic information to electronic medical records to tailor medical care to individuals. This project aims to "critically assess the potential for bias and discrimination in health data collection, sharing, and interpretation." A keystone *report* coming out of the project recommends that researchers explicitly focus on recruiting and actively engaging diverse patients as active participants in medical research in response to the historical lack of representation in medical research. Authors also recommend considering geographic and socioeconomic diversity and continental ancestry alongside race when collecting health data.

[7] The Documenting COVID Project and USA Today Network, "Uncounted: Inaccurate death certificates across the country hide the true toll of COVID-19, USA Today, December 21, 2021.

[8] Documenting COVID-19, April 17, 2022. Accessed May 18, 2022 at *www.documentingcovid19.io*.

Three Key Questions About the Use of Data in Society

As research about the data-society interface has developed in multiple disciplinary and interdisciplinary contexts, scholars have posed distinct questions to which they have addressed their inquiries. This section identifies and discusses three key questions found in this diverse scholarship.

QUESTION 1

How does the use of new data and analytic methods by corporations, governments, and other organizations reset the boundary between these actors' efforts to shape the choices and opportunities we face, and individuals' desires for equity, freedom, and privacy?

This question is a central focus of the field of "surveillance studies."[9] Scholars working in this area draw important theoretical inspiration from the philosopher and historian Michel Foucault, who developed key ideas about how the rationalization of modern society provides governments with tools to monitor, channel, and direct the activities and beliefs of their citizens.[10] Such tools often depend on the quantification of complex human behaviors, i.e., the creation of data that become the basis for subsequent analysis and decision-making. Scholars have argued that an overreliance on quantification excises core human characteristics like dignity, individuality, and emotion from those decisions, but governments have long drawn on such tools in their efforts to manage groups, places, organizations, and societies.

Surveillance studies scholars contend that in the current moment, for-profit corporations are equally if not more important than governments as developers and users of large-scale quantification and surveillance techniques, and that these innovations pose important risks to freedom, privacy, and democracy. For example, tools developed by social media platforms allow advertisers to "microtarget" individuals—that is, to classify individuals based on their online behaviors, such as which websites they visit online, how much time they spend looking at specific webpages, or which products they buy. Recent press and scholarly investigations have shown that housing, employment, and credit companies can use microtargeted advertising to exclude certain groups of consumers in ways that violate federal equal protection laws.[11] This might mean preventing people who spend time on African-American-oriented websites from seeing ads about available housing in majority-white neighborhoods; such practice would violate the Fair Housing Act, but continues to be an issue today. Large-scale data thus can be used in ways that lead us towards an ever-growing surveillance society, the corporate side of which is largely unregulated and moving much faster than governments can keep up.

[9] Ball K., Haggerty, K.D., & Lyon, D. (eds). 2012. Routledge Handbook of Surveillance Studies. Wood, D.M. 2009. "The 'Surveillance Society': Questions of History, Place, Culture." *European Journal of Criminology* 6.

[10] Foucault, Michel. 2003 [1976]. *Society Must Be Defended: Lectures at the College de France, 1975-1976*. New York: Picador. Foucault, Michel. 2009 [1978]. *Security, Territory, Population: Lectures at the College de France, 1977-1978*. New York: Picador.

[11] Angwin, Julia and Terry Parris, Jr. 2016. "Facebook Lets Advertisers Exclude Users by Race." *ProPublica*, October 26. Angwin, Julia, Ariana Tobin, and Madeleine Varner. 2017. "Facebook (Still) Letting Housing Advertisers Exclude Users by Race." *ProPublica*, November 21. Zang, Jinyan. 2021. Case Studies in Public Interest Technology. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

QUESTION 2

As new forms of data and analytic methods are developed, how can we remain conscious of the ways that human values and choices are driving the emergence and use of these seemingly neutral, objective, and scientific techniques?

This question has been taken up by an interdisciplinary field of study usually referred to as "science and technology studies," or STS.[12] The theoretical focus of STS scholars is on the social production of techno-scientific objects. Such objects—including machines, pharmaceuticals, and algorithms—are not cold and neutral, but rather should be recognized as manifestations of human activity, negotiation, and translation made within institutional networks of resources and knowledge.

For example, while researchers, practitioners and the public at large increasingly recognize that algorithms can embed bias in their decision-making, the most common approach to this problem is to seek technical solutions for it: to build a more accurate algorithm.[13] Insights from STS scholars show us that although some technical improvements can be made, in fact the ideal bias-free algorithm does not exist. This is because algorithms are built by humans, who are not bias-free creatures and rely on data from social contexts already marked by inequality.

The goal, then, should be to remain aware of how human activity is shaping the creation of techno-scientific objects, as well as how humans draw on their specific social contexts to interpret technical outputs and decide how to act. For example, how should a decision be made about whether a person who is charged with a crime should be released prior to trial or held in jail?

The usual practice of judges making those decisions increasingly is being replaced (or at least augmented) by algorithmic decision-making tools,[14] a process that has received much recent attention and critique,[15] and which we discuss further later in this report. The rationale for using algorithms to support pre-trial release decision-making turns on the allegedly unbiased nature of these algorithms, especially when compared to human judges. It is critical, however, for people to monitor and resist this inclination to attribute a sense of infallibility to techno-scientific objects, instead remembering that they contain human foundations just like other forms of decision-making. Furthermore, we should remain aware that the reasons we take up such tools are often linked to the human pursuit of specific interests—such as a desire to influence public policy or a motivation to sell new products—and are not simply self-evident forms of progress.

[12] Gillespie 2014; Lee, F. & Larsen, L.B. 2019. "How should we theorize algorithms? Five ideal types in analyzing algorithmic normativities." *Big Data & Society*. Ames, M.G. 2018. "Editorial: Deconstructing the algorithmic sublime." *Big Data & Society*. See also *Big Data & Society* special issues May 2018 and August 2019.

[13] Gillespie 2014.

[14] Kleinberg, J. et al. 2017. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133. Lapowsky, I. 2017. "One State's Bail Reform Exposes the Promises and Pitfalls of Tech-Driven Justice. Wired." *https://www.wired.com/story/bail-reform-tech-justice/*

[15] Angwin, J. et al. 2016. "Machine Bias." ProPublica. Chouldechova, A. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data 5*; Washington, A.L. 2019. "How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate." Colo. Tech. L.J., 17(131)

QUESTION 3

How does the widespread take-up of new data and analytics change what human beings see as important, self-evident, or true?

This question has framed the inquiries of scholars in the humanities, especially history and new media studies.[16] Scholarship in these traditions often has examined how mass adoption of media technologies, such as the printing press, radio, film, television, and the Internet, have changed both the methods available for communication, and the content of what we communicate to each other. In the area of the data-society interface, the humanities perspective asks how the rise of data, technology, and digital media are changing how we think about and experience the world. This approach also draws our attention to how those new understandings and experiences produce new social phenomena, and even change the very nature of what it means to be human.

SUMMING UP

The three above questions frame different forms of inquiry into the data-society interface. Each question, and the research that is motivated by it, begins from a different theoretical position and offers a particular set of insights. Surveillance studies researchers examine how new technologies extend government and corporate power while restricting individual and collective autonomy. STS researchers emphasize that new science and technology is always shaped by human society rather than existing as something neutral and outside of it. Humanities researchers ask how new technology can deeply reshape how humans experience the world, and thus go on to act in it.

Though each perspective approaches the data-society interface from a different angle, all fundamentally are concerned with challenging our tendency to take new technology for granted, and especially to assume that technology always brings progress. They emphasize the relationship between human life and the technology it uses, pushing us to consider the deeper social implications and how we might better develop technologies whose goal is not simply to maximize efficiency or effectiveness, but rather to serve human needs holistically.

[16] Gitelman 2013; Ananny & Crawford 2018; Brunton & Nissenbaum 2011; Chun, W. 2016. "Big Data as Drama." *English Literary History* 83.

FEATURED SOLUTION

County Health Rankings & Roadmaps

The County Health Rankings & Roadmaps program of the University of Wisconsin Population Health Institute was developed to increase awareness of the determinants of population health, engage multiple sectors in population health improvement efforts, and ultimately present the data in a way that can shape how we collectively think and talk about health and what influences it. The County Health Rankings are based on a *model* of community health that emphasizes the many factors that influence how long and how well we live, including social and economic factors like education and community safety as well as clinical factors like access to care and elements of the physical environment like housing and transportation. According to the program's *10-year reflection report*, "The media has been a critical audience for disseminating the Rankings and, through the program's messaging, contributing to shifts in the public narrative about health and equity." A rigorous and ongoing message analysis of the media coverage of the Rankings found that over time, messages like "where we live matters to our health" became more pervasive in coverage, and the kinds of experts quoted in articles shifted from primarily public health officials to include more leaders of other sectors such as economic development, housing advocates and public officials, potentially showing more sectors seeing their role in advancing population health.

Data Analysis

With a more nuanced understanding of how data is created and shaped by societal standards and biases, it's important to turn next to how that data is analyzed and mobilized to shape our collective decisions. The volume, velocity, variety, and relationality (or interconnected nature) of data that characterizes our era of datafication has given rise to new approaches to data analysis. While traditional statistical methods remain useful, computational methods, such as data mining, machine learning, and data visualization increasingly are being used to shed light on the social world. This section reviews the basic distinctions between statistical and computational approaches to analysis, shows how different forms of analysis can shed light on different aspects of a social problem, and considers the connection between analysis and action. The section concludes by identifying key issues to which consumers of these different forms of data analysis should pay attention.

What Is the Analytical Goal?

Data has long been used by scientists and social scientists in the pursuit of *explanation* about the world: an understanding of why things we observe in the world exist or occur the way they do. For example, what individual and social factors explain income levels in a nation, the incarceration rate of a population, or the health status of a city? Why do people volunteer in their communities, drink too much alcohol, or have different numbers of children? Such knowledge gives us a better understanding of how we got to where we are; it can also be useful for helping us think about how to create change.

Other approaches to data analysis aim for *prediction:* knowledge about what will occur in the future. Which intersections in our city will experience more traffic accidents? Will this patient develop diabetes? When and where will severe weather occur this week? With these types of questions, an explanation is not necessary. We just want to know the answer so that we can decide how to anticipate or respond to events or conditions that are likely (though not guaranteed) to occur in the future.

Explanation: Descriptions and Causes

Data analysis that seeks explanation usually is either *descriptive* or *causal.* Both of these approaches help us understand existing conditions. Each one offers different information on how existing conditions might be changed in the future.

Descriptive data analysis allows us to understand the contours of particular items ("variables") of interest. Univariate description focuses on one variable at a time, such as average income in a country, or the number of housing units in a city. Bivariate and multivariate description look at the relationship between an outcome of interest (the "dependent variable") and one or more other items ("independent variables") thought to have some relationship to or influence on the item of interest.

For example, the average income in a country is related to that country's average level of educational attainment; or, a person's weight is thought to be influenced by factors such as genetics, food intake, exercise, and sleep. Descriptive analysis offers a form of explanation that helps identify which factors *may* be more or less influential in shaping the outcome of interest. The old adage "correlation does not imply causation" is important to remember in descriptive analysis, given that any observed *association* between two variables does not provide evidence that one of those variables *caused* the other.

Causal data analysis goes beyond description in that it aims to identify whether one variable actually causes the outcome of interest. More specifically, causal analysis asks whether *changing* one variable leads to change in the outcome of interest. In other words, if we make a specific intervention, does that produce a "causal effect" on the outcome of interest?[17] Again, while descriptive analysis can show *associations* between independent variables and an outcome, it cannot demonstrate that those variables *caused* the outcome. In contrast, causal approaches to data analysis ask targeted questions about what might be directly responsible for certain outcomes. For example, does taking a certain medication lower blood pressure? Does

[17] This framing of the question is what is referred to as the "effects of causes" paradigm in causal analysis. This paradigm asks questions of the form: "what is the result ("effect") of this intervention ("cause")"? This would be distinguished from a "causes of effects" approach, which would ask questions of the form "what is responsible ("causes") for this outcome ("effect")?" The latter type of question cannot be answered through existing causal inference methods, such as randomized controlled trials.

participating in a micro-savings program lead to increased food intake? Does experiencing a traumatic event during pregnancy result a baby being born with lower birth weight? By identifying these kinds of relationships, causal analysis provides guidance on how changes in the outcome of interest might be achieved.

Efforts to evaluate place-based interventions like Purpose Built Communities (PBC), a holistic neighborhood revitalization model to transform areas of concentrated urban poverty, provides an example of how complex it can be to isolate causal connections in real-world settings. According to an Urban Institute report, "a cross-sector coalition of organizations applied the Purpose Built Community model to the East Lake neighborhood in Atlanta and documented a wide range of positive changes, from reductions in crime rate to increases in reading levels and home values." The researchers used a method, known as synthetic comparison group analysis, which is widely accepted by researchers as the most appropriate method for examining place-based interventions. However, even with this method the researchers noted that more research is needed to determine whether positive outcomes are due to improvements in outcomes for existing residents or the result of new people with different characteristics moving into the neighborhood. Thus, the causal effect of the PBC model remains uncertain.

Prediction: Forecasting the Future

New types and quantities of data, along with vastly increased computing power, have given rise to widespread use of *predictive* data analysis. Prediction approaches, such as machine learning, ask which combination of variables, including transformations of variables, give the best model for accurately predicting a future outcome of interest. For example, if we are interested in knowing how much pollution will be in the air tomorrow, or which entering ninth graders are most likely to drop out of high school, or which patients would be the best candidates for a particular surgery, we would seek to identify what combination of characteristics (the "model") predict our outcome of interest with the greatest accuracy. Vast amounts of data, including variables that might not seem related to the outcome, can be processed in search of the algorithmic model that achieves the highest predictive power.

> An algorithmic model that *predicts* an outcome cannot be used to determine the *cause* of that outcome.

Importantly, the prediction paradigm does not seek to *explain* any of the relationships between variables in the model and the outcome; rather, it seeks to find the set of characteristics that does best at *predicting* the outcome. This means that knowing which variables are part of the *predictive* model is not the same as knowing which factors *cause* the outcome of interest.

Leveraging Data to Inform Action

Data analysis can usefully explore different kinds of questions in order to help guide us in developing strategies for action. An important first step in the analytical process is to ask whether a particular question seeks an explanation or a prediction; the choice of analytic approach and method will hinge on the answer. The search for solutions to social problems contains both explanatory and predictive questions, within and across different social problem areas. As in all research, then, the way the question is asked will drive what data analysis can reveal. Both causal and predictive analysis can leverage data to inform action. Users and consumers of new data and new methods to inform action must exercise analytic caution, however, in order to recognize what kind of question is being asked and then choose whether causal or predictive analysis is better suited to developing a strategy for action. All of these analytic considerations also should be informed by how well the data to be analyzed actually represents the social phenomenon under study, including issues of data bias, as discussed in prior sections of this report.

Causal analysis uses experimental (e.g., randomized controlled trials) and quasi-experimental (e.g., difference-in-difference, regression discontinuity, instrumental variables) methods to isolate the effect of a *specific cause* and ascertain the magnitude by which it changes a desired outcome. For example, does giving people experiencing homelessness cash assistance reduce their entry into expensive helping systems such as shelters and emergency rooms? If a causal analysis finds this to be true, then policymakers may be willing to bet that expanding this intervention to more people is likely to have the same effect for them, and ultimately will reduce public spending on these expensive helping systems. Causal analysis is now widespread as a method for testing whether particular interventions in fact improve outcomes across a wide range of social issues, including health, employment, public safety, education, and so on.

There are two main downsides to the causal approach. First, the so-called “gold standard” causal methodology, the randomized controlled trial (RCT), is quite difficult, time-consuming, and expensive to implement. This means that implementation challenges and limited resources may make it impossible to assemble causal evidence for many effective interventions. Second, many RCTs suffer from a lack of external validity, i.e., a failure to replicate the experimental effect beyond the study population. In other words, many interventions have produced a clear effect during a causal evaluation, but then that success has not extended to other situations beyond the experimental setting. Researchers and practitioners are aware of these two challenges of the causal approach, and are actively seeking ways to address them.

Predictive analysis is based on an entirely different epistemology from causal analysis. Prediction attempts to forecast the future based on what has happened in the past. For example, how can we predict which individuals who have been arrested are likely to show up for trial if released into the community without bail? A predictive analysis would use data on past arrestees, whose appearance at trial (or not) is already known. A model would be built to identify which characteristics most accurately predicted who showed up for trial. Based on these findings, judges and other criminal justice personnel may decide to release individuals whose characteristics match those of the prior arrestees who did in fact show up for trial, while detaining individuals without those characteristics.

It is critical to refrain from confusing prediction (which is a form of correlation) with causation. A predictive analysis may indicate which characteristics identify people who are likely to fail to show up for trial, but it cannot say what criminal justice personnel should do to get those same people to show up for trial. For example, one factor that might *predict* someone is less likely to show up for trial might be that a person is unemployed. This does not necessarily mean, however, that giving someone a job will *cause* them to appear at trial; sometimes employed people fail to show up for trial, too. Only actions that respond to a *situation as it exists* can be informed by predictive analysis; prediction cannot offer guidance on what actions will *change the situation*. At the same time, even if predictive analysis is used to address an appropriate question, a downside of the predictive approach is that it ties individual fates—such as release or detention after being arrested—to the average behavior of groups. For example, judicial decisions about whether to grant bail to current defendants with a certain set of characteristics are dependent on what past defendants with that same set of characteristics have done. Some legal scholars question whether such an approach comports with law regarding the rights of individuals to be judged as individuals, rather than as members of groups.

How Different Analytic Approaches Affect Decision-Making: Mass Incarceration as an Example

Most social problems have multiple aspects and can be framed in multiple ways. This section illustrates how different analytical approaches (descriptive, causal, and predictive) to a single social issue—mass incarceration in this example—can lead to different understandings of the problem and different implications for solutions.

Mass incarceration is a significant social problem in the U.S. and produces particularly negative consequences for Black men and Black communities. There are a number of different components to this problem, but one that has generated a lot of interest from activists and researchers alike concerns the use of pre-trial detention: holding someone in jail after they have been arrested but before they have been convicted of a crime. Detention can wreak havoc on a person’s life, leading to job loss, strains on personal relationships, homelessness, the removal of one’s children by the state, and other stark consequences. These costs of pre-trial detention must be borne even when a person is ultimately found not guilty of the crime with which they have been charged. Disturbingly, even when there is no conviction, the costs of pre-trial detention are rarely repaid.

When a person is arrested for a crime, they either are released into the community until their case goes to trial, or they are ordered to be held in pre-trial detention; in the latter case, cash bail is sometimes offered to facilitate release while also ensuring that a person shows up for trial. The post-arrest decision point can have far-reaching consequences for individuals, and analysts have applied descriptive, causal, and predictive approaches to both increase our understanding of the problem and to attempt to improve decision-making.

Descriptive analysis has demonstrated, among other things, the racial disproportionality in pre-trial detention: Black people are over-represented in pre-trial detention compared to their numbers in the

FEATURED SOLUTION

CARDIFF Violence Prevention Model

More than half of violent crime in the United States is not reported to law enforcement, according to the U.S. Department of Justice, further illustrating why arrest data is insufficient for conducting causal analyses on crime. According to the *CARDIFF Model Toolkit* from CDC and the University of Pennsylvania, "That means cities and communities lack a complete understanding of where violence occurs, which limits the ability to develop successful solutions. The Cardiff Violence Prevention Model provides a way for communities to gain a clearer picture about where violence is occurring by combining and mapping both hospital and police data on violence. But more than just an approach to map and understand violence, the Cardiff Model provides a straightforward framework for hospitals, law enforcement agencies, public health agencies, community groups, and others interested in violence prevention to work together and develop collaborative violence prevention strategies."

population at large. Descriptive analysis also shows that when Black people and white people are arrested for the same crime, Black people are more likely to be held in pre-trial detention. These and other descriptive findings have informed arguments that the disparate racial impact of pre-trial detention needs to be addressed in the interest of justice.

Causal analysis of this issue has attempted to identify what might be driving decisions that contribute to worse outcomes for Black defendants. There are many decision points that could be contributing, such as who police officers choose to arrest, how police managers allocate officers to certain neighborhoods, which defendants judges award bail to, and so on. Causal analysis attempts to identify the impact of specific decisions like these to see if interventions at those points might change criminal justice outcomes. For example, a recent study[18] looked at the effects of pre-trial detention on convictions, future criminal activity, and future employment. Using a quasi-experimental design with an instrumental variables approach, the study found that pre-trial detention significantly increased convictions, primarily as a result of incentivizing defendants to plead guilty. In addition, pre-trial detention did not lead to any net decrease in crime, but it did lead to worse employment outcomes for those who were detained. The analysis thus makes a cautious recommendation that reducing pre-trial detention would cause both improved outcomes for many individual defendants, and a net societal benefit.

With mounting causal evidence that pre-trial detention is harmful for defendants, *predictive* approaches to this issue have asked how to improve decision-making about who is detained pre-trial. If the function of pre-trial detention is to (1) ensure that defendants show up to trial and (2) prevent defendants from being re-arrested, then a predictive analysis would build an algorithmic model to identify which defendants were likely to fall into either of these two categories. Defendants predicted to either jump bail or recidivate would be held in pre-trial detention; all others would be released into the community—without the further hurdle of cash bail. This analysis is not interested in figuring out how to get a person who is likely to fail to appear at trial to instead show up; it only seeks to identify which individuals are sufficiently at risk for not showing up, in order to detain them. It is important to note that this approach disregards the question of racial disproportionality in pre-trial detention, and that the results of this analysis might either reduce or increase such disproportionalities, even as it is likely to reduce the numbers of people held in pre-trial detention overall.

Cautions for Causal and Predictive Analysis

The sense that information presented as numbers implies certainty of knowledge is seductive. So too are the allure that "big data" captures more of the world than has previously been possible, or that new computational methods offer solutions to chronic analytic conundrums. We increasingly recognize, however, that new

[18] Dobbie, Will, Jacob Goldin and Crystal S. Yang. 2018. "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." *American Economic Review* 108: 201-240.

data and methods must confront many of the same problems as old data and methods, while also having unique challenges of their own.

Both causal and predictive analysis have some limitations as a basis for taking action to solve social problems. On the causal side, for any outcome of interest—like addressing decision points that contribute to racial disproportionality in pre-trial detention—the effects of hypothesized causes may be small. As such, even interventions that are based on empirically established causal effects may produce small changes. Such interventions also may have limited external validity; that is, they might be effective in one setting, but not in another. For predictive analysis, it is important to note that because predictions about the future are based on what happened in the past, the accuracy of those predictions will depend on how much conditions in the future replicate conditions of the past. Given the nature of social change, it may be difficult to use data adjustment strategies to identify and address disjunctions between past and future.

In addition, data used in predictive models may not accurately reflect the social processes underlying their production, leading to biased data. Data quality is a familiar concept to users of descriptive and causal analysis, and remains important in predictive analysis. Whether a data element accurately represents reality directly affects the validity of the subsequent analysis. For example, data on arrests is a faulty representation of crime, due to various biases in how policing is conducted: certain types of crime are more likely to occur in public and thus be more easily visible to police; policing may be more aggressive in certain areas than in others; police officers may be more likely to arrest members of particular racial or ethnic groups than others; and so on. Thus, a predictive analysis that uses arrest data to determine where to concentrate police resources may over-estimate the amount of crime in certain areas, send more police to those areas, and thereby lead to even more arrests in those areas, further exacerbating the problem.

This type of process often is referred to as the “analytic bias” in algorithms: when predictive algorithms use data that misrepresents the phenomenon of interest, and then reproduce the biases in that data by suggesting actions that take us further down a path that might have been interrupted by different choices. Supporters of automated decision-making often counter this view by arguing that human decision-makers, such as judges, have biases and other forms of fallibility that render their decisions less than optimal, and sometimes harmful. This certainly can be the case. At the same time, however, it is not simply the case that the deep reservoir of data on which automated tools draw can offer a perfect form of decision-making, free of human error. Predictive analysis, like any kind of data analysis, makes assumptions about how specific data bits represent the world. In our time of increasingly explosive data accumulation, we must interrogate whether the data we use accurately capture the social processes we claim they do, and develop approaches for monitoring this critical issue.

SUMMING UP

The explosion of data has increased our collective appetite for data-driven decision-making on a wide range of social issues. Just as we must be aware of the benefits and shortcomings of the datasets we choose to gain insight into these issues, we must be clear about our goals when we select a method for data analysis. Long traditions of using data for descriptive and causal analysis exist, and consumers of these types of analysis should be clear about what these methods can and cannot offer. The same goes for newer methods that are predictive in nature, such as machine learning. In particular, predictive analysis should not be confused with causal analysis: the former discovers (at times unexpected) associations, whereas the latter identifies causal mechanisms, and these are not the same. No one analytical method tells us everything we need to know, and it is our questions about the world that should lead us to select the type of data and analysis that will best serve our efforts to inform decision-making with data.

Urgent Need for Ethics and Regulation at the Data-Society Interface

As datafication and new analytic methods become increasingly commonplace, we are faced with questions and challenges that have accompanied every other new wave of technology: how do these new tools and their associated practices pose risks and consequences for human life?

> "The rapid pace at which datafication and new analytic applications are emerging virtually ensures that regulation of these issues will inevitably lag behind practice and innovation. This sharpens the need for a robust engagement with ethics."

The rapid pace at which datafication and new analytic applications are emerging virtually ensures that regulation of these issues will inevitably lag behind practice and innovation. This sharpens the need for a robust engagement with ethics, as practitioners and consumers at the data-society interface often will need to rely on their own ethical compass. We begin this section with a brief discussion of ethics, and then point to several of the major ethical issues of the data-society interface: privacy, analytic interpretability, and fairness.

Practical Ethics for the Data-Society Interface: Should We? Can We?

Principles of ethics generally can be classified into two camps: consequentialist and deontological. Consequentialism holds that ethical choices should be assessed solely by the states of affairs they bring about: moral choices are those which increase what society considers intrinsically valuable (often shorthanded as "The Good"). How "The Good" is defined differs across contexts, and this challenge has been a focus of much consequentialist debate. In contrast, the deontological perspective argues that the morality of choices is independent of the states of affairs those choices bring about: there is right and wrong, and we can be clear which is which regardless of the outcome. This means that a choice that is not in accordance with "The Right" cannot be made even if it were to enhance "The Good." In other words, as the saying goes, the ends do not justify the means.

These opposing perspectives often give way to a practical application of ethics conceived of as *ethical reasoning*. This means specifying well-founded standards for action, and then (regularly) assessing whether our actions live up to those standards. While some actions might be marked as clearly unethical, many actions cannot be. The latter instead require us to engage in ethical reasoning. New ethical challenges pervade the data-society interface, as choices that have rarely or never been seen before present themselves, and actions must be taken. No easy checklist exists for how to make these choices ethically, and the reality is that as new situations emerge, our ethical reasoning may change, and we may need to make new trade-offs between principles. In doing so, two questions can help guide the choices we make at the data-society interface.[19]

First: *should we be doing this?* This is a deontological question. It asks us to be clear about the commitments we're obligated by society or our organization to uphold. Those commitments may come from various sources; three of the most common are professional ethical codes, such as in medicine or social work; the legal strictures of nation-states; and principles of universal human rights. There are, thus, different ways to go about answering the question "should we be doing this."

Second: *can we do this right?* This is more of a consequentialist question, particularly when considering that data-driven decision-making increasingly is being used to make high-stakes decisions for individuals and groups. For example, predictive risk models are being used by government child protective services to determine whether a family should be investigated for possible abuse or neglect of a child. Putting a family in contact with a coercive state system known for its racial bias in how it treats families represents

[19] Leslie, David. 2019. *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector.* The Alan Turing Institute. https://zenodo.org/record/3240529.

a high-stakes decision for that family—particularly if the family is Black. It is thus critical to ask whether this approach to decision-making can be "done right," including using high-quality data, conducting careful data pre-processing, designing the predictive model deliberately, testing and validating the model diligently, and training users of the model to ensure trustworthy implementation of indicated decisions.[20]

Beyond these general principles of ethics and an approach to ethical reasoning, we face several specific ethical challenges in the use of data to improve social conditions. We discuss three of those here: privacy, analytic interpretability, and fairness.

Privacy

Widespread datafication of transactions, images, activities and more means our everyday actions become digital traces—fine-grained records of us that are stored as never before, and potentially subject to investigation and use by actors of all kinds. The privacy issues raised by datafication are myriad and far from settled. Outside of scientific research settings—where existing regulations protect the rights of people participating in research—data collection, sharing, and resale protocols are in their infancy. Today, when so much personally-identifiable data is gathered by private companies, often as a byproduct of wide-ranging user agreements that few people ever read, privacy issues are increasingly pressing.

One recent concern is the potential risk of reidentification that occurs when multiple datasets are linked together for analysis. While individuals represented in a single dataset may not be connected to any identifying information (e.g., name, birthdate, address, etc.), it is increasingly possible to link datasets that include some identifying information, and thus to build a composite image of an individual which includes details that are private and sensitive.

Privacy issues may be governed by public agencies, private associations, or nobody at all. While privacy governance entities have attempted to keep pace with the developing frontier of technology, it appears nearly impossible to do so.

Analytic Interpretability: Is There a Right to a Human Decision?

Part of the risk of computational analysis is a lack of transparency in how data are analyzed in computational models. Many machine learning algorithms are so-called "black box" algorithms: they can learn from themselves, constantly iterating over their own analysis and refining themselves in increasingly complex ways.[21] Their exact function therefore becomes obscured even to the researchers who set them in motion.

This kind of power can analyze data at larger scales and faster than ever before, but there are also risks when analysis is not transparent. For example, machine

[20] Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York: St. Martin's Press; Leslie 2019; Leslie, David, Lisa Holmes, Christina Hitrova and Eleanor Ott. 2020. *Ethics Review of Machine Learning in Children's Social Care.*

[21] Rudin, Cynthia and Joanna Radlin. 2019. "Why Are We Using Black Box AI Models When We Don't Need To? A Lesson from an Explainable AI Competition." *Harvard Data Science Review* 1, 2.

FEATURED SOLUTION

European Union's Approach to Data Privacy

The most far-reaching privacy effort to date, the *European Union's General Data Protection Regulation* (GDPR), was passed in 2018 to restrict the data collected regarding EU citizens.[1] The GDPR affirmed EU citizens' right to digital privacy and legally requires that data only be collected for certain purposes and as minimally as feasible for those purposes. It represents the first major step by a public governing body to regulate a technology that is developing faster than relevant law and regulatory systems.[2]

[1] *Text of GDPR.*

[2] Tutt, Andrew. 2017. "An FDA for Algorithms." *Administrative Law Review* 69: 83-123.

learning is certainly much more effective at determining whether an email should or should not be sent to a spam folder than would be a human performing the same task by hand. This is because spam identification algorithms have access to large quantities of data documenting the features of emails which users previously have hand-marked as spam. Over time, the algorithm increasingly can identify the characteristics common to spam-flagged emails and learn to automatically filter emails including those characteristics out of a user's inbox. However, when algorithmic automation reaches a certain point it becomes unclear exactly what characteristics it is using to separate spam from non-spam. The designers of the algorithm have not programmed this selection process into it; rather, the algorithm now is selecting its own criteria.

This process might be acceptable for a mundane task like filtering spam emails, but what about for higher-stakes issues with greater influence over human lives? What criteria are being used in analysis? Can we be sure rights to privacy or against ethnic or sex discrimination are not being violated? Many data ethicists have insisted that algorithmic decisions that affect people must be transparent. Article 22 of the GDPR provides that individuals have the right not to be subject to solely automated decisions. Recent scholarship on the American context, however, has argued that there is no firm legal basis for establishing a right to a human decision.[22]

Algorithms can greatly boost our capacity to make decisions, but when we reach the point of passing decisions entirely over to an algorithm, we cross a very significant threshold. Constraining researchers to develop algorithms that are transparent and can always be interpreted and explained might mean losing some raw prediction power (though some researchers argue this is not always the case).[23] Nevertheless, users of algorithms for making high-stakes human decisions also must weigh the importance of their accountability to the public.

The Myth of Algorithmic Neutrality

Shifting decision-making from humans to algorithms raises important questions about whether algorithmic decision-making does a better or worse job than the human decision-makers it might replace. Just as human decisions often are scrutinized for signs of fairness or bias, so too must we address the ways in which algorithmic decision-making may systematically disadvantage certain individuals or groups.[24] There are a number of reasons why algorithms systematically fail to produce the kinds of neutral outcomes that we often imagine data products are capable of achieving. While there is no clear consensus on how to think about these issues, we suggest two broad categories that we believe capture much of the current discussion: training bias and fairness definition.

Training bias refers to faults in the process by which algorithms learn to recognize objects, build classifications, and discern patterns. The basis of this training process is existing, available data that has already been classified ("labeled") by humans, such as records of people who have been approved for mortgages, pictures of objects that have been named, or grades of students who have dropped out of high school. It is critical to note that the data we use in most algorithmic training efforts reflects the ways that humans have constructed social life. Human biases and limits are reproduced in social practices, leading to data imbued with those biases, and thus their encoding in the algorithms we build.[25]

For example, if a company uses its employment records to train an algorithm to identify the best candidates for open positions, the longstanding history of occupational sex segregation—women being more likely to hold lower-status, lower-paying jobs compared to men—will be reproduced in the algorithm's choice of top job candidates. That is, because certain characteristics associated with females—such as names, college majors, extracurricular activities, etc.—are also more likely to have been associated with lower-status jobs in the past,

[22] Huq, Aziz Z. 2020. "A Right to a Human Decision." *Virginia Law Review* 106: 611-688.

[23] Rudin & Radlin 2019; Salganik et al. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117: 8398-8403.

[24] O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Benjamin, Ruha. 2019. "Assessing Risk, Automating Racism." *Science* 366: 421-422.

[25] Caliskan, Aylin, Joanna J. Bryson and Arvind Narayanan. 2017. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356: 183-186. MacCarthy, Mark. 2019. "Fairness in Algorithmic Decision- Making." Washington, DC: Brookings Institute; Buolamwini, Joy and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Gender Classification." *Proceedings of Machine Learning Research* 81: 1-15.

the algorithm will select more female candidates for present-day lower-status job openings.[26] Recent examples of this kind of training bias abound. So too do examples where available training data are insufficient to capture reality, leading to algorithms that omit certain aspects of social life. For example, one recent study found that standard image datasets used to train facial recognition algorithms proved far less able to accurately identify darker-skinned faces due to the low representation of darker-skinned faces in the training data.[27] One potential consequence of this training bias is that darker-complexioned people are more likely to be erroneously identified as matching the face of a person being sought by law enforcement, as the algorithm is worse at distinguishing between darker-skinned faces than between lighter-skinned faces. And in the health field, skin cancer detection models used to help with efficient diagnosis are being trained using mostly images of lighter-complexioned skin—again, this training bias can have dire consequences if skin cancer diagnoses are missed for patients with darker skin tones.

A second type of ethical challenge concerns the *fairness definition* used in any application of algorithmic models to decision-making. The core of the fairness discussion rests on the question of whether algorithms treat members of different groups—e.g., men and women, Blacks and Whites—the same. However, fairness can be defined in multiple ways, and thinking about fairness in the context of algorithmic models forces us to explicitly articulate which fairness definition we are using. Just as human decision-making can choose to assert one definition of fairness over others, so too can algorithmic models be constructed to privilege one definition of fairness.

A key recent debate of this issue has shown that when applying the same algorithmic model to two different groups, a choice must be made as to whether the algorithm will give both groups the same *outcome rate* or the same *error rate.*[28] For example, if an algorithm is determining which applicants can be approved for a loan, an equal outcome rate means the algorithm will recommend that the same proportion of different groups of applicants—say, men and women—should be approved.[29] In contrast, if the algorithm is programmed to obtain an equal error rate for male and female applicants, it means that the same proportion of men and women will be erroneously rejected for a loan—i.e., will be classified as unqualified for a loan when in fact they are qualified.

Mathematically, it is impossible for an algorithm to achieve both an equal outcome rate and an equal error rate. This means that algorithm designers must choose which of these definitions of fairness—or some other definition of fairness—the model should be trained to obtain. Some scholars also argue that the best solution to the challenge of fairness definition is to require that algorithms be transparent, thereby ensuring that others can understand which fairness definition is being used.

SUMMING UP

As the landscape of the data-society interface is evolving rapidly, new ethical challenges and new choices arise with astonishing frequency. There is no easy checklist for how to make these choices ethically. As a society, we need to continuously come back to two central questions when it comes to our collective use of data: should we be doing this? And can we do this right? These ethical questions can guide us through high-stakes decisions around the use of data and its implications for privacy, analytic interpretability, and fairness. As we go forward, we have to contend with hard questions on how people's data can be used and shared, whether algorithms alone should be responsible for high-stakes decisions without being transparent and interpretable by the people they affect, and how we can more intentionally scrutinize the ways in which algorithmic decision-making may introduce bias and systematically disadvantage certain individuals or groups.

[26] Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability and Transparency,* pp. 469-481.

[27] Buolamwini and Gebru 2018.

[28] Angwin et al. 2016; Chouldechova 2017; Paulus, Jessica K. and David M. Kent. 2020. "Predictably Unequal: Understanding and Addressing Concerns that Algorithmic Clinical Prediction May Increase Health Disparities." *npj Digital Medicine* 3:99.

[29] An algorithmic model with an equal outcome rate is also referred to as a "well-calibrated" model.

Conclusion

This report illuminates key issues posed by the use of data to inform priorities and decisions in our society. Data is being produced, analyzed, and leveraged to inform action at a breadth and scale that far exceeds anything we have seen even in the recent past. Despite this transformation, data remains what it always has been: a tool that should be used in service to human contemplation, assessment, and decision-making. We believe that viewing data as a solution in itself, as a substitute for the debate that occurs in the realms of ethics and politics, is a grave mistake. It is thus incumbent upon us to understand how and why we use data, how we make meaning from it, how we can hold these processes accountable to our values and serve our broader collective goals for improving society.