66 STAN. L. REV. ONLINE 41
September 3, 2013

THREE PARADOXES OF BIG DATA

Neil M. Richards & Jonathan H. King*

INTRODUCTION

Big data is all the rage. Its proponents tout the use of sophisticated analytics to mine large data sets for insight as the solution to many of our society's problems. These big data evangelists insist that data-driven decision-making can now give us better predictions in areas ranging from college admissions to dating to hiring.[1] And it might one day help us better conserve precious resources, track and cure lethal diseases, and make our lives vastly safer and more efficient. Big data is not just for corporations. Smartphones and wearable sensors enable believers in the "Quantified Self" to measure their lives in order to improve sleep, lose weight, and get fitter.[2] And recent revelations about the National Security Agency's efforts to collect a database of all caller records suggest that big data may hold the answer to keeping us safe from terrorism as well.

Consider *The Human Face of Big Data*, a glossy coffee table book that appeared last holiday season, which is also available as an iPad app. Such products are thinly disguised advertisements for big data's potential to revolutionize society. The book argues that "Big Data is an extraordinary knowledge revolution that's sweeping, almost invisibly, through business, academia, government, healthcare, and everyday life."[3] The app opens with a statement that frames both the promise and the peril of big data: "Every animate and inanimate object on earth will soon be generating data, including our homes, our cars, and yes, even our bodies." Yet the app and the book, like so many

* Neil M. Richards is Professor of Law, Washington University. Jonathan H. King is Vice President, Cloud Strategy and Business Development, Savvis, a CenturyLink Company.

1. *See, e.g.*, Adam Bryant, *In Head-Hunting, Big Data May Not Be Such a Big Deal*, N.Y. TIMES (June 19, 2013), http://www.nytimes.com/2013/06/20/business/in-head-hunting-big-data-may-not-be-such-a-big-deal.html?pagewanted=all&_r=0.

2. *See* Emily Singer, *Is "Self-tracking" the Secret to Living Better?*, MIT TECH. REV. (June 9, 2011), http://www.technologyreview.com/view/424252/is-self-tracking-the-secret-to-living-better.

3. RICK SMOLAN & JENNIFER ERWITT, THE HUMAN FACE OF BIG DATA 3 (2012).

proponents of big data, provide no meaningful analysis of its potential perils, only the promise.

We don't deny that big data holds substantial potential for the future, and that large dataset analysis has important uses today. But we would like to sound a cautionary note and pause to consider big data's potential more critically. In particular, we want to highlight three paradoxes in the current rhetoric about big data to help move us toward a more complete understanding of the big data picture. First, while big data pervasively collects all manner of private information, the operations of big data itself are almost entirely shrouded in legal and commercial secrecy. We call this the *Transparency Paradox*. Second, though big data evangelists talk in terms of miraculous outcomes, this rhetoric ignores the fact that big data seeks to identify at the expense of individual and collective identity. We call this the *Identity Paradox*. And third, the rhetoric of big data is characterized by its power to transform society, but big data has power effects of its own, which privilege large government and corporate entities at the expense of ordinary individuals. We call this the *Power Paradox.* Recognizing the paradoxes of big data, which show its perils alongside its potential, will help us to better understand this revolution. It may also allow us to craft solutions to produce a revolution that will be as good as its evangelists predict.

I. THE TRANSPARENCY PARADOX

Big data analytics depend on small data inputs, including information about people, places, and things collected by sensors, cell phones, click patterns, and the like. These small data inputs are aggregated to produce large datasets which analytic techniques mine for insight. This data collection happens invisibly and it is only accelerating. Moving past the Internet of Things to the "Internet of Everything," Cisco projects that thirty-seven billion intelligent devices will connect to the Internet by 2020.[4] These devices and sensors drive exponentially growing mobile data traffic, which in 2012 was almost twelve times larger than all global Internet traffic was in 2000.[5] Highly secure data centers house these datasets on high-performance, low-cost infrastructure to enable real-time or near real-time big data analytics.

This is the Transparency Paradox. Big data promises to use this data to make the world more transparent, but its collection is invisible, and its tools and techniques are opaque, shrouded by layers of physical, legal, and technical

4. Dave Evans, *How the Internet of Everything Will Change the World . . . for the Better #IoE [Infographic]*, CISCO BLOGS (Nov. 7, 2012, 9:58 AM PST), http://blogs.cisco.com/news/how-the-internet-of-everything-will-change-the-worldfor-the-better-infographic.

5. *See* CISCO, CISCO VISUAL NETWORKING INDEX: GLOBAL MOBILE DATA TRAFFIC FORECAST UPDATE, 2012-2017, at 1 (2013), *available at* http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf.

privacy by design. If big data spells the end of privacy, then why is the big data revolution occurring mostly in secret?

Of course, there are legitimate arguments for some level of big data secrecy (just as there remain legitimate arguments for personal privacy in the big data era). To make them work fully, commercial and government big data systems which are constantly pulling private information from the growing Internet of Everything are also often connected to highly sensitive intellectual property and national security assets. Big data profitability can depend on trade secrets, and the existence of sensitive personal data in big databases also counsels for meaningful privacy and security. But when big data analytics are increasingly being used to make decisions about individual people, those people have a right to know on what basis those decisions are made. Danielle Citron's call for "Technological Due Process"[6] is particularly important in the big data context, and it should apply to both government and corporate decisions.

We are not proposing that these systems be stored insecurely or opened to the public *en masse*. But we must acknowledge the Transparency Paradox and bring legal, technical, business, government, and political leaders together to develop the right technical, commercial, ethical, and legal safeguards for big data and for individuals.[7] We cannot have a system, or even the appearance of a system, where surveillance is secret,[8] or where decisions are made about individuals by a Kafkaesque system of opaque and unreviewable decision-makers.[9]

II. THE IDENTITY PARADOX

Big data seeks to *identify*, but it also threatens *identity*. This is the Identity Paradox. We instinctively desire sovereignty over our personal identity. Whereas the important right to privacy harkens from the right to be left alone,[10] the right to identity originates from the right to free choice about who we are. This is the right to define who "I am." I am me; I am anonymous. I am here; I am there. I am watching; I am buying. I am a supporter; I am a critic. I am voting; I am abstaining. I am for; I am against. I like; I do not like. I am a permanent resident alien; I am an American citizen.

How will our right to identity, our right to say "I am," fare in the big data era? With even the most basic access to a combination of big data pools like phone records, surfing history, buying history, social networking posts, and

---

6. Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008).

7. *See* Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW J. TECH. & INTELL. PROP. 239, 270-72 (2013).

8. *See* Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934, 1959-61 (2013).

9. *Cf.* DANIEL J. SOLOVE, THE DIGITAL PERSON (2005).

10. *See* Julie E. Cohen, *What Privacy Is For*, 126 HARV. L. REV. 1904, 1906 (2013).

others, "I am" and "I like" risk becoming "you are" and "you will like." Every Google user is already influenced by big-data-fed feedback loops from Google's tailored search results, which risk producing individual and collective echo chambers of thought. In his article, *How Netflix Is Turning Viewers into Puppets*, Andrew Leonard explains how:

> The companies that figure out how to generate intelligence from that data will know more about us than we know ourselves, and will be able to craft techniques that push us toward where they want us to go, rather than where we would go by ourselves if left to our own devices.[11]

Taking it further, by applying advances in personal genomics to academic and career screening, the dystopian future portrayed in the movie *Gattaca*[12] might not be that outlandish. In *Gattaca*, an aspiring starship pilot is forced to assume the identity of another because a test determines him to be genetically inferior. Without developing big data identity protections now, "you are" and "you will like" risk becoming "you cannot" and "you will not". The power of big data is thus the power to use information to nudge, to persuade, to influence, and even to restrict our identities.[13]

Such influence over our individual and collective identities risks eroding the vigor and quality of our democracy. If we lack the power to individually say who "I am," if filters and nudges and personalized recommendations undermine our intellectual choices, we will have become identified but lose our identities as we have defined and cherished them in the past.

III. THE POWER PARADOX

The power to shape our identities for us suggests a third paradox of big data. Big data is touted as a powerful tool that enables its users to view a sharper and clearer picture of the world.[14] For example, many Arab Spring protesters and commentators credited social media for helping protesters to organize. But big data sensors and big data pools are predominantly in the hands of powerful intermediary institutions, not ordinary people. Seeming to learn from Arab Spring organizers, the Syrian regime feigned the removal of restrictions on its citizens' Facebook, Twitter, and YouTube usage only to secretly profile, track, and round up dissidents. [15]

---

11. Andrew Leonard, *How Netflix Is Turning Viewers into Puppets*, SALON (Feb. 1, 2013, 7:45 AM EST), http://www.salon.com/2013/02/01/how_netflix_is_turning_viewers_into_puppets.

12. GATTACA (Columbia Pictures 1997).

13. *See* RICHARD H. THALER & CASS R. SUNSTEIN, NUDGE: IMPROVING DECISIONS ABOUT HEALTH, WEALTH, AND HAPPINESS (2009); Richards, *supra* note 8, at 1955-56.

14. *See* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 11 (2013).

15. *See* Stephan Faris, *The Hackers of Damascus*, BLOOMBERG BUSINESSWEEK (Nov. 15, 2012), http://www.businessweek.com/articles/2012-11-15/the-hackers-of-damascus.

This is the Power Paradox. Big data will create winners and losers, and it is likely to benefit the institutions who wield its tools over the individuals being mined, analyzed, and sorted. Not knowing the appropriate legal or technical boundaries, each side is left guessing. Individuals succumb to denial while governments and corporations get away with what they can by default, until they are left reeling from scandal after shock of disclosure. The result is an uneasy, uncertain state of affairs that is not healthy for anyone and leaves individual rights eroded and our democracy diminished.

If we do not build privacy, transparency, autonomy, and identity protections into big data from the outset, the Power Paradox will diminish big data's lofty ambitions. We need a healthier balance of power between those who generate the data and those who make inferences and decisions based on it, so that one doesn't come to unduly revolt or control the other.

CONCLUSION

Almost two decades ago, Internet evangelist John Perry Barlow penned *A Declaration of the Independence of Cyberspace*, declaring the Internet to be a "new home of [the] Mind" in which governments would have no jurisdiction.[16] Barlow was one of many cyber-exceptionalists who argued that the Internet would change everything. He was mostly right—the Internet did change pretty much everything, and it did create a new home for the mind. But the rhetoric of cyber-exceptionalism was too optimistic, too dismissive of the human realities of cyberspace, the problems it would cause, and the inevitability (and potential utility) of government regulation.

We think something similar is happening in the rhetoric of big data, in which utopian claims are being made that overstate its potential and understate the values on the other side of the equation, particularly individual privacy, identity, and checks on power. Our purpose in this Essay is thus twofold.

First, we want to suggest that the utopian rhetoric of big data is frequently overblown, and that a less wild-eyed and more pragmatic discussion of big data would be more helpful. It isn't too much to ask sometimes for data-based decisions about data-based decisionmaking.

Second, we must recognize not just big data's potential, but also some of the dangers that powerful big data analytics will unleash upon society. The utopian ideal of cyberspace needed to yield to human reality, especially when it revealed problems like identity theft, spam, and cyber-bullying. Regulation of the Internet's excesses was (and is) necessary in order to gain the benefits of its substantial breakthroughs. Something similar must happen with big data, so that we can take advantage of the good things it can do, while avoiding as much of the bad as possible. The solution to this problem is beyond the scope of this short symposium essay, but we think the answer must lie in the development of

16. John Perry Barlow, *A Declaration of the Independence of Cyberspace*, ELEC. FRONTIER FOUND. (Feb. 8, 1996), https://projects.eff.org/~barlow/Declaration-Final.html.

a concept of "Big Data Ethics"—a social understanding of the times and contexts when big data analytics are appropriate, and of the times and contexts when they are not.

Big data will be revolutionary, but we should ensure that it is a revolution that we want, and one that is consistent with values we have long cherished like privacy, identity, and individual power. Only if we do that will big data's potential start to approach the story we are hearing from its evangelists.