Perspectives on Big Data, Ethics, and Society

Report Prepared By: Jacob Metcalf, Emily F. Keller, and danah boyd

I. EXECUTIVE SUMMARY

The Council for Big Data, Ethics, and Society was convened to bring together researchers from diverse fields who were thinking deeply about ethical, social and policy challenges associated with the rise of "big data" research and industry, with an eye toward developing recommendations about future directions for the field. Our reports, meetings, and ongoing conversations have consistently indicated that there is a disjunction between the familiar concepts and infrastructures of science and engineering, on the one hand, and the epistemic, social, and ethical dynamics of big data research and practice, on the other. We contend that facilitating ethical conduct in data science and related endeavors requires careful consideration of big data's broad technical, social, and political contexts.

Big data is marked by technical advances in storage capacity, speed, and price points of data collection and analysis, and by a move towards understanding data as continuously collected, almost-infinitely networkable, and highly flexible. The ability to analyze datasets from highly disparate contexts and generate new, unanticipated knowledge sets the stage for both the power and peril of big data research and broader data science-related analysis. The conceptual, regulatory, and institutional resources of research ethics developed over the last 70 years were premised on assumptions about human data research practices that sometimes do not easily apply to data analytics work done under the umbrella of big data. This results in conflicts over whether big data research methods should be excluded from or forced to meet existing norms, whether existing norms should be made to accommodate the special circumstances of big data, or whether entirely new norms and institutional commitments are needed.

The Council's findings, outputs, and recommendations—including those described in this white paper as well as those in earlier reports—address concrete manifestations of these disjunctions between big data research methods and existing research ethics paradigms. We have identified policy changes that would encourage greater engagement and reflection on ethics topics. We have indicated a number of pedagogical needs for data science instructors, and endeavored to fulfill some of them. We have also explored cultural and institutional barriers to collaboration between ethicists, social scientists, and data scientists in academia and industry around ethics challenges. Overall, our recommendations are geared toward those who are invested in a future for data science, big data analytics, and artificial intelligence guided by ethical considerations along with technical merit.

II. BACKGROUND

Introduction

The explosion of data collection, sharing, and analytics known as "big data" is a rapidly sprawling phenomenon that promises to have tremendous impacts on economics, policing, security, science, education, policy, governance, health care, public health, and much more. Within a relatively short window, the infrastructures and methods of big data have wrought significant intellectual and organizational changes for many academic disciplines, government bodies, philanthropic and nonprofit organizations, and private enterprises. Given its newly expansive reach, big data has still-unfolding consequences for how various stakeholders access and wield power, and therefore how social and political goods are distributed. New industrial, educational, and governmental investments in data science and artificial intelligence highlight how big data is collapsing historically segmented technical approaches. Because of both the hype and the reality of these new developments, the proliferation of big data raises ethical[1] issues that demand deliberation. Big data's broad ethical consequences strain the familiar conceptual and infrastructural resources of science and technology ethics.

Knowledge production utilizing big data methods bends many central commitments of regulatory and compliance schemes. For many U.S. scholars in medicine, biology, and social science, the commitment to ethical research involving human subjects starts with an obligation to the ethical principles underpinning the Common Rule, the federal regulation that most Institutional Review Boards (IRBs) use to assess the ethical practice of work involving human participants. Yet, this rule is not designed for the type of work typically done under the purview of big data, raising significant questions for consideration.[2] For example, the Common Rule exempts research using public datasets from review, due to a long-standing—but now misplaced—assumption that the "publicness" of data renders future informational harms minor or inherently low-risk. However, big data's central power and peril is the ability to network and re-analyze datasets from highly disparate contexts—often in concert—to generate unanticipated insights. Datasets can no longer be considered static archives because they are now capable of generating new insights for researchers, and consequences for human subjects, indefinitely. Thus, research ethics regulations or principles that focus on the status of the dataset as "public," rather than focusing on the potential uses of the dataset, will miss a plausible category of harms. This example—one among many—demonstrates that epistemic conditions that were baked into research ethics regulation no longer hold in light of big data methods of knowledge production.

---

[1] Throughout this document, we will use "ethics" in the broadest possible sense, to include any type of inquiry that addresses normative questions, regardless of discipline.
[2] Metcalf J and Crawford K (2016) Where are Human Subjects in Big Data Research? The Emerging Ethics Divide. *Big Data & Society.* Spring 2016. DOI: 10.1177/2053951716650211

The Council has also found that big data sidesteps many of the informal modes of ethics regulation found in other science and technology communities. For example, three of the disciplines that inform the nascent field of data science (computer science, physics, and applied mathematics) have long been considered outside of human-subjects-related ethics concerns because their work and contributions have historically been about systems and not people. As a result, the content of the datasets at play are considered irrelevant to the substantive research questions—the data could be about quasars or about social networks without much difference to the root mathematical questions. Yet as big data inexorably draws these disciplines closer to sensitive human phenomena, the field of data science is finding that it does not have the ethics curricula or training materials developed for handling ethical challenges. Similarly, because these fields have fallen outside of the IRB system created by the Common Rule for better or for worse, attempts at self regulation have been largely ad hoc, such as ethics review panels for computer science academic conference committees. Because these fields have not been required to practically grapple with ethics requirements, they often lack access to pedagogical resources about research ethics that are widespread in other fields.

None of this is to say that members of these technical fields aren't engaged with ethical concerns. Rather, what is unfolding is emerging through informal networks and is unevenly distributed. Furthermore, while some of the issues they are encountering have long histories of consideration and debate in other fields, still others highlight disconnects between different fields. For example, the boundaries of experimental work in psychology are approached differently than ethnographic work in anthropology and statistical work in sociology. Interventions in big data systems can have properties of all three simultaneously, making visible the limitations of existing approaches.

This leads to the main finding of this Council: there is a substantial disjunction between the familiar infrastructures and conceptual frameworks of research ethics and the emerging epistemic conditions of big data. There are often good historical and institutional reasons for this disjunction, but it demands attention nonetheless. The Council's recommendations regarding future policy and research agendas are focused on establishing the intellectual resources and practical models necessary to address the consequences of this disjunction.

The unique ethical conditions of big data

The complex phenomenon we identify as "big data" has roots in the challenges of handling and analyzing large datasets that tax the memory or storage of a single computer, such as those in physics, finance, and the Census.[3] Without a doubt, rapidly expanding, ever-cheaper, and highly networked

---

[3] Likely the first use of "big data" to describe a coherent problem was in a publication by Michael Cox and David Ellsworth in 1997 attributing the term to the challenge of visualizing large datasets. See Cox M and Ellsworth D (1997) Application-controlled demand paging for out-of-core

computing capacity has set the stage for the rise of big data. However, getting a grasp on the ethics of big data requires theorizing big data as something more than a technological artifact. A number of scholars have noted the importance of placing big data within a historical framework that grounds it as an epistemic change beyond its substantial industry hype. Kate Crawford, Kate Miltner, and Mary Gray (2014) encourage us to not let rapid technological change do too much explanatory work when analyzing big data because that enables a shallow view of big data as a unitary, purely technical phenomenon. Rather, they argue that we should treat big data as a "mythological artifact" that constitutes a worldview, and endeavor to trace out the many ways that big data draws on, replicates, and alters ethically fraught matters of knowledge and governance.[4] Tom Boellstorff similarly argues that big data should be "taken down a notch" from a unitary phenomenon and instead framed as a conceptual rubric so that theorists can address "issues of time, context, and power." Geoffrey C. Bowker has argued that raw data is an oxymoron because all databases are richly contextual, with specific temporalities, spatialities, and materialities that require critical attention.[5] These critical historical analyses are writing against both industry hype that frames big data as a new service that can be sold off the shelf and other theorists who have identified big data with the "end of theory" and the rise of hypothesis-free science.[6]

It is clear that *something* about the ways that we create and use knowledge has changed. Despite the prefix, "big" data is not simply data massively scaled up in quantity. Rather, the bigness of big data points to the newly expansive capabilities to connect disparate datasets through algorithmic analysis, forging unpredictable relationships between data collected at different times and places and for different purposes. Although it is impossible to identify a single characteristic that makes data "big," the emergent properties of massive, connected, and heterogeneous datasets are different than those of "traditional" datasets that remain restricted to a context much closer to their original point of collection. If we focus on its effects rather that its properties, Viktor Mayer-Schönberger and Kenneth Cukier (2013) suggest that big data can be thought of as the "things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value, in ways that change markets, organizations, the relationship between citizens and governments, and more."[7] Rob Kitchin (2014) similarly writes that as opposed to traditional large datasets such as censuses, "Big Data is characterized by being generated continuously, seeking to be exhaustive and fine-grained in scope, and flexible and scalable in its production."[8] danah boyd and Kate Crawford (2012) argue that treating

---

visualization. In: *Proceedings of the 8th conference on Visualization'97*, IEEE Computer Society Press, p. 235–ff. Available from: http://dl.acm.org/citation.cfm?id=267068 (accessed 9 February 2016).

[4] Crawford K, Gray ML and Miltner K (2014) Critiquing Big Data: Politics, Ethics, Epistemology. *International Journal of Communication* 8(0): 10.

[5] Bowker GC (2005) *Memory practices in the sciences*. MIT Press Cambridge, MA. See also: Gitelman L (ed.) (2013) *Raw Data Is an Oxymoron*. Cambridge, MA: MIT Press.

[6] Anderson C (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *WIRED.* Available from: http://www.wired.com/2008/06/pb-theory/ (accessed 9 February 2016).
See also: Graham M (2012) Big data and the end of theory? *The Guardian*, 9th March. Available from: http://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory (accessed 9 February 2016).

[7] Mayer-Schönberger V and Cukier K (2013) Big Data: A Revolution that Will Transform how We Live, Work, and Think. Houghton Mifflin Harcourt.

[8] Kitchin R (2014) Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1(1): 2053951714528481.

big data as a route to theory-free objective knowledge simply because of its scale risks *apophenia*, or "seeing patterns where none actually exist, simply because enormous quantities of data can offer connections that radiate in all directions."[9] Avoiding *apophenia* requires attending to the conditions, limits, and histories of our machine tools.

At the heart of big data's consequences is thus both a change in scales—speed, capacity, continuous generation—as well as a change in the relationality, flexibility, repurposing, and de-contextualization of data. Although the reality is much messier, in an ideal sense big data infrastructures are designed with the intention of relating any set of data to any other set of data. This capability leads to sometimes headline-grabbing and head-scratching correlations as these techniques are integrated into everything from marketing to policing. At the same time, critics have highlighted how this fast-paced rollout has bulldozed thoughtful consideration of bias, statistical meaning, and grounded interpretation. To underscore this point, Harvard Law student Tyler Vigen mocked the hype surrounding big data through a website called Spurious Correlations, where he showed data correlating seemingly disconnected topics like "Number of people who drowned by falling into a pool" with "Number of films Nicolas Cage appeared in."[10]

This dramatic move toward building and understanding datasets as potentially infinitely relational is at the heart of the epistemic shift toward algorithmic knowledge production. Insofar as research ethics regulations and conceptual frameworks are responding to the conditions of knowledge production that precede this shift, we should expect to find mismatches between the big data research and the extant research ethics regimes. Big data stretches our concepts of ethical research by moving inquiry away from familiar categories of harm, such as physical pain or psychological distress, to other categories, such as perennial surveillance, individual and group discrimination, and "predictive privacy harms," where privacy invasions occur through inference rather than direct collection of personal data (Crawford and Schultz, 2013).[11] The massive aggregation of research data also turns our concept of a human subject away from individuals and toward distributed groupings or classifications. Together, these shifts are hard to quantify, creating mismatches between ethics paradigms focused on traditional categories of harm and thereby frustrating attempts to account for and ameliorate novel types of informational harm.[12]

The Council has identified a number of areas where complex negotiations about the relationship between ethics and epistemology are shaping the uptake of big data. For example, ethical regimes around data tend to implicitly assume that data stays put within a specific context and temporal

---

[9] boyd danah and Crawford K (2012) Critical Questions for Big Data. *Information, Communication & Society* 15(5): 662–679.
[10] http://www.tylervigen.com/spurious-correlations
[11] Crawford and Schultz, 2013.
[12] Zwitter A (2014) Big Data ethics. *Big Data & Society* 1(2): 2053951714559253.

timeframe. Informed consent, the most common point of ethical exchange between researchers and subjects, is front-loaded in the research process: consent occurs at the point of collection before a subject's body or data is used for research.[13] Yet as Rasmus Helles and Klaus Jensen (2013) argue, "data creation is a process that is extended in time and across spatial and institutional settings."[14] This is true for all sorts of data[15], but the infrastructures of big data make this temporal stretching of data ever more obvious. As it becomes cheaper to collect, store, and re-analyze large datasets, it has become clear that informed consent at the beginning of research cannot adequately capture the possible benefits and (potentially unknown) risks of consenting to the uses of one's data. Even experts struggle to determine what they are consenting to because big data, and its sister disciplines such as biobanking, now stretch the utility of data far beyond the horizon.[16] If the most prominent point of ethical exchange between researchers and subjects is thus becoming questionable or even impossible on principle, then what is to be done? The debates about how to handle consent in big data research have generated a remarkable diversity of positions, ranging from jettisoning informed consent altogether in non-interventional research that makes use of already existing or passively collected datasets, to calls to develop statistical methods and research infrastructures should be developed to accommodate more dynamic notions of consent.[17,18,19,20,21] Prominent efforts such as the Personal Genome Project, 23andMe, and Patients Like Me have likewise sought to leverage informed consent to change research subjects' approach to informational risk and privacy, in service of their end goal of putting comprehensive medical datasets closer to the public domain. Notably, those who are attempting to build research infrastructures to take full advantage of big data methods are often staking a complementary claim about developing new ethical modes that emphasize the right or duty of individuals to share useful medical or personal data widely. These new research infrastructures and ethics claims often intentionally fall outside of the contested, but largely robust, research ethics regulations developed over the last 70 years in the U.S., which is heavily oriented toward federal funding and university-based researchers. To the extent that human-subjects research has strong

---

[13] Metcalf J (2015) *Human-Subjects Protections and Big Data: Open Questions and Changing Landscapes.* Council for Big Data, Ethics, and Society. Available from: http://bdes.datasociety.net/council-output/human-subjects-protections-and-big-data-open-questions-and-changing-landscapes/ (accessed 11 December 2015).

[14] Helles R and Jensen KB (2013) Introduction to the special issue: 'Making data-Big data and beyond'. *First Monday* 18(10). Available from: http://ojs-prod-lib.cc.uic.edu/ojs/index.php/fm/article/view/4860/3748 (accessed 9 February 2016).

[15] Bowker (2005).

[16] Reardon J (2013) Should patients understand that they are research subjects? *San Francisco Chronicle*, 3rd March. Available from: http://www.sfgate.com/opinion/article/Should-patients-understand-that-they-are-research-4321242.php (accessed 10 February 2016).

[17] Rothstein MA and Shoben AB (2013) Does Consent Bias Research? *The American Journal of Bioethics* 13(4): 27–37.

[18] For an example of similar claims in social science see: Neuhaus F and Webmoor T (2012) Agile Ethics for Massified Research and Visualization. *Information, Communication & Society* 15(1): 43–65.

[19] Committee on Revisions to the Common Rule for the Protection of, Board on Behavioral, Cognitive, and Sensory Sciences, Committee on National Statistics, et al. (2014) *Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences.* Available from: http://www.nap.edu/read/18614/chapter/1 (accessed 21 October 2015).

[20] Marsolo K, Corsmo J, Barnes MG, et al. (2012) Challenges in creating an opt-in biobank with a registrar-based consent process and a commercial EHR. *Journal of the American Medical Informatics Association* 19(6): 1115–1118; Rothstein MA and Shoben AB (2013) Does Consent Bias Research? *The American Journal of Bioethics* 13(4): 27–37.

[21] McGuire AL, Basford M, Dressler LG, et al. (2011) Ethical and practical challenges of sharing data from genome-wide association studies: The eMERGE Consortium experience. *Genome Research* 21(7): 1001–1007.

protections, it is usually in the medical context, and even then those protections are often not well-suited to the specificities of biomedical or public health research utilizing big data techniques.

John P. A. Ioannidis describes this aspect of the big data revolution as the oxymoron of "research that is not research."[22] Repeatedly in conversations around ethics regulations we see data science slipping into an uncanny valley between "research" and "not research," where the algorithmic production of knowledge technically falls outside of regulatory or philosophical definitions of scientific research yet looks for all intents and purposes to be the production of new knowledge. Nowhere is this clearer than the question of whether big data research should be treated as human-subjects research governed by the Common Rule. As the rise of data science has drawn mathematics and computer science disciplines closer to sensitive human data, researchers have debated whether data science should be treated as if it were human-subjects research qualifying for the core research ethics protections applicable to other disciplines. However, the large majority of data science that makes use of existing datasets largely avoids these regulations by not quite qualifying as "human subjects" because it does not involve an intervention in a subject's life, and not qualifying as "research" because it does not collect new data in pursuit of generalizable knowledge.[23,24]

The play between "research" and "not research" was at the center of the most notorious case of data ethics in private enterprise, the "Facebook emotional contagion study," published in the *Proceedings of the National Academies of Science* in 2014.[25] In this paper, Facebook data scientist Adam Kramer and Cornell University social scientists Jamie Guillory and Jeff Hancock present the results of experimentally modifying the Facebook feed algorithm for 689,003 people. They demonstrated that the negative or positive emotional valence of the posts that show up on users' News Feed alters the emotional valence of the posts that the user him or herself makes. This supported the hypothesis that emotional contagion—spreading emotional states through social contact—occurs on a massive scale (though with relatively small individual effects) in social networks. It also launched a public controversy about big data research ethics, in large part because Facebook users whose feeds were altered did not directly consent to participating but rather were covered by Facebook's terms of services when they signed up for the service. This study brought a remarkable range of criticism and defense from thought leaders and scholars, highlighting the unsettled nature of these matters even among the experts. Some were deeply critical of Internet services experimenting on their users'

---

[22] Ioannidis JPA (2013) Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research. *The American Journal of Bioethics* 13(4): 40–42.
[23] Metcalf and Crawford, 2016.
[24] King JL (2015) Humans in computing: growing responsibilities for researchers. *Communications of the ACM* 58(3): 31–33.
[25] Kramer ADI, Guillory JE and Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–8790.

emotional states with thin standards for consent.[26] Others claimed that the very possibility of algorithmically driven Internet services depends on the A/B testing at the heart of Facebook's experiment, a practice that is core to research as well as innovation.[27] Indeed, the major difference between this incident and standard practice inside the Internet economy is that the results of the Facebook experiment were made public through publication in a scientific journal. It also served as an opportunity to call attention to the importance of dialogue with corporations about ethics review.[28] The emotional contagion study controversy thus illustrated a series of substantive disputes that are central to data ethics inquiry moving forward. If iterative, algorithmic research methods do not have a temporally distinct beginning or conclusion, then at what point should ethics controls be applied? Does the A/B testing at the heart of Internet services and other big data endeavors count as an "intervention" in the life of human subjects, and does it therefore count as "research"? How does this differ from other interventions that are common in industry (e.g., headline manipulation) and research (e.g., public health studies)? Which disciplinary or professional norms and values should shape new data science practices? To what extent should academic researchers partner with researchers within private enterprises that have troves of data collected under conditions that do not rise to the level of informed consent? How should the practice of data analysis be governed, regardless of sector?

Both the present version of the Common Rule and the proposed revisions explicitly assume that any research methods using existing public datasets pose such miniscule risks to individual human subjects that researchers should not face scrutiny by IRBs.[29] Yet this is not adequate for understanding data analytics techniques that can create a composite picture of persons from widely disparate datasets that may be innocuous on their own but produce deeply personal insights when combined.[30] Similarly, the assumption (codified in law) that individual harm is the only type of risk that researchers are required to track and mitigate undercuts our ability to see harms that affect communities or produce "networked harms," such as data discrimination.[31] In the Council's public comment regarding the proposed revisions to the Common Rule submitted to the Department of Health and Human Services,

---

[26] Grimmelmann J (2015) Ethical Culture Clashes in Social Media Research. *2d.laboratorium*. Available from: http://2d.laboratorium.net/post/108480841510/ethical-culture-clashes-in-social-media-ressearch (accessed 6 November 2015).

[27] Meyer MN (2015) *Two Cheers for Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation*. SSRN Scholarly Paper, Rochester, NY: Social Science Research Network. Available from: http://papers.ssrn.com/abstract=2605132 (accessed 19 October 2015) Meyer MN and Chabris CF (2015) Please, Corporations, Experiment on Us. *The New York Times*, 19 June 2015. Available from: http://www.nytimes.com/2015/06/21/opinion/sunday/please-corporations-experiment-on-us.html (accessed 19 October 2015).

[28] boyd, d. (2015). "Untangling Research and Practice: What Facebook's "Emotional Contagion" Study Teaches Us." *Research Ethics* 12(1): 4-13.; Crawford K (2014) The Test We Can—and Should—Run on Facebook. *The Atlantic*. Available from: http://www.theatlantic.com/technology/archive/2014/07/the-test-we-canand-shouldrun-on-facebook/373819/ (accessed 21 January 2015); Calo R (2013) Consumer Subject Review Boards: A Thought Experiment. *Stanford Law Review Online* 66: 97; Polonetsky J, Tene O and Jerome J (2015) Beyond the Common Rule: Ethical Structures for Data Research in Non-Academic Settings. *Colorado Technology Law Journal* 13.

[29] Metcalf and Crawford (2016).

[30] Crawford K and Schultz J (2014) Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev*. 55: 93.

[31] boyd danah, Levy K and Marwick AE (2014) The Networked Nature of Algorithmic Discrimination. In: Gangadharan SP, Eubanks V, and Barocas S (eds), *Data and Discrimination: Collected Essays*, New America, pp. 43–57. Available from: http://www.newamerica.org/downloads/OTI-Data-an-Discrimination-FINAL-small.pdf.; Levy, K and boyd d (2014) "Networked Rights and Networked Harms." Presented at *Privacy Law Scholars Conference*, 4 June 2014.

the Council notes that regulators would be mistaken to assert that research should be exempt from oversight on the basis of the public status of a dataset.[32] The expansive relationality of datasets should change how we anticipate and mitigate harms: whether a dataset is public is of much less consequence than what is done with it and what other datasets it is merged with.

The interoperability of datasets also creates new types of pressure on privacy that are not simply a matter of disclosure risk from a single dataset or individual control of personal data. Datasets that would otherwise be innocuous and adequately anonymized on their own can be used to reveal highly sensitive data when analyzed in conjunction with other datasets. Privacy of personal data in large datasets thus depends not only on the safeguards applied to the dataset itself, but also the safeguards used in all other auxiliary datasets that could be used by an adversary.[33]

III. RECOMMENDATIONS

The Council offers the following recommendations to begin to address the conceptual and institutional disjunctions between big data and existing norms and infrastructures of research ethics. Our overarching recommendation is that entities engaged in data science practice and regulation should endeavor to create more spaces for inclusive deliberation around data ethics. At this stage, data ethics scholarship is largely engaged in identifying broad questions and establishing forward-looking research agendas. Thus, every organization that has stakes in data ethics should aim to build capacity for collaborative deliberation among ethicists, social scientists, data scientists, regulators, and community members in order to chart a diversity of paths forward.

Policy matters

1. Ensure the Common Rule clearly addresses regulation of data science.

As discussed above, the Council has reservations about the proposed revisions to the Common Rule centered on the assumption that public datasets pose few ethical risks to human subjects. The long time-scale and highly networked nature of big data means that the status of a dataset as "public" or "already existing" is a poor proxy for measuring possible informational harms. While the Council is divided over whether data science should fall under the Common Rule at

---

[32] Metcalf J (2016) *Letter on Proposed Changes to the Common Rule*. Council for Big Data, Ethics, and Society. Available from: http://bdes.datasociety.net/council-output/letter-on-proposed-changes-to-the-common-rule/ (accessed 11 January 2016).
[33] Narayanan A, Huey J and Felten EW (2016) A Precautionary Approach to Big Data Privacy. In: *Data Protection on the Move*, Springer, pp. 357–385. Available from: http://link.springer.com/chapter/10.1007/978-94-017-7376-8_13 (accessed 9 February 2016).

all, we do have a consensus that research ethics regulations should be built on sound empirical grounds, which clearly indicate that even seemingly innocuous public datasets can disclose highly personal data when networked with other datasets. Therefore, ethics regulations should focus on what will be or could be done with datasets. Furthermore, we note that the Common Rule has historically had significant influence even outside of its official purview because it establishes norms and expectations that operate across industry and academia and substantially informs the ethics training that most professionals receive at university. We express concern that the new rule's exclusion of all data science from ethics regulations on the basis of empirically suspect assumptions would weaken efforts to develop ethics review even outside of the Common Rule's purview.

2. Seek ways to facilitate new approaches to ethics review inside academia and industry.

Although they have been historically essential for introducing ethical review to scientific research, university-based IRBs are a cumbersome and contested method for ethics regulation. Their scope is limited to individual harms by statute, thus excluding the distributed harms that are more relevant to big data research. By avoiding discussion of distributed social outcomes and group harms, IRBs have functionally limited ethics engagement to legalistic matters of compliance. As big data researchers increasingly follow the interesting datasets outside of the university setting and into industry (and back again), more research will fall outside of the purview of IRBs or will involve complicated collaborations across academia and industry, as happened in the Facebook emotional contagion study. Thus, we recommend that policy makers and leaders in enterprise seek ways to encourage experimentation around ethics review inside and outside of university settings. Although many question the utility of IRBs in their current iteration for regulating data science, the function of pausing for independent review and deliberation is indispensable. We encourage trying new approaches that consider potential group harms that may arise in big data research, taking account of power differences between researchers and subjects, and, where possible, including input from affected populations. For example, IRBs could be better calibrated to the conditions and norms of disciplines such as data science if panels had ad hoc members with subject-area expertise.

3. Develop mechanisms of ethical assessment calibrated to the practices of big data.

Ethics review practices in industry and academia can vary widely. For example, in industry, ethics review often happens after product development and before product launch in order to catch potential or unexpected risks. In academia, ethics review typically occurs earlier, at the point of a project proposal before any research occurs. Big data methods disrupt both approaches by collecting data at a much higher velocity and altering products iteratively. Therefore, while up-front ethical review can and should play an important part in big data analysis, it should not stop there. Rather, we should continue to look for social, structural, and

technical mechanisms to assess the ethical implications of a system throughout the entire development and analysis lifecycle. There should be more investment in developing technical interventions for the practical barriers to ethics review, such as algorithmic auditing and values-oriented formal verification. Better understanding of the organizational, cultural, and technical decisions that go into the development of big data systems is also needed and requires further research investment. Finally, it is important to consider critical junctures in big data research and practice where ethical considerations might be most fruitful. In industry, quality assurance testing might be a site where ethics issues can be meaningfully addressed. In academia, developing protocols for peer reviewers and conference committee meetings might be another key site for deeper investment.

4. Integrate data ethics concerns into NSF program solicitations and the grant-making process.

According to our analysis of projects funded by the NSF BIGDATA program, ethics matters are largely absent in project proposals beyond the requisite Broader Impacts section. As discussed above, few if any data science projects are reviewed by university IRBs, for better or worse. Thus, we recommend that NSF program directors seek opportunities to encourage ethical reflection by Principal Investigators (PI) and other grant recipients in programs utilizing big data research methodologies. This may include writing program solicitations to instruct applicants to account for broad ethical questions, such as provenance, disclosure risks, and what other plausible harms might come from networking their datasets with other datasets. Although support for the idea appears mixed, some have suggested that a revision to the Data Management Plan (DMP) requirement may include more explicit ethics consideration. DMPs have thus far focused on where data will be stored and shared, however they might plausibly accommodate other concerns moving forward. Including ethics in program solicitations and DMPs signals to peer reviewers that ethics are legitimate criteria by which to judge the funding-worthiness of a project. The NSF should also consider creating a solicitation for empirical and theoretical study of research ethics challenges and innovative solutions in data science, perhaps as a collaboration between the CISE and SBE Directorates. Collaborative proposals between social and technical researchers and projects that embed social researchers in technical contexts are valuable approaches. Another avenue for encouraging ethics engagement by PIs is including ethics panels at the mandatory PI meetings at the NSF.

Pedagogical Interventions

1. Create and distribute high quality data ethics case studies that address difficulties faced by data scientists and practitioners.

Our survey of Council members, available syllabi, and online resources such as the National

Online Ethics Center indicated a dearth of case studies addressing the dilemmas faced by data science researchers in practice. Case studies are a valuable pedagogical resource in applied ethics because they facilitate collaborative discussion inside the classroom and lengthier reflection as writing assignments. The Council decided to address this by soliciting researchers for case studies based on their experiences. The resulting case studies will be distributed via the Council website, journals, and the National Online Ethics Center database. The Council recommends that other entities, funders, and publishers make a point of developing case study resources moving forward. Although other models are also valuable, we chose to emphasize case studies representing actual dilemmas experienced by researchers. Our surveys of case studies in other disciplines indicated that hypothetical case studies tend not to capture the complexity of the decisions made by engineers as they address ethical, social, and political dynamics of their research and practice.

2. Develop and support data science curricula with integrative approaches to ethics education.

Big data is an inherently multi-disciplinary endeavor, and even more so when considering human data. However, the conditions of contemporary universities often work against interdisciplinarity in the classroom due to disciplinary silos and lack of common language and methods. There are promising pedagogical models that focus on the design process as a gathering point for multiple forms of expertise, such as Values in Design.[34] Other models emphasize justice and ethics as leverage for opening collaboration around empirical projects, such as the Science & Justice program.[35] Whether ethical considerations are treated as a stand-alone module or integrated into the very essence of the course, it is clear that ethics needs to be a cornerstone of big data education. Because moving big data research ethics forward will require a diversity of approaches, the Council recommends that funders support creative efforts at bridging intellectual and institutional divides across disciplines.

3. Train librarians to achieve and promulgate data science literacy.

To strengthen and expand the assistance that research librarians provide to researchers, support the development of their technical knowledge. Librarians' involvement in areas such as data sharing, metadata creation, data curation and preservation, and copyright, access, and legal use of information provides an opportunity for ethical discussions.[36,37] Topics may

---

[34] See: http://www.nyu.edu/projects/nissenbaum/vid/. See also: Knobel C and Bowker GC (2011) Values in design. *Communications of the ACM* 54(7): 26–28.

[35] Science & Justice Research Center (Collaborative Writing Group) (2013) Experiments in Collaboration: Interdisciplinary Graduate Education in Science and Justice. *PLOS Biol* 11(7): e1001619.

[36] Goodman A, Pepe A, Blocker AW, et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. *PLOS Comput Biol* 10(4): e1003542.

include whether Terms of Service must be followed and how; whether consent forms are sufficient in the case of unforeseen future aggregations and potential re-identification; the ethics of secondary data use, particularly for data acquired through company internships or jobs that have not received university IRB review; web scraping and crawling; and data reuse and manipulation. This additional technical knowledge helps librarians to build trust with researchers. Librarians also function as part of a triage system at some universities, helping to refer researchers out to the appropriate departments for issues beyond their scope of knowledge.

4. Strengthen ethics-oriented activities within professional associations.

Professional fields—such as medicine and law—often put ethics at the center of education, accreditation, and practice. While the targets of such efforts are primarily practitioners, the impact is felt throughout the entire field, including in the research community. Although there are plenty of professional associations addressing big data, there is no accreditation process for data scientists, nor are there normalized educational mandates. Although associations like the Association for Computing Machinery (ACM) have a Code of Ethics, most computer scientists are not aware of its existence. More work is needed at the professional association level to ensure ethical commitments in research and practice.

Network Building: Development of cultures of ethics engagement

1. Create hybrid spaces for ethics engagement.

Simply put, there are relatively few areas where ethicists and engineers mingle and develop professional ties. "Ethics" in the broadest sense is not simply deciding what is right and wrong, but indicates literacy about one's context and how one's decisions and actions affect others. Thus, while formal bodies are certainly required for ethical decision making, they are premised on the possibility of collaboration and networking between people with a diversity of expertise. Entities that wish to encourage data ethics—funders, universities, employers, and disciplinary associations—should treat networking and collaboration as necessary components of establishing ethics capacity.

2. Build models of internal and external ethics regulation bodies in industry.

While individual designers may consider ethics issues in their work, product and business decisions are generally centered on building the best product, requiring tradeoffs that aren't

---

[37] Borgman CL (2015) *Big Data, Little Data, No Data: Scholarship in the Networked World.* MIT Press.

always articulated formally. Likewise, industry research projects on social issues are often driven by product design needs rather than overall social needs or considerations. As a result, data ethicists and legal scholars have recently proposed models for ethics review bodies in industry that could serve a function similar to IRBs.[38] Ryan Calo suggested that businesses should consider "Consumer Subject Review Boards" that would enable users of data-driven Internet services to have some say over how their data may be used.[39] This approach would focus on incorporating review by representatives or members of affected communities into ethics governance structures. Omer Tene, Jules Polonetsky, and Joseph Jerome have argued for a two-track model of internal and external ethics review for data companies.[40] Though underexplored with respect to big data ethics, community IRBs may also provide insight into how to consider ethical dimensions of big data research. Yet consideration of an internal mechanism for ethics oversight similar to academia poses challenges, such as resistance to fitting emerging research into predetermined categories. Industry research on sensitive populations may already undergo review by legal, public policy, or engineering leaders, but defining what constitutes sensitive material can change in response to external factors, such as political issues having heightened emotion around an election. Industry leaders emphasize the value of flexibility in new or existing review mechanisms to accommodate the fluid and fast-moving nature of research projects and methods, and the importance of steering away from the attempted application of one strict set of standards to all departments or teams. Without internal, external, or legal repercussions, voluntary ethics review mechanisms could be difficult to enforce. Another approach would be to consider the structures necessary for fiduciary responsibility, including the possibility of board oversight and external auditing of practice. While it is not yet clear what might be the best way forward, more work is definitely needed to determine how explicit attention to ethics can be meaningfully integrated into industrial contexts.

3. Set standards for responsible cross-sector data sharing.

In response to academic researchers' frequent requests for corporate data, companies discuss potential ways to release their data to contribute to the scientific community. They must weigh potential benefits against past experiences of coming under public scrutiny for decisions to share anonymized data that backfired. Students who intern or work with companies have strict rules on use of the data they have access to, which includes the inability to share with others,

---

[38] Dove ES, Townend D, Meslin EM, et al. (2016) Ethics review for international data-intensive research. *Science* 351(6280): 1399–1400.
[39] Calo R (2013) Consumer Subject Review Boards: A Thought Experiment. *Stanford Law Review Online* 66: 97.
[40] Polonetsky J, Tene O and Jerome J (2015) Beyond the Common Rule: Ethical Structures for Data Research in Non-Academic Settings. *Colorado Technology Law Journal* 13.

but professors have cited instances in which students overlooked or misunderstood these requirements. Even after being scrubbed, sensitive corporate data is uncontrollable after release and can lead to identifiable information and legal problems following combination with other data sets. Discussions take place in search of ways to allow people to run queries and get answers without seeing the data. Restricting the data release for an IRB-reviewed academic purpose and prohibiting combination with other data sets is one possibility allowing greater control, but concerns remain about feasibility or potential backlash against the research community if a perceived violation of trust or privacy results. Vetting researchers for individual data release can also lead to perceived bias if elite institutions have greater leeway in gaining access to data sets than other researchers or members of the public. In other large-scale data-sharing efforts, such as the government's efforts around Census data, third party organizations have been created to help enable ethical and responsible data analysis. Such an avenue should be considered for other sensitive corporate data.

IV. Areas for further research

Data ethics is a quickly growing research area. The Council has identified important topics for future research agendas, and encourages federal funders and foundations to pay particular attention to these needs.

1. Should human data science be regarded as human-subjects research?

The Council has emphasized the disjunctions between the existing norms and institutions of research ethics and the epistemic conditions of data science. It is now clear that the methods of data science are at best an awkward fit for prevailing research ethics norms and regulations; what is not clear is how that gap should be addressed. At the heart of these disjunctions is the question of when big data analytics about human data should or should not be considered human-subjects research, and therefore fall under the purview of these norms and institutions. This is both a philosophical (what is morally owed to data science research subjects?) and practical matter (what present and future regulations should apply to data science?), and Council members hold a diversity of opinions. Both the philosophical and practical concerns present valuable research tracks. We anticipate that research ethics will undergo significant changes as big data methods continue to evolve, and funders should seek to stay in front of these changes in order to provide useful guidance to researchers, policy makers, and research subjects.

2. What are the quantifiable risks posed by correlative and/or predictive data research?

The proposal to revise the Common Rule addresses the long-running claim by social scientists that IRBs are a cumbersome way to regulate their research practices. In particular, IRBs demonstrate a pattern of over-regulation of non-biomedical research through a lens of maximal and hypothetical risk of informational harm. The Notice of Proposed Rulemaking (NPRM) seeks to provide clearer guidance to IRBs about scaling regulation to reflect empirical research about how to measure risk of informational harm. While this is welcome, it also has the effect of highlighting just how little is known about the specific risks posed by data science research methods. Data privacy and security researchers have demonstrated *how* informational risks are generated, but there are few accounts of how those risks should be weighted and mitigated. How should we engage with research that poses very small individual risks spread over an enormous number of people? What metrics are appropriate for defining informational risk when research subjects' privacy expectations may vary widely? Additionally, we note the need to develop metrics for useful theoretical terms that are becoming common in data ethics literature, such as "creepiness" or "sensitive." Such empirical and theoretical work must be a high priority if we are to understand the consequences for human subjects and develop trust over time.

3. Similarly, how should we account for the risk of sharing datasets when we cannot know what auxiliary datasets they will be combined (munged) with in the future?

Does the risk differ with public datasets? With biomedical data? The power and peril of large-scale data analytics is the capacity to combine datasets from highly different contexts with relative ease, enabling data to become (at least theoretically) infinitely flexible and perpetually available for repurposing. Most troubling to account for is the risk posed by unknown and unpredictable auxiliary datasets that can be used to re-identify personal data in a research dataset. This means that the risks faced by subjects of data research are not limited to the context and lifespan of the project itself. Researchers have not historically been accountable for such far-reaching consequences. Thus research ethics faces an unprecedented question: how should we account for and mitigate unknowable future risks resulting from a research project? This challenging question indicates an important area for future research, particularly focused on empirical measurement and/or consistent accounting of such risks, and developing iterative procedures for accounting and mitigation.[41]

4. How is big data redefining both when and how the public benefits from research? And what are more precise ways to assess public benefit or justice considerations in big data research?

---

[41] Wan Z, Vorobeychik Y, Xia W, et al. (2015) A Game Theoretic Framework for Analyzing Re-Identification Risk. *PLOS ONE* 10(3): e0120592.

Most conspicuously, because of the issue of relationality of datasets (see above), big data research may face significant challenges in weighing benefits of a stated research program, as well as comparing risks to benefits. The question of beneficence and justice may not be adequately judged at the outset of a research endeavor, but rather may require periodic or even frequent inspection. The issue of relationality and interoperability also begs the question of which publics or affected communities are most likely to benefit: sensitive attributes revealed in the process of merging and analyzing certain kinds of data may have implications for specific kinds of groups, upending original forecasts that predict improvements in a group or population's well-being. Needless to say, the challenge of scoping who benefits, when, or how is a significant one and requires deeper examination. A research agenda focused on beneficence and justice could have broad pedagogical value as well as practical impact in the development of governance structures in big data research.

5. How should data privacy and security scientists approach illicitly gained, publicly available data?

Illicitly gained datasets released openly on the Internet are a potentially rich source of research data, especially in computer security. Such datasets occupy an ambiguous space because they are "public" in the sense that anyone has access to them, but "not-public" in the sense that no one should have access to them. For example, some independent security researchers have collected large sets of hacked login and password data on the presumption that harm has already been done, and some good can come from analysis of individuals' security behaviors. In other cases, hacked datasets can have information that has nothing to do with security but would otherwise be unobtainable to those with legitimate research purposes. Despite the growing prevalence of this problem, there are no clear professional guidelines.

6. What are the options for self-regulation in data science?

The precursor disciplines of data science have not historically fallen under the purview of ethics review at universities, and it does not appear to be likely that data science will be routinely regulated in that fashion moving forward. Additionally, university-based IRBs have well-demonstrated shortcomings that reduce their effectiveness at fostering ethical data science methods. Most notably, university-based IRBs do not provide proactive advice about project structure and are forbidden from considering societal harms. Thus there is a need to innovate models of self-regulation and enforcement of professional norms outside of university-based IRB's. Currently, the most prominent examples of self-regulation are conference paper review committees that determine which papers are accepted to the prominent disciplinary conferences, sometimes utilizing ethical reasons to determine

rejections. However, this model is problematic because it occurs after the research is already conducted and the conferences do not provide ethical guidelines that researchers can rely on in advance. External IRBs, which independent and university-based researchers may contract with to review their research proposals, are better able than university-based IRBs to provide proactive advice about improving research. Additional research is needed to determine how proactive, iterative, and consistent ethical advice can be provided to data science researchers.

7. What resources are needed in the university context to encourage engagement with data ethics issues, particularly outside of the IRB?

For the many reasons enumerated above, university-based IRBs are a contentious route for regulating data science research ethics. Additionally, there are long-standing shortcomings with the IRB model, and thus alternative or complementary approaches should be encouraged. One promising route is utilizing research librarians. Universities are increasingly turning toward their libraries for data repository management and their librarians for counseling faculty and students on data management issues. Because they are often becoming an obligatory passage point for research data or DMPs, librarians may be well placed to provide data ethics guidance to researchers. Likewise, neutral interlocutors within and beyond the library are helpful for providing non-judgmental spaces to discuss ethical concerns arising from big data projects. For example, the emergence of a handful of university data clinics, based on the statistics clinic model, offer an informal, conversational drop-in space for researchers to address ethics issues in the context of their overall research goals. These data clinics lack the formal scrutiny of an IRB review, but may provide more relevant advice.

8. How can integrative approaches to data ethics be fostered in classroom environments? What pedagogical resources are needed?

Pedagogical best practices research indicates that integrative approaches to science and engineering ethics education are more effective than stand-alone modules and courses. Ethics can be taught as an aspect of problem solving, rather than as an external constraint. As data science grows, it will be best served by developing resources that enable instructors to foster ethics reasoning throughout the curriculum. Funders interested in science and engineering ethics pedagogy should emphasize the pressing need for resources in data science classrooms. The guidance received in the classroom may extend to future research decisions within industry, where graduates often take their careers.

9. What are the ecological and environmental impacts of a rise in big data research and industry?

Indefinitely maintaining enormous and ever-growing databases in a state ready for immediate

access requires remarkable quantities of energy and water, as well as causing direct and indirect carbon emissions. Although data centers run by major players in the tech industry have made strides in energy efficiency, the propagation of data intensive Internet applications drives energy usage in contexts outside of the control of industry leaders. A sustainable data industry will need to carefully consider the tradeoffs between performance and energy efficiency implicit in engineering decisions.

10. How can ethical issues be integrated into core technical research?

Technical researchers are coming together through workshops and conferences to think about how to turn ethical issues into tractable technical endeavors. The "Privacy Enhancing Technologies" (PETS) community is already quite strong, while a new group has emerged recently under the banner of "Fairness, Accountability, and Transparency in Machine Learning" (FATML).[42] Such approaches drive ethical issues into the core of technical research, raising ethical issues in technical communities through algorithmic innovation and mathematical analysis. Although this work is nascent, there is significant opportunity for high impact research in this area.

11. What motivates data scientists—and their colleagues and employers—in industry to establish ethics processes? Which ethics review structures do and do not work inside industry?

Industry leaders have indicated that a diversity of motivating factors are involved in the development of data ethics processes in private enterprise. As discussed above, some of the large industry players have begun to develop substantive ethics programs and that has provided preliminary insight into what motivates ethical programs in industry and what types of programs may work. Industry researchers and their employers have differing motivations for engaging with research ethics than those in academia, and this sometimes causes conflict and at other times prompts collaboration. A more concerted research program is needed to track these changes and account for which programs are effective. There is also a need for models of ethical accountability that can be translated between large and small companies. The effectiveness of ethics review programs in industry is an area ripe for social science and/or business management research.

12. What is the proper purview of "research ethics" as a topic in the age of big data?

---

[42] http://www.fatml.org/

A widespread frustration with "privacy" as the common regulatory, legal, and ethical framework is that it has been forced to carry too many undifferentiated concerns about the long-run consequences of big data analytics. It appears that data ethics, and particularly research ethics, has now taken up some of that slack. However, it would be similarly problematic to expect research ethics to function as a catch-all for the social, ethical, and political challenges of big data because not all ethically contentious uses of data analytics are directly related to research.

V. CONCLUSION

The Council for Big Data, Ethics, and Society has identified a broad need for development of research ethics norms and institutions calibrated to methods and assumptions of large-scale data analytics. As data science continues to mature rapidly into an indispensable component of scientific research and industrial practice, it is critically important that we nurture sustained conversations about the ethical, social, and legal complexities of data analytics. We have identified a number of disjunctions between the epistemic conditions of data science and the familiar norms and infrastructures of research ethics and regulations. Foundational assumptions about how research is done and what constitutes human-subjects research have been deeply contested by big data, and thus our tools for regulating research ethics are now miscalibrated.

However, this need not be a source of retrenchment around familiar models of research ethics. The development of large-scale data analytics can also provide fodder for productively rethinking the structures of research ethics and regulations while also reaffirming basic principles of respect, justice, and accountability. We have a historically unique opportunity to shape for the better the trajectory of a technoscientific revolution. Scholars, funders, and regulators should seek opportunities for collaborative, integrative, and innovative approaches to ensuring that big data analytics in science and industry are responsive to foundational ethical concerns.

Council reports have examined big data projects supported by the National Science Foundation; the history of ethical codes from mid-20th century; debates about personal genomics, the use of student data for personalized learning, and social media algorithms; the history of data management plans as a tool for integrating ethics principles; an examination of data ethics courses and modules in fields such as statistics, business, computer science, and journalism; the application of human-subjects protections from the social sciences to big data analytics; and proposed changes to the Common Rule by the U.S. Department of Health and Human Services. The Council is currently building an expanded network as well as a collection of data ethics case studies. For complete information and outputs visit: http://bdes.datasociety.net/.

Principal Investigators:

danah boyd, Data & Society / Microsoft Research (co-PI)

Kate Crawford, Microsoft Research / New York University (co-PI)

Geoffrey C. Bowker, University of California, Irvine (co-PI)

Helen Nissenbaum, New York University (co-PI)

Council Members:

Alessandro Acquisti,
Heinz College, Carnegie Mellon
University

Mark Andrejevic,
Pomona College

Solon Barocas,
Princeton University

Edward Felten,
Princeton University

Alyssa Goodman,
Harvard University

Rachelle Hollander,
National Academy of Engineering

Barbara Koenig,
University of California, San Francisco

Eric Meslin,
Indiana University Center for Bioethics

Arvind Narayanan,
Princeton University

Alondra Nelson,
Columbia University

Paul Ohm,
University of Colorado Law School

Frank Pasquale,
University of Maryland

Seeta Peña Gangadharan,
London School of Economics and
Political Science/ Data & Society

Latanya Sweeney,
Harvard University

Sharon Traweek,
University of California at Los Angeles

Matt Zook,
University of Kentucky

Staff:

Emily F. Keller, Project Coordinator

Jacob Metcalf, Post-doc

Supported by:

NSF Grant# 1413864