Redesigning consent: Big data, bigger risks

Harvard Kennedy School Misinformation Review[1]
January 2020, Volume 1, Issue 1
Creative Commons Attribution 4.0 International (CC BY 4.0)
Reprints and permissions: misinforeview@hks.harvard.edu
DOI: https://doi.org/10.37016/mr-2020-006
Website: misinforeview.hks.harvard.edu

Commentary

Redesigning consent: Big data, bigger risks

Over the last decade, the rapid proliferation of social media platforms coupled with the advancement of computational methods for collecting, processing, and analyzing big datasets created new opportunities for social science. But alongside new insights about the behaviors of individuals and groups, these practices raise new questions regarding what constitutes ethical research. Most critically, current disinformation scholars face a replication crisis exacerbated by uneven access to datasets, where social media users are unaware of their participation in academic research. Establishing scientific norms, where research is shared with the individuals whose data are accessed and processed in the name of science, involves redesigning consent and providing universal public access to databases. Ultimately, without methodological norms for disinformation studies, the field will continue to be dominated by corporate interests, further endangering the public's trust in disinformation research.

Author: Joan Donovan
Affiliation: Shorenstein Center on Media, Politics and Public Policy, Harvard Kennedy School, USA

How to cite: Donovan, J. (2020). Redesigning consent: Big data, bigger risks. *Harvard Kennedy School (HKS) Misinformation Review*, *1*(1).
Received: October 10th, 2019. Accepted: January 3rd, 2020. Published: January 14th, 2020.

A multi-billion-dollar information economy exists, and yet, most of us have no idea who is purchasing our data or how to ensure our privacy. While this is worrisome for the public who are adopting new communications technologies faster than ever before, this market now impacts social science research in disturbing new ways. Researchers rely heavily on access to social media datasets to inform studies of group dynamics. Following the 2016 US election, a new field of disinformation studies was born. Different from communication studies of propaganda or persuasive media, *disinformation studies* is characterized by uncovering the intentional abuse of social media platforms and the open web to manipulate the public's understanding of political issues.

The field of disinformation studies crystalized around a shared goal of revealing how malicious actors utilize information and communication technologies, bridging the fields of sociology, media studies, and political science, with more technical domains of data science and cybersecurity. While methods for studying the phenomenon of disinformation are still in flux, the corporate control of social media data has led to unevenly distributed access and an inability to replicate studies. At the same time, social media consumers are largely unaware that their behavior is being tracked and used for academic research, which has reinvigorated an ethical challenge of managing user privacy versus seeking consent.

Many researchers are blinded by the hope and hype cycle that typically envelops new scientific endeavors. As Hedgecoe (2004) points out, in the 1990s scientific studies of genetics overestimated the

[1] *A publication of the Shorenstein Center for Media, Politics and Public Policy, at Harvard University, John F. Kennedy School of Government.*

possibility of big data identifying rare diseases, while also misjudging the risks patients took on. In this piece, I argue that the collection and processing of large social media datasets calls into question the ethics of responsible research. Without clear limits on how social media data can be used coupled with a process for meaningful consent, researchers can unintentionally expose research subjects to a range of harms, including identity theft, financial fraud, harassment, abuse, or reidentification (Hedgecoe, 2004). To understand the ever-growing risks of big data, researchers must begin to work alongside civil society partners to safeguard fairness, accountability, and transparency in the ethical production of datasets. When data is the primary way of seeing society, the impacts of big data on vulnerable people are rarely foregrounded.

Big data today

In April 2018, Facebook in conjunction with the Social Science Research Council (SSRC) and several philanthropic foundations announced an unprecedented initiative to provide user data to academic researchers. While on the surface this effort looked like a step in the right direction for Facebook, who recently had several scandals related to data misappropriation, researchers were in a bind when Facebook failed to produce the data a year and a half later. While Facebook promised one of the biggest social media datasets of all time, the problem was that nobody, not even Facebook itself, knew how to ensure the privacy of its users.

The call for researchers to study Facebook coincided with Zuckerberg testifying in front of Congress regarding Cambridge Analytica and the Russian Troll Factory known as the Internet Research Agency. Both entities used Facebook's products to gather information from users, which was later weaponized to distort issues in the 2016 US election. In 2018, the SSRC, a non-profit, put out a call on behalf of Facebook asking researchers to describe how they would study disinformation on Facebook. Run by the SSRC, the Social Data Initiative[2] would bootstrap the prize-winning applicants with hundreds of thousands of dollars in funding for staff and associated costs. Facebook agreed to provide a database of 32 million "misinformation" URLs shared on Facebook to the winners of this call. The big dataset would be packaged by Facebook and held within a new start-up called Social Science One,[3] a non-profit incubated within Harvard's Institute for Quantitative Social Science.[4] Because of the size of the dataset, teams featuring data scientists were by-in-large the most competitive applicants.

The goal of the consortium was to create access for academic researchers to anonymized and privacy-protected social media data and metadata featuring URLs coded as misinformation by third-party fact-checkers. But, eighteen months later, Facebook had only issued synthetic data, i.e., data that looks similar to the expected dataset, and then very recently produced a "light-table" release[5] after pressure from funders. A New York Times article described how the project became imperiled, where some researchers were stuck until they obtained the controversial dataset (Kang, Hsu & Frenkel, 2018). Recently, Social Science One issued a statement calling the available data "extremely limited in scientific value," suggesting that the entire enterprise was stopped (The European Advisory Committee Social Science One and Science One, 2019).

[2] See the Social Data Initiative at the Social Science Research Council (SSRC)
[3] See Social Science One at Harvard's Institute for Quantitative Social
[4] See the Institute for Quantitative Social Science at Harvard
[5] See "Facebook Privacy-Protected URLs Light Table Release" by Solomon Messing, Christina DeGregorio, Bennett Hilebrand, Gary King, Saurav Mahanti, Chaya Nayak, Nathaniel Persily, Bogdan State, and Arjun Wilkins (Harvard Dataverse, 2019).

Pulling the emergency brake, though, provided an opportunity to publicly acknowledge the risks to social media users, particularly the potential for de-anonymization. Do researchers know how social media consumers may be stigmatized after being associated with a disinformation campaign? Moreover, because Facebook had started to provide data access to some research partners, it created a fissure between research teams who have privileged access to platform data and those who had followed the process but would not receive the promised data.

Corporate control of social media data creates several problems for the advancement of social science, including fracturing development of a shared replication methodology, thus making evaluation of findings impossible. It also encourages secret transactions between researchers and companies, contradicting the scientific principles of openness, public good, and peer-review.[6] Compounding this methodological crisis is another serious conflict; because social media companies have been reluctant to provide transparency around data reuse by researchers and others, it has been impossible to create standards for informed consent; thus, adding to the public skepticism of data and technology.

Big data's past

Sociologists and data scientists have faced this ethical problem before. In 2006, prior to incubating Social Science One, Harvard's Institute for Quantitative Social Science's launched another data archiving service called Dataverse – an academic research data sharing portal. Dataverse was embroiled in a controversy after publishing an "anonymized" dataset of college students' Facebook data, titled ''Tastes, Ties, and Time,'' or otherwise known as the T3 study (Parry, 2011; Lewis, Kaufman, Gonzalez, Wimmer, Christakis, 2008).

Michael Zimmer (2010), a privacy researcher and data ethicist, quickly de-anonymized the T3 dataset and the cohort was identified as Harvard undergrad students. Ironically, one of the studies published using the T3 dataset detailed students' adoption of privacy settings. Zimmer wrote about the reaction of the T3 researchers to the reidentification of the students in the T3 dataset (Zimmer, 2010). One T3 researcher replied that "the data was already public" and that doing research on Facebook is like being in the public square (Lewis et al., 2008). This laissez-faire response was indicative of the time, and to their credit, the T3 researchers followed ethical standards and consulted with other experts on the release.

Zimmer (2010) points out that the problem was both social and technical. It's not that "the data was already public," but that Facebook users did not expect their data to be reused in such a way by the T3 researchers. He writes: "The data was made available on Facebook for the purpose of social networking among friends and colleagues, not to be used as fodder for academic research." While there have been many changes to Facebook's terms of service since 2008 to include researcher and secondary reuse of personal data, we still do not know if users fully understand what that meant then or today. Indeed, few users even read the terms of service (Zimmer 2010; Berreby, 2017).

As the academic field of big data studies developed over the last decade, social scientists, STEM researchers, marketers, cybersecurity specialists, and technologists avoided questions of consent in favor of drawing together large swaths of data from all types of sources, including social media, web traffic, as well as digitizing paper materials like medical records, credit card reports, police records, and more. This situation has left many with no choice and no recourse except to participate in big data regimes. With each data entry, these professionals neglected to recognize that data are traces of people's behaviors and that without adequate protections data can become valuable for the wrong reasons. Now, places like

[6] For a critical overview of the scientific method, see for example "Author's Preface" by Robert K. Merton (The University of Chicago Press, 1973).

schools and hospitals are targets of hackers, who take hold of their data systems and resell records, often supplementing these stolen records with social media data (McGee, 2016).

In many ways, what we are going through today is predicated on researchers ignoring the same ethical questions of consent to use data over a decade ago, except now collection is largely automated and data sales are a billion-dollar industry. Controversially, the T3 study design relied on research assistants, who were also part of the Harvard cohort on Facebook, to visit their friends' pages and manually harvest data (Zimmer, 2010). However, data harvesting is now a feature of the platform. After a major investment in development of advertising features, Facebook has automated and monetized how to create discrete datasets for microtargeting segmented audiences. These datasets are useful to researchers, political operatives, and all manner of marketers alike. Just as big data is different from individual data points, automation is different because it increases speed and scale, which are features often weaponized in manipulation campaigns (Alim, 2014).

Big data's future

We must bear in mind that even with ethical standards and protocols in place, data is not static. Given how much data is collected, packaged, sold, and used for myriad reasons – not all of which are harmless, there must be a new privacy pact drawn up between Internet users, tech companies, and researchers alike (Nadler, Crain & Donovan, 2018). Big data is different. Its value is in the hidden patterns and relationships contained within it. As a result, harms can be both individual and collective. For example, an intelligence gathering operation harvesting data about a political opponent's social media audiences proved useful in an effort to spread microtargeted negative media coverage to sway public opinion (Monaco, 2017). In another disinformation campaign, Russian operatives sought to involve American activists in broadcasting disinformation and participating in public protests (Friedberg & Donovan, 2019).

Risk can vary by location. For example, anti-corruption activists currently working in India fear that the government is using the topic of disinformation as an excuse to wrangle big datasets from social media platforms to target and suppress dissenters. Where activists believed they would find likely allies in researchers to defend against the government collection of data, the field is split where some researchers describe the trade-offs as "privacy vs democracy," or "privacy vs good social science." Instead, researchers must understand privacy as a watchword for the disenfranchised. As Eubanks (2018) points out, the poor and marginalized are first to trade their privacy to access basic resources, including communication technologies. Globally, activists are fighting against the rapid expansion of the collection capacities of platform companies, data brokers, government and intelligence agencies. What position should scholars in disinformation studies take?

It is the responsibility of researchers to put the needs of subjects before their desires to collect big data for their own ends. Even as Social Science One admirably strives to follow the highest standards for big data release by abiding by differential privacy protocols, there is still no way for ordinary folks to resist inclusion.[7] Further, as Mary Gray, a senior researcher at Microsoft Research, has stated, "If you're afraid to ask your subjects for their permission to conduct the research, there's probably a deeper ethical issue that must be considered" (Goel, 2014, para. 23).

Data are partial representations of people; quantified, aggregated, and ready-made for discrimination. That has not changed with the introduction of social media. What is different today is how easy it is to collect, sell, and combine datasets to identify, attack, and divide different groups across the information ecosystem (Nadler et al., 2018). The founder of Data for Black Lives,[8] Yeshimabeit Milner

[7] See Apple's "Differential Privacy" overview

[8] See Data for Black Lives

(2019), uses a provocative slogan: “Abolish Big Data!” But the call to action is not to get rid of platforms or data science. Data for Black Lives flips the paradigm of corporate ownership of big data by endeavoring to “put data into the hands of those who need it most” (Milner, 2019, para. 9). How could disinformation studies reorient research protocols to adhere to a principle of minimal data collection until such time that meaningful consent is possible?

In the search for bigger data, some researchers have advocated for individuals and groups to take on bigger risks for a better understanding of society. The challenge of today is for disinformation studies to rise to a scientific standard that achieves openness, public good, and peer-review, while at the same time refusing privileged and restricted access to data that undermines the entire field. New methodological norms should include engaging technologists, community partners, and civil society organizations in the production of data pipelines, beginning from the ethical principles laid out by the Association of Internet Researchers, which includes advocating for meaningful informed consent throughout all stages of research (Association of Internet Researchers, 2019). Otherwise, we researchers face a risk of our own; becoming untrusted by the very public we endeavor to serve.

Bibliography

Alim, S. (2014). An initial exploration of ethical research practices regarding automated data extraction from online social media user profiles. *First Monday*, *19*(7). https://doi.org/10.5210/fm.v19i7.5382

Association of Internet Researchers. (2019). *Internet research: Ethical guidelines 3.0*. https://aoir.org/reports/ethics3.pdf

Berreby, D. (2017, March 3). Click to agree with what? No one reads terms of services, studies confirm. *The Guardian*. https://www.theguardian.com/technology/2017/mar/03/terms-of-service-online-contracts-fine-print

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. MacMillan.

Friedberg, B., & Donovan, J. (2019). On the internet, nobody knows you’re a bot: Pseudoanonymous influence operations and networked social movements. *Journal of Design and Science*, 6. https://doi.org/10.21428/7808da6b.45957184

Hedgecoe, A. (2004). *The politics of personalized medicine: Pharmacogenetics in the clinic*. Cambridge University Press. https://doi.org/10.1017/CBO9780511489136

Kang, C., Hsu, T., & Frenkel, S. (2018, April 8). Mark Zuckerberg meets with top lawmakers before hearing. *The New York Times*. https://www.nytimes.com/2018/04/09/technology/mark-zuckerberg-facebook.html?module=inline

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks 30*(4), 330-342. https://doi.org/10.1016/j.socnet.2008.07.002

McGee, M. K. (2016, September 27). *Research reveals why hacked patient records are so valuable*. Data Breach Today. https://www.databreachtoday.com/interviews/research-reveals-hacked-patient-records-are-so-valuable-i-3341

Milner, Y. (2019, July 8). *Abolish big data*. Medium. https://medium.com/@YESHICAN/abolish-big-data-ad0871579a41

Monaco, M. J. (2017). *Computing propaganda in Taiwan: Where digital democracy meets automated autocracy* (Working Paper No. 2017.2). The Project on Computational Propaganda. https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Comprop-Taiwan-2.pdf

Nalder, A., Crain, M., & Donovan, J. (2018). *Weaponizing the digital influence machine: The political perils of online ad tech.* Data & Society. https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf

Parry, M. (2011, July 10). Harvard researchers accused of breaching students' privacy. *The Chronicle of Higher Education*. https://www.chronicle.com/article/harvard-researchers-accused-of-breaching-students-privacy/

The European Advisory Committee Social Science One, & Social Science. (2019, December 11). *Public statement from the co-chairs and European advisory committee of social science one.* Social Science One. https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one

Zimmer, M. (2010). "But the data is already public:" On the ethics of research in Facebook. *Ethics and Information Technology, 12*(4), 313-325. https://doi.org/10.1007/s10676-010-9227-5