Data&Society

Supporting Ethical Data Research: An Exploratory Study of Emerging Issues in Big Data and Technical Research

Working Paper 08.04.2016

DANAH BOYD, EMILY F. KELLER, BONNIE TIJERINA

Purpose

New and complex data sets raise challenging ethical questions about risk to individuals that are not sufficiently covered by computer science training, ethics codes, or Institutional Review Boards (IRBs). The use of publicly available, corporate, and government data sets in research projects may reveal human practices, behaviors, and interactions in unintended ways, creating the need for new kinds of ethical support.

Technical researchers using this data in their research are navigating issues and making ethical decisions in ways that are not often taught in their discipline. Many have only their peers to turn to for difficult questions that could have long-term impacts on their research or reputation. Research librarians are one set of actors within an ecosystem of support that university researchers rely on during their research. As providers of a growing set of Research Data Management (RDM) services, librarians help researchers meet federal mandates to produce Data Management Plans (DMPs) for their projects. In addition, the values of privacy, ethics, and equitable access to data are core to librarianship, making this partnership unique.

This project was initiated to surface complex issues faced by researchers and to explore current and potential interactivity between librarians and computer science researchers at different phases of the research process as they come across questions of privacy, potential harms, and security in data

Workshop materials related to this report can be found at
http://datasociety.net/output/supporting-ethics-in-data-research/

storage. We conducted interviews and campus visits with academic computer scientists and librarians to surface emerging ethical issues brought about by big data research, to examine the ecosystem of formal and informal support systems for researchers, and to consider the role of research librarians in assisting technical researchers as they navigate ethical issues.

This report provides valuable insights into the current state of collaboration between librarians and computer science researchers on issues of "big data"[1] ethics. Statements and assertions represent information provided by participants, in combination with a literature review and additional formal and informal research. This report is not meant to be conclusive or comprehensive about all data science research, as we purposefully limited the scope of our work to a narrow band of institutions and actors. Yet, our findings do offer important insights that open up challenging questions and require future exploration.

Methodology

Given our interest in understanding when and how librarians and computer scientists can work together, we began by mapping out research institutions that had both strong computer science programs doing big data work and research libraries that had begun exploring ways to support researchers on privacy issues. We narrowed potential sites of inquiry by focusing on schools that had information science programs (which often bridges computer science and librarianship thinking) and researchers who had been thinking about ethics issues in different ways. We selected eight R1 universities, all of which were publicly funded. We interviewed nine researchers and 12 librarians at those institutions. After hearing about a particular project related to what we were investigating, we did an additional interview with a researcher at a private university. Our interviews were semi-structured phone and Skype interviews.

After doing these interviews, we conducted two on-campus group workshops with 13 people each at two of the eight universities. Interviewees generally represented academic libraries or were professors trained in computer science. In-person workshops also included technical faculty who also had appointments in information science, statistics, and social science departments, as well as students and information technology specialists. We have anonymized the individuals and institutions that participated in this research and removed identifying details.

Prior to conducting interviews and workshops, we reviewed literature to examine prior cases of data ethics controversies, campus resources and guidance for supporting data ethics and data management, the expanding role of academic libraries in providing RDM services, and the ethics trainings, courses,

1 For this purposes of this study, we left "big data" loosely defined. The researchers we talked with all worked with large quantities of data using and contributing to a variety of computer science subfields (e.g., information retrieval, machine learning, security, etc.).

and requirements at each university. We also researched the ethics codes of three professional computer science associations: the Association for Computing Machinery (ACM), the Association for the Advancement of Artificial Intelligence (AAAI), and the Institute of Electrical and Electronic Engineers (IEEE).

Finally, combining the interviews and literature review, we mapped the network of formal and informal support for those researchers, identifying the gaps, and looking for what, if any, role the library could play in supporting these researchers at various points of the research process.

Key Findings

There are new ethical dilemmas that computer science researchers are faced with when they conduct big data research, especially research using social media data or scraping "public" information off the internet. Some researchers expressed concern that results from data aggregations could cause harms to individuals or communities in ways that are not adequately addressed by current regulations, ethics codes, or best practices in their field. Concerns around consent, use of secondary data, data sharing, and data storage are emerging as issues for some technical researchers.

Ethical decision-making takes place in a range of circumstances and phases within a research project. While high-profile violations and controversies receive the most attention, day-to-day ethical quandaries are often buried within routine tasks such as data management and sharing protocols, web scraping and publishing of public data, anonymization of research subjects, and storage procedures for confidential information. Questionable ethics decisions often signify lack of clear communication or coordination rather than malicious intent. However, competitiveness, assumptions of weak enforcement mechanisms, varying belief systems about what is right within a rapidly changing landscape, and social norms that support violating certain rules (such as Terms of Service) also play a role in generating common behaviors that cause ethics concerns. Below are a few examples of complex ethical quandaries that computer science researchers often face.

Web Scraping and Terms of Service

Acquisition of online public data carries Terms of Service (TOS) requirements that raise logistical and ethical challenges about issues like replication, identification, and consent.[2] For example, Twitter prohibits passing on tweets after users have deleted them, in turn promising this protection to users. However, research norms differ from TOS rules, with many researchers continuing to use data containing subsequently deleted tweets, as re-doing a large study to eliminate them would be

2 Danyel Fisher, David W. Mcdonald, Andrew L. Brooks, and Elizabeth F. Churchill. "Terms of Service, Ethics, and Bias: Tapping the Social Web for CSCW Research," *CSCW 2010*, February 6–10, 2010, Savannah, Georgia: 603-606. ACM 978-1-60558-795-0/10/02.

logistically challenging and there is no mechanism of notification to researchers when an individual tweet has been removed. This raises questions about the practicality and enforceability of the TOS. Researchers do not always understand or follow the TOS, with students often trusting their advisor or choosing the most convenient path in lieu of clear institutional guidelines, oversight, or legal ramifications. In many cases no one checks whether the TOS were followed, providing a disincentive for making extra efforts to adhere to the rules.

In addition, some researchers feel that politicians, as public figures, should not have the right to prevent their deleted tweets from future use, and should be held to a higher standard of transparency. In the same vein, some researchers feel that government agencies have no right to prevent public use of data on their sites through TOS restrictions, given that they serve the public.

Growing interest in web scraping of online data raises ethics questions about issues such as copyright, control, and ownership. Companies that publish data often seek a balancing act between making their data findable and trying to prevent mass downloads. Researchers may use a separate email address when they scrape content from the internet, and are taught to "pretend to be a browser." However, web crawlers transmit an indication to the source that a site has been crawled when machines are hit repeatedly, which may get researchers in trouble. A librarian stated the library has a role in teaching people how materials are licensed and the rules of mass downloading of data, as well as negotiating access.

As for potential repercussions for researchers' violations of online TOS, the risk calculation for a university is whether they could be sued, for the researcher it is whether the research can be done without violating the TOS (or whether they'll get caught), and for the library it is whether that resource is open.

Consent

The traditional model of seeking informed consent at the beginning of a research study is often insufficient when it comes to big data research. There is an ongoing debate about whether big data research should fall under the Institutional Review Board (IRB) and oversight mechanisms for human subjects research, such as informed consent requirements, when the connection to humans is indirect.[3] Consent forms for big data research may offer a false sense of protection when the researcher cannot prevent the data collected from becoming part of a future aggregation that threatens the subject's anonymity.

When researchers gather public online and social media data, consent forms are often bypassed

3 Jacob Metcalf, Emily F. Keller, and danah boyd. "Perspectives on Big Data, Ethics, and Society," Council for Big Data, Ethics, and Society report, May 23, 2016. Accessed May 26, 2016. http://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/

altogether based on the assumption that posting to a public forum constitutes authorization for data reuse. Researchers often view publicly accessible data as observational data and offer weaker protections to the people who are represented than traditional protections for studies that involve a clear experiment. However, many social media users do not intend or foresee having a public audience when communicating socially.[4] The increased visibility can lead to harm. For example, publishing quotes from small public forums in which patients discuss their medical conditions could lead to re-identification by searching for the exact quotes through search engines.[5] Expectations around consent in a changing landscape of personal data exchange via social media has unwritten components that go beyond user agreements. Although the legal agreement that a social media site like Facebook has with users may permit experimentation, the tacit agreement with users and implicit understanding and expectations may be quite different.

Following the backlash to the Facebook "emotional contagion study" that involved manipulating thousands of users' News Feeds to assess whether seeing more negative or positive posts influenced one's updates, there have been increased calls for informed consent, as well as skepticism toward big data research.[6] Researchers are concerned that the fallout from the study is having a restrictive impact on research, from data access to publishing. One professor who is using public data said he is facing extreme opposition, discomfort, and calls for consent from peers and publication venues, which he attributes to reactions to the Facebook study. "There's a very good chance that this paper will die, even though we're excited about it and think it's kind of cool and has no real negative impact, but people don't really want to touch it," he said. Another professor said, "The problem is that so much of the data that we all want to use is locked up in corporations, so, one way or another, if we don't solve this problem we all lose big time."

In deciding whether to seek informed consent, researchers must weigh if doing so could unintentionally reduce the quality of their research.[7] One professor explained, "There's tension both in terms of helping the subject understand what data is being used that they are generating, but also wanting to preserve a semi-clean, externally valid population." For example, one researcher who uses data from communications devices to infer characteristics about subjects said that explaining the details of the study to participants would be problematic, potentially leading them to change their behavior. This could cause the appearance of having a selected and abnormal sample as he seeks to develop new measures for policy makers based on the results.

4 boyd, danah, *It's Complicated: The Social Lives of Networked Teens*, Yale University Press, 2014.

5 Jessica Vitak, Katie Shilton, and Zahra Ashktorab. "Beyond the Belmont Principles: Ethical Challenges, Practices, and Beliefs in the Online Data Research Community," *CSCW '16 Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*: 941-953. http://dl.acm.org/citation.cfm?doid=2818048.2820078

6 James Grimmelmann coordinated a bibliography of relevant articles at: http://laboratorium.net/archive/2014/06/30/the_facebook_emotional_manipulation_study_source His approach to this issue can be found here: Grimmelmann, James, "The Law and Ethics of Experiments on Social Media Users," 13 Colo. Tech. L.J. 219, 2015; University of Maryland Legal Studies Research Paper No. 2015-15, May 7, 2015. Available at SSRN: http://ssrn.com/abstract=2604168

7 Graham Crow, Rose Wiles, Sue Heath, and Vikki Charles, "Research Ethics and Data Quality: The Implications of Informed Consent," *International Journal of Social Research Methodology*, Volume 9, Issue 2, 2006; Special Issue: Quality in Social Research: 83-95. DOI:10.1080/13645570600595231

Secondary Data Use

Secondary use of human subjects data collected by a third party falls into an IRB gray area, as it is considered "exempt" and not reviewed by IRBs since the data was already collected. However, some researchers consider that a loophole and advocate for greater IRB oversight of this frequent practice due to the threat of re-identification or privacy violations that become possible through the continued analysis or aggregation of data. The line between observational or unobtrusive data collection and experimental interventions can be fuzzy. Many researchers utilize secondary data, saying it does not make sense to omit already published data from their studies, while others fear the increased exposure could legitimize studies with unknown research methods or further impact human subjects.

Some students attempt to use data gathered at corporate jobs or internships, which may include a Non-Disclosure Agreement (NDA) or partial employment structure, and are unaware of the limits once they return to campus. One professor said, "I feel like some of the biggest ethical violations are when I talk to students who did some internship and then maybe they have their hard drive with a terabyte of super sensitive data on it that they just brought back with them and didn't really realize that this is not a good thing to do." This can also create an oversight problem when students use corporate data in their theses, linking the university's name to research that was not reviewed by its IRB.

Those we spoke with cited IRBs as an important tool for oversight, but many researchers said the IRB had an insufficient understanding of key technical issues. One professor said, "To what extent do you have any obligation to delve into the details of how the data was originally collected?...Our IRB will say, 'Oh, you're just using a data set someone else created. That's fine. It's not human subjects research.' However, what if the way it was collected was not okay in the first place?" Many computer science researchers feel as though they know more about the ethical challenges presented by their work than oversight bodies such as IRBs. In asking researchers about computer science professors who serve on IRBs, we could only find one example that anyone even knew about.

Data Sharing

There is a growing set of requirements for sharing raw data with journals for replicability, and sharing and disseminating federally funded research with the public for potential reuse.[8] Some professors are looking towards providing scholarly credit for the creation of reusable data sets, which could be counted like a publication, or having PhD students learn to do research initially by replication. Interest in reproducibility has focused historically on whether repeating work produces the same results, while documented cases of data reuse for research are limited. Although data reuse is common in astronomy,

8 "Expanding Public Access to the Results of Federally Funded Research," the White House, February 22, 2013. Accessed May 31, 2016. https://www.whitehouse.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research

which has widely accepted data standards, it requires significant documentation of research context to work in other fields.

Fulfilling data sharing mandates is complicated, ambiguous, and potentially risky. Sharing requirements cause concern about potential privacy impacts such as re-identification. Some researchers fear that sharing will lead others to misinterpret or draw problematic conclusions from their data, potentially introducing new types of risk that they cannot control. Other may not fulfill the mandate without oversight due to confusion over how to do so. One barrier to data management and curation, particularly metadata generation to prepare data for reuse, is, "Nobody wants to do it because it is not for their benefit; it is for somebody else's benefit, and why would I care if somebody else can reuse my data?" A combination of education, incentives, and infrastructure are needed for data sharing to be successful.

Christine L. Borgman writes in *Big Data, Little Data, No Data: Scholarship in the Networked World* that scholars need a change in incentives to keep more data in a reusable form that would improve data access for others.[9] "They need tools, services, and assistance in archiving their own data in ways they can reuse them, which increases the likelihood that their data will be useful to others later," she writes. Borgman explains that this includes building and sustaining knowledge infrastructures collectively and investing in the human infrastructure that holds them together through invisible work.

Librarians provide some assistance, but their conception of data sharing varies from that of computer science researchers. Librarians tend to think of sharing in terms of open access, while researchers usually think of sharing only with fellow researchers, colleagues, or other institutions.[10] Though data may be shared in response to email requests or with trusted colleagues, researchers are hesitant to share their data widely due to doubts about its usefulness to others, concern about consent issues, and fears about misuse. When data is shared, it may not be packaged, cleaned, or documented. One researcher said his ideal scenario would be to have a line item in National Science Foundation (NSF) grant budgets to fund a program for researchers to hand over their completed data and IRB documents to be cleaned and posted publicly by NSF on a data.gov site. Most researchers see this packaging work as outside of their purview or skillset.

Data Storage and Security

When making decisions about data storage, researchers must take into account current security issues as well as unknown future possibilities for data breaches and re-identification. Finding the right repository or program involves many factors. One university offers a resource that matches project

9 Borgman, Christine L. *Big Data, Little Data, No Data: Scholarship in the Networked World* (Massachusetts: The MIT Press, 2015), 282.
10 Jillian C. Wallis, Elizabeth Rolando, and Christine L. Borgman. "If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology," *PLOS*, July 23, 2013. Accessed May 31, 2016. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0067332

characteristics with the appropriate storage program, such as Amazon Web Services or Dropbox. Choosing a program in which the university has a license may protect the researcher by placing liability on the university if something goes wrong.

One problem that arises frequently after libraries complete negotiations to acquire or license data for researchers is lack of a central repository for storage. Many schools are in the process of creating new repositories for big data, with libraries playing a primary role. Funding this work can happen via multiple routes, such as having storage costs written into grants. One librarian noted, "If they are going to manage their data, park it, and prepare it, somebody has to pay for the time for those researchers to do that."

Regardless of the storage location chosen, there is a widespread and prevalent concern over whether data is truly secure. One professor bought his own servers to store data rather than using university servers that can be accessed by Information Technology (IT) staff. Acting as his own system administrator takes several hours per week. Once the data is anonymized and aggregated, he stores it on the university supercomputers. Speaking about data security, he said, "I spend a lot of time worrying about that and thinking about it...something that keeps me up at night is I am sure my data is not as secure as it could be." He would love for the school to manage this for him, but feels as though his colleagues in computer science know far more about security than anyone on the IT staff.

Data access procedures and trainings vary depending on the data set, from public data with no clear guidelines to Census data kept in protected rooms that are disconnected from the network, and medical data servers protected through a firewall and encrypted email. Beyond the most sensitive and well protected data, there is ambiguity around what instructions or criteria a researcher should follow in deciding when to take more protective measures. As a result, students are known to keep anonymized data directly on their hard drives. IT and other departments offer guidance, but inconsistencies and confusion remain, as advice may not always be sought, followed, or clearly conveyed. One professor said, "Definitely if you want a really tight airlock, no air gap security, you can do that, but [if] it's the middle ground, how do you [protect the data]?" He said guidelines are difficult to find because security promises could lead to liability issues if the information is hacked.

Navigating Unexpected and Potential Harms

Regardless of whether a researcher has good intentions, future reuse or aggregations using their data can lead to unintended outcomes. One professor said a big anxiety in research is the possibility that, "People could take the methods I am working on to do things that I wouldn't want done." For example, in an oppressive regime, what if data on individuals that was collected to examine poverty could instead be used to infer ethnicity?

The anonymous release of data brings the threat of potential re-identification. For example, when researchers publicly released an anonymized data set based on Facebook profiles for about 1,700 college students in 2008, the data source was quickly discovered to be Harvard College, despite the use

of privacy protections and IRB approval. According to Michael Zimmer, "This incident "reveals considerable conceptual gaps in the understanding of the privacy implications of research in social networking spaces."[11]

The complexity of weighing risks versus benefits in the age of big data is well demonstrated by security research, which involves penetrating privacy protections to identify areas of weakness and proving that subjects can be re-identified without actually naming them. This can be difficult to explain to the IRB. For example, one computer science professor experienced resistance for using controversial data, which was publicly released, to examine the effectiveness of anonymization. "There were people in the scientific community that basically said, 'You shouldn't touch it. We know it's out there, but it's problematic in all these ways, and so it's not okay.' And I really wrestled, for a long time, with whether I should use that data in this research in this way." He determined that his work was justified since it was intended to benefit the public and the scientific community by demonstrating how to better encrypt data to prevent identification.

Sometimes data collection reveals potential harms that exist independently of a study, such as threats of suicide or self-harm expressed on social media or online forums. Researchers struggle over whether to contact law enforcement, counselors, or online platforms out of concern for individuals, or they may experience emotional impacts from viewing graphic content.[12] One professor asked whether the IRB should seek to protect researchers who handle data that may harm them.

Unintentional harms may also arise during the research process, regardless of whether the findings are ultimately published. Research using Amazon's Mechanical Turk,[13] a crowdsourcing online marketplace, is not reviewed by the IRB, but researchers worry about the ethics of using work that pays very low wages. One professor, following an incident in which a reporter googled a tweet from her study and publicly re-identified a participant prior to publication, has developed strategies for protecting anonymity: she withholds or masks user IDs or sequences them to random numbers; strips metadata and only shares content; and tweaks quotes over a certain number of words for papers.

Future Unknowns

One professor explained why new safeguards are needed, saying, "I think it is not just that more data is available and it is harder to protect people's privacy...It is that we [are] actually much better able now and want to, as a result, manipulate people as experimental subjects, and that creates a lot of issues that we haven't had to deal with that much in the past." As another person stated, "We don't have any

11 Zimmer, Michael, "'But the data is already public': on the ethics of research in Facebook," *Ethics Inf Technol* (2010) 12:313–325, June 4, 2010.
12 Stern, Susannah R., "Encountering Distressing Information in Online Research: A Consideration of Legal and Ethical Responsibilities," *New Media & Society*, June 2003, vol. 5 no. 2: 249-266. DOI: 10.1177/1461444803005002006
13 Hitlin, Paul, "Research in the Crowdsourcing Age, a Case Study," Pew Research Center, July 11, 2016. Accessed July13, 2016. http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/

systematic training or protections for what I would say are the newer, emerging problems." The rise of big data and manipulation also raises the question of when modeling or statistical consulting should warrant co-authorship.

Another professor summarized the issue by saying, "Part of it is just making them aware of what the well-known canonical questions are. The thing that worries me more is that as we develop new methods and new techniques and new types of research, new potentials for harm, new ethical questions are coming up for which we don't have many examples yet and people are not aware - the unknown 'gotchas' that I worry about people falling into."

An Ecosystem of Support

Computer science researchers have little to no formal or systematic ethics training in their education, compared with medicine or psychology. They often use informal networks or conversations to make ethical decisions in their work, or learn from their advisors in an apprentice relationship as they encounter issues for the first time. Requirements such as IRB training or Responsible Conduct of Research (RCR) provide some basics, but most computer science researchers feel as though this is inadequate and irrelevant to the challenges they face. However, researchers often learn ethics on the job, through good and bad experiences, and from ad hoc conversations with other graduate students or peers. One interviewee said they learned through self-investigation and self-study, and that, "It came out of a need to know."

A variety of formal and informal structures and services help to fill these gaps on campus, addressing issues such as data management, collection, and sharing. However, these mechanisms may only be conveyed through word of mouth, not universally used, or made visible only following a violation. Although researchers receive announcements, they may not use available resources unless they know someone who has used them, or they may have incomplete information about the consortia of options available or how they fit together when a particular need arises. Many researchers complain that these services are ill-equipped to address the unique concerns that arise from big data research.

Formal Oversight Mechanisms

Oversight mechanisms exist to help support researchers, protect universities, and get grantees to consider the long-term existence of their data. These mechanisms all play a part in overseeing big data research needs, but more support is needed with the growingly complex world researchers often find themselves in.

Data Sharing Mandates and Data Management Plans

Federal funders require researchers to complete Data Management Plans (DMPs) at the start of a project to describe data sharing and storage plans. Requirements vary by funder.[14] While DMPs incorporate ethical issues and are part of a formal process, researchers see them as a hoop to jump through and tend to informally swap and copy sections from their peers' DMPs, then adapt and reuse them. DMPs should help researchers think about how people will access their data when their research is done, but they are completed ad hoc and without formal training. One professor said, "It certainly was not something that I put a lot of thought into." This sometimes results in confusion over responsibility. For example, data management and sharing in multi-institution partnerships can be complicated when researchers feel little attachment to the data or believe data management is the responsibility of a principal investigator (PI) at another institution.

Assistance is available to help researchers understand the components and meaning of DMPs, especially in the early stages of their careers. Some universities provide templates or writing assistance for grant submissions and initial phases of research-funder relationships. One university provides pre-filled out templates and materials at specific deadlines for each grant, along with suggesting changes to help an application obtain funding. Libraries assist a handful to a few dozen researchers seeking help each year, while others turn to online templates such as DMPTool. One library plans to review all DMPs in order to have greater oversight at the beginning of a research project, giving librarians a chance to raise red flags about grant proposals that could prevent problems later. That change seeks to address a common scenario in which researchers come to the library only at the end of their project for needs like repositories and archiving.

Christine L. Borgman writes, "However, it is not clear how much of the data management burden scholars are willing to bear. Many, if not most, view time and resources devoted to managing their data as effort lost on their research. They may prefer to delegate these duties to library or archival staff, although such partnerships also take time to develop. Libraries are stretched to provide current services, and not all view data management as being within their purview. Publishers are more willing to index data and link to repositories than to curate data."[15]

Borgman continued, "When present-day researchers are asked whether they are willing to share their data, most say yes, they are willing to do so. When the same researchers are asked if they do release their data, they typically acknowledge that they have not done so. Willingness does not equal action…"[16]

14 "Dissemination and Sharing of Research Results," the National Science Foundation. Accessed May 31, 2016. http://www.nsf.gov/bfa/dias/policy/dmp.jsp
15 Borgman, Big Data, Little Data, No Data, 284.
16 Borgman, Big Data, Little Data, No Data, 205.

Institutional Review Boards

Institutional Review Boards (IRBs) are the legal oversight mechanism for human subjects protections at universities and federally funded institutions that determine research project approvals, denials, and modifications. While researchers may learn ethical principles through the restraints of the IRB and value its legal and procedural oversight, many researchers say the IRB is not the best mechanism for considering potential ramifications of big data ethics overall, as human subjects protections are just one component of ethics. One person said, "The act of articulating things to an IRB helps you think it through for yourself, but in the end I think you are in dangerous territory if you proxy the ethical decision to some other."

According to the February 2016 webinar, "Big Data Research: Practical Solutions to Emerging Problems for IRBs," by PRIM&R (Public Responsibility in Medicine and Research), IRBs struggle with questions such as whether de-identified data qualifies as human subjects data, when a review should be expedited to protect human subjects, and how to de-identify data while still retaining its research value. With regards to consent, the IRB must weigh: what kind of consent is required for retrospective (or secondary use) data, in what cases is achieving consent impractical due to volume, and when can consent be waived due to minimal risk? Threat level must be assessed in comparison to everyday risk. Factors that increase the threat level and consequent data restrictions include HIV status, mental health diagnoses, financial or banking information, and, potentially, education level. Combined data sets, including historical data sets that were collected with different standards than are currently used, are particularly challenging. Security laws vary by state, and researchers must follow the most stringent applicable state law for an entire study of multi-state aggregate data, as explained in the PRIM&R webinar by presenters Betsy Draper and Sean Owen.[17]

IRBs are criticized as lenient, understaffed, bureaucratic, and slow, which can tempt researchers to cut corners. One computer science student said most computer scientists do not know the IRB exists since they assume they are not working with people. IRBs are sometimes modeled on the traditional - medical and psychology - model rather than computer science. Critics said: it is probably easy to sneak something past the IRB due to their lack of understanding; the IRB over-emphasizes privacy and disclosure and overlooks other issues; the IRB has an adversarial relationship with researchers in terms of what it will allow; it focuses more on alignment with a grant or funding source than ethics; and its primary purpose is to provide cover and liability. One professor said, "They don't understand big data sets and unobtrusive data collection, what it means to get data from outside...There's a lot of lawyers and a lot of people who don't really deal with this kind of data or have thought about it in any deep way."

17 Draper, Betsy and Owen, Sean, "Big Data Research: Practical Solutions to Emerging Problems for IRBs" (webinar), PRIM&R (Public Responsibility in Medicine and Research), February 10, 2016.

Some researchers actively avoid disclosing ethical issues in their work to the IRB. One professor noted, "When I write my IRB application, I know what the real ethical issues are and when the comments come back, I am always really surprised that they are about something completely different and mundane and not about the ethical issues...maybe what I should be doing is actually pointing out the ethical issues myself in the application. But who does that?"

The IRB is also limited in the types of research it covers. Secondary data use is generally considered exempt by IRBs and not part of traditional review, but changes to research methods resulting from big data have drawn this into question, as they have blurred the distinction between primary and secondary research.[18] Researchers called the IRB too permissive on secondary data use and public data sharing in particular, even as they benefited from their leniency.

One professor said the idea of deferring ethical responsibility to the IRB is a "fundamentally broken paradigm," as this makes the relationship with the IRB transactional. Instead, the IRB should be advisory and the underlying ethical responsibility should belong to the researcher, since they will be in the spotlight if a violation occurs. Many ethical issues fall outside the IRB's responsibility of protecting human subjects, such as corporate intellectual property. Educating board members on best practices in security and big data could help when these issues arise.

There are concerns about what is being left out of formal IRB and RCR trainings. One person said, "We don't have systematic training or protections for what are emerging issues." With a large number of international graduate students, there are concerns about whether ethics trainings and conversations are culturally sensitive, realistic, and understandable. Existing ethics trainings are seen as too focused on compliance and lacking engagement and sufficient real-world examples for students.

Associations and Conferences

Journal review boards and conference program committees are one potential structure for providing additional oversight by determining whether submitted work is ethical. However, opinions vary on whether they should have that role, and what should be required to make their process fair and consistent. Some reviewers say it is not their place to criticize the ethics of something approved by an IRB, but others disagree. Professional associations often lack review policies, leaving the protocol up to individual reviewers. Ethical issues are frequently absent from association and conference agendas. Conferences and publications vary on whether they allow presentations that violate TOS. Some claim to disallow TOS violations but end up publishing research with violations unknowingly. Program committee reviewers have inconsistent approaches and often rely on trusting the researcher.

18 Metcalf, Jacob, "Big Data Analytics and Revision of the Common Rule," *Communications of the ACM*, Vol. 59 No. 7: 31-33. 10.1145/2935882

Challenges of the Ethics Frame

Convincing researchers to take extra precautions that slow down their research when the threat of a bad outcome is minimal is a hard sell, one professor said. He explained, “I think there are definitely people, if it’s low overhead, people want to do the right thing. It’s just they don’t even know exactly what the right thing is and trying to figure that out is time consuming.” Another professor noted that going to a discussion forum with an ethical concern that leads to a constraint being placed on her research would not make her want to return, but receiving support materials would be helpful. Similarly, one professor noted that most people do good faith research and do not commit terrible violations, but they may have a question about best practices for a particular task. How can the research community create a space for those conversations?

Suggesting a conversation about ethics often makes people feel like they are doing something wrong, as the most memorable ethical topics are extreme breaches. Yet ethics issues are often discussed as part of routine decision-making between peers, or with advisors, without necessarily using the word “ethics.” Finding mechanisms to introduce the topic more explicitly is complicated, as the word “ethics” may have a deterrent effect if the conversation is optional, or it may be perceived as a bureaucratic hurdle or checklist when introduced as a requirement. For example, one professor explained, “The ethics training we do is somewhat helpful, but it’s also just this formality that people don't take seriously, that isn't updated enough, that isn't nuanced enough." Meanwhile, one of the challenges of voluntary ethics workshops are that they attract people are who already thinking about ethics, while “the people who need help are the people who are not asking for it.” Another professor said, "The people who attend ethics workshops are not the people who need to attend ethics workshops, I think, by definition.” Emerging data clinics are one promising way to strike the balance between these imperatives and embed ethics in technical practice.

Some researchers may not feel that ethics trainings or deliberations are even relevant to their work. For example, one professor said that Machine Learning researchers view themselves as being "in the plumbing," similar to the way that FedEx is not concerned with libel laws when delivering a package containing libelous print.

Librarians as well as computer science researchers do not necessarily think of their work as ethical or unethical. Librarians are often engaged in work that involves ethical decision-making, such as long-term preservation, legal or contractual use of resources, and copyright issues, without necessarily using the word “ethics” to describe it. Though IRBs are more likely to be credited with overseeing ethics, librarians and others who support researchers along the way in their projects may be the ones helping them to make their nuanced ethical decisions.

Rather than a set of ethical rules, one professor said she would like to have support tools that explain the options and constraints - such as legal requirements, ownership, cost, and ethics - for making contextually driven decisions for particular project goals. She said, “I think [that] is the way we need to go, because the normative approach is problematic, because we haven't dealt with some of these

problems before, and we're making the norms ourselves through our practice." Indeed, many of the researchers we interviewed who thought deeply about ethics said it would be better to use notions like trade-offs or focus specifically on privacy, re-identification, or security concerns.

Informal and Educational Support

Campus support services are another piece of the puzzle. Libraries have expanded their services and new data drop-in clinics are appearing. A key is engaging computer science students and researchers in formal and informal education as early as possible and ensuring ethics training is understood in the context of their work.

Research Data Management Services at Libraries

Libraries have had a growing role in providing Research Data Management (RDM) services at various phases of the research lifecycle through dedicated units or ad hoc assistance at many institutions. Librarians provide advice on choices regarding metadata and format for deposit and reuse, and help with negotiating access, data archiving, DMPs, data storage, citation management, intellectual property, preservation, and future access. However, researchers have varying levels of awareness that these services are available. Acting as an intermediary to refer researchers to other units across campus is also a function of many research libraries.

The "Research Data Management Services" SPEC Kit by the Association of Research Libraries (ARL) documented the prevalence of RDM services in academic libraries based on a 2010 survey with 73 institutional respondents. The survey showed that 74 percent of responding libraries provided RDM services, and 23 percent planned to do so. The DMP mandate from NSF, which took effect in 2010, inspired some of those libraries to introduce the services, which included DMP assistance. Three out of four respondents provided links to the popular DMPTool. Librarians providing RDM services had a range of educational backgrounds. When asked which skills RDM services staff needed advanced training on, 29 percent of respondents said ethical and legal issues. They also noted deeper technical skills in data acquisition, analysis, interpretation, and visualization.[19]

Campus partners for dedicated RDM services units at the library include IT, research offices, supercomputer centers, digital and subject librarians, and representatives from the IRB, research technology, and survey research. Services and activities include trainings and workshops, metadata, curation, data management consultation either by appointment or drop-in arrangements, forums on open data and reproducibility, outreach about data management funding requirements, and

19 "SPEC Kit 334. Research Data Management Services," Association of Research Libraries, July 2013.

infrastructure and repository development. Staff and assistants may include data curators, research data management librarians, or library liaisons to scientific disciplines who have data experience.

Discussions with researchers often revolve around general data management issues such as available options for storage or working in the cloud, or relate to the overwhelming nature of organizational research needs, such as sharing information in off-campus collaborations. The library's RDM services have a greater appeal to researchers from departments with fewer data-related resources or smaller grants, typically excluding computer science departments. One graduate student stumbled upon the library as a resource when his work accidentally shut down access to a publication. The library contacted him and helped him to gain access, which led him to return to the library for that service again.

However, many computer science researchers view data management as a task belonging to the individual researcher or their unit rather than looking to the library. Others contact the library only as a last step for curation and storage, which puts libraries at a disadvantage for having meaningful input or ensuring prior steps in the research process are in sync with what will be needed at the end. One librarian said, "Depending on who you talk to, there are the people that think, in the library side and in the research side, that the stuff that we're calling RDM is the stuff that happens after you're done with your research and a place to put your data so that others can use it. Our consulting service has a more holistic view of what RDM means. We're talking about data management throughout the data lifecycle, ideally beginning when people are writing their grants so they have a good sense of what resources are available for them and how they'll be doing things in an efficient way from beginning to end, with the ultimate goal being long-term preservation of the data unless there's some reason why they can't do that."

Data Clinics

Data clinics based on the statistical clinic model have emerged at some universities, providing drop-in centers to discuss issues such as collecting, cleaning, and manipulating data, which often surface ethical issues. Additionally, some statistics clinics have started doing data clinic work. Clinics offer workshops on topics like data carpentry, R (a programming language), web scraping, Hadoop, and Amazon Web Services. Staff members may include professional statisticians and people with statistics degrees, Masters and PhD level professional consulting staff, and students who typically do the lower level work.

Data clinics can also offer an opportunity to discuss ethical challenges without the sense of judgment experienced in an IRB review. They can provide a more formal structure and consistent student access to conversations that often happen informally when researchers seek out colleagues or statisticians one-on-one. Referring to office hours services that would be useful, one person said, "Those are the services that you need anyway to do your research and you have to get that work done anyway to publish. And if somehow the conversation around ethics could happen around those things that you need to do anyway and around those relationships that you need to build anyway, that's one way of trying to ensure that it actually happens." However, attracting people to focus on ethical issues might take work.

People are likely to come to a statistics clinic with questions in which the answers are needed to finish a dissertation, publish papers, or complete graduation requirements. Issues of ethical risk, on the other hand, may be discounted as ambiguous, far into the future, or unlikely to lead to being caught or facing severe consequences.

Funding models for statistics and data clinics vary. Some clinics offer limited consulting for free and charge for additional sessions. Others charge departments individually, though large departments with their own internal resources may choose not to pay for their students to utilize the service, leaving the students to pay per visit. Researchers may write the assistance as a line item into their grants. In general, computer scientists and others with significant resources in their departments, or strong confidence in their knowledge of the field, are less likely to seek out voluntary resources and are thus harder to reach through these mechanisms.

One professor said future data clinics or office hours arrangements should address functions such as, "Connect with them about how to put their policy together [and] how they would help us post the data and collect it, or where it would be stored. They would probably have this automated system to help anonymize parts of it really clearly." The touch points could be IRB approval, data hosting, data storage, data management, access to software and software licenses, and data analysis and sharing. The scope should include both quantitative and qualitative issues and instruction on campus policy regarding licenses for various tools. Data office hours would be helpful for discussing security, privacy, archival method, and who has access to services like Dropbox or Google Docs.

Teaching ethics

Ethics is sometimes taught through dedicated courses on data or research ethics, or embedded within a range of existing courses. Professors spoke of the importance of teaching ethical research principles such as transparency about research methods and reporting, conducting research that contributes to the public good, and balancing inquisitiveness with risk to individuals. Their approaches include:

- Present moral dilemmas with no clear answer and open the topics up for debate. Seek students' perspectives. "I ask the students ethical questions I don't know the answer to."
- Explore the application of philosophical theories to ethical dilemmas.
- Discuss consent and privacy of data collected by companies in a way that challenges students to think about the moral and conceptual ambiguity around it, rather than relying on a "pre-ordained" approach common to computer science ethics training.
- Make people aware of the information that can be learned from the data they collect, including in combination with social media data. Have students look at self-reported medical information in social media groups without strict privacy settings to teach about re-identification, exploring topics such as linkability of information and quasi-identifiers.

- In prototyping technology, have students think about alternative uses and commentary outside the typical research agenda; identify core metaphors in the field and think about communities who are excluded or marginalized by them; then build alternatives. Class can be an exploration of "alternative worldviews and philosophical systems and value systems and what they have to say about technology, and what the implications might be for design, for engineering, and for entrepreneurship."
- Have students design a project showing the "darker sides of technology," which is intentionally wasteful or promotes deception, to show potential misuses of their outputs and what can happen when providing quick deliverables. Then have the students discuss how "some of those things might also appear in designs that are meant to be very ethical, that in fact if you're not aware of what that other landscape looks like, you might not be aware of how close you are to operating near it as a designer or as a computer scientist."
- "I just try to embed [ethics] in every class I teach. It will look different in different classes. In the intro web development class, it will look like accessibility, it will look like working properly with clients. It will work with the fact that designers make decisions. In the computational methods class, it will talk about biases and limitations, protecting people's rights."

Additionally, professors discuss values such as providing fair treatment to research subjects, especially disadvantaged or at-risk groups, and doing participatory action research that makes the researcher accountable to what participants who lack an academic background hope to gain from the experience. Did the research subjects benefit from the research findings? Will they become beneficiaries of the output or product if the project is successful? Likewise, social media research should have the long-range goal of improving the user experience.

In one person's work with data that is generated by people, ethics means: "ensuring that when we utilize that data for answering a research question, we ensure that the data is not compromised, the data is secure, and the data is presented in a way that doesn't show the person in an inappropriate light or jeopardize their identity or create avenues of discriminating against those individuals in a certain way in the broader society."

Professor Katie Shilton of the University of Maryland's College of Information Studies has introduced the concept of "values levers," or "practices that open new conversations about social values, and encourage consensus around those values as design criteria" in her study of the social and ethical

implications of emerging technologies.[20] She uses this term to describe ethical issues that surface within design processes, such as internal technology or software testing, seeking user feedback, or navigating IRB mandates. Shilton explains that backbone technical infrastructure presents ethical ambiguities, as it is designed to be flexible for various things to be built on top of it. This creates unique challenges for designing values in infrastructure.

There is widespread interest in increased guidance and resources, such as best practices, for data ethics. Upfront training before problems appear could relieve professors from the repetition of teaching students these lessons one-on-one and could expand students' knowledge beyond that of their professors or advisors. Researchers also suggested providing resources around political, technical, ethical, and legal issues of web scraping and crawling.

Ethical mistakes or violations are often due to lack of knowledge rather than contentious behavior, making education key. But some ambiguities are difficult to resolve. One person asked for a concrete division of responsibility, "A conclusion of who's ultimately responsible and how incentives are managed in terms of the mix of corporate ethics, academic ethics, the journals, the grant agencies," and others. The person said that while stakeholders each have their own corner, it's not ultimately clear "who has responsibility and if those boundaries can be defined in a clearer way."

Some interviewees cautioned against teaching data ethics too early, saying students will forget the lessons by the time they encounter the issues. Another person suggested that ethics needs to have a constant presence, saying, "For me, the idea of ethics is a very relative one, relative to a specific situation and cultural, social context. So you have to keep learning it over and over again, at least the specifics of it."

A report by the multi-disciplinary Council for Big Data, Ethics, and Society, hosted by Data & Society Research Institute, found that data ethics courses and modules are sprouting up across a diverse set of disciplines such as mathematics, data science and analytics, statistics, computer science, astronomy, journalism, law, biomedicine, and media studies.[21] Topics of study include privacy, professional ethics codes, cybersecurity, plagiarism and fraud, objectivity and bias, data aggregation and representation, data mining, and predictive analytics. The Council concluded that ethics education should include interactive discussions and nuanced case studies that go beyond the RCR course model and other training methods that are centered on rule-following. The Council advocates for integrative approaches that include ethics as an element in design or problem solving processes within each course rather than stand-alone ethics modules that place ethics reasoning outside of research and engineering practices.

---

20 "Katie Shilton: Finding Values Levers: Building Ethics into Emerging Technologies," University of Maryland, April 16, 2013. Accessed May 5, 2016. http://mith.umd.edu/dialogues/katie-shilton-finding-values-levers-building-ethics-into-emerging-technologies/

21 Jacob Metcalf, Kate Crawford, and Emily F. Keller. "Pedagogical Approaches to Data Ethics." Council for Big Data, Ethics, and Society report, April 21, 2015. Accessed May 5, 2016. http://bdes.datasociety.net/council-output/pedagogical-approaches-to-data-ethics-2/

Role of the Library

When helping researchers with their data management needs, librarians do not always have the technical skill set or vocabulary to be one of the main sources of support to these researchers at this point. Increasing data science education for librarians and hiring scholars to act as liaisons are helping to strengthen this role.

The library's growing role in RDM services has generated concerns among some researchers and librarians based on librarians' frequent lack of data research experience or subject expertise. Many researchers do not think of the library when they think of their data-related needs, relying instead on resources within their departments or seeking help informally from colleagues.

An empirical investigation into the research data services at U.S. and Canadian academic research libraries found that 28 percent of 223 respondents considered those services to be "integral" to their job, while 41 percent interacted with faculty, students, and staff on research data services issues only occasionally, and 32 percent did not view those services as a regular part of their job. Of those who chose the latter, 60 percent said they did not feel they have the skill, knowledge, and training needed to provide research data services.[22]

Gaining greater technical skills would boost librarians' credibility and trust with researchers. Areas of training could include learning more about big data research design to understand the principles and assumptions that go into data collection, as well as data science, data visualization, and statistics. This would strengthen librarians' abilities to ask questions to assist researchers or evaluate a data set. Some librarians are pursuing these skills through workshops or online courses. One librarian said she would be uncomfortable with a large data set without this additional training. To help bridge the gap, some libraries hire data scientists to act as liaisons between librarians and researchers, or host postdoctoral fellows from the Council on Library and Information Resources (CLIR), which places PhDs in humanities, sciences, and social sciences in academic libraries to promote interdisciplinary collaboration.[23].

While some librarians are happily jumping in to maximize their role in RDM services and raise their profile around campus, barriers and resistance remain for others. One librarian explained, "Different groups on campus have varying levels of receptivity to libraries being involved in research generally." Some researchers seek significant assistance while others have no interest. Librarians also have varying opinions on how highly data science should be prioritized among their tasks. One librarian said, "Subject and reference librarians have always been involved logically in helping faculty and students

22 Carol Tenopir, Robert J. Sandusky, Suzie Allard, and Ben Birch. "Academic librarians and research data services: preparation and attitudes," *International Federation of Library Associations and Institutions* 39(1) 70–78, 2012.

23 "Fellowships in Academic Libraries," Council on Library and Information Resources. Accessed April 21, 2016. http://www.clir.org/fellowships/postdoc/applicants/acad.

with their research, but this recent influx of digital scholarship onto our campuses has been a big step for a lot of librarians." There is a funding challenge for libraries to provide new services while continuing to provide existing ones, such as adhering to traditional methodologies for preservation and access.

Some additional barriers to librarians getting involved with data issues are lack of familiarity, fear of stepping on people's toes, and concern about data scandals. One librarian explained how data management can generate fear and resistance among librarians by saying, "It sounds like one of those situations where you make one wrong move and you expose someone's data. Every time you hear 'data' in the news, it is a frightening thing."

Advising researchers on managing data is complicated by the fact that each person has a unique process, which one librarian compared to the way that people manage their vacation photographs, asking, "so how do you interject yourself in that, and how do you give them really explicit advice?" The time commitment deters researchers from data management, and librarians can help to relieve some of that burden, taking the approach that, "They're supposed to make the data, not worry about what it's like in twenty years."

Future of Libraries' Support for Researchers

As RDM services increase, one person said the library does not want to become a compliance monitor, as that is the role of the IRB, but could provide someone to talk to for issues of uncertainty or understanding options for sharing data while protecting confidentiality, as well as nuances between sharing all or none of their data. She said, "I think that the librarians have a really good ability to look at the big picture and say, 'Okay, now have we addressed all these issues?' Someone could be thinking about storage and file naming and even thinking about how they're going to disseminate their results, but the librarian, when they talk to the librarian, could say, 'Have you thought about confidentiality? Have you thought about privacy?' and bringing that holistic view of the work and that being one piece of it into people's minds."

To increase the publicity of its services, one library wants to present the assistance it offers as concrete and benefiting researchers in the lab by saving them time, effort, or investment and making them better, stronger, and faster.

One professor said, "Thinking totally outside of the box, if I trusted them enough, it would be fantastic to give something like a library the raw data I get from the [data source] while I am in my fanciful dream world and they would even negotiate the NDAs and figure out the back end networking that's required to get the data from the [data source]; sanitize it; anonymize it according to best practices; and then provide it in a nice, structured database for me to use. I honestly cannot see that happening in the next 20 years."

For the library representatives we interviewed, future library goals include creating a network of

services to go beyond the library, with centralized communication, relationship building, and an integrated approach across silos; and being a connector or catalyst to link researchers to each other and relevant departments. One library is working on a matchmaking service to recommended repositories for specific data sets and people to connect to on campus or beyond, and to provide a persistent index of data sets to extend the value of negotiations for access beyond single uses. A librarian described this challenge by saying, "A lot of the time these access to data sets are negotiated by departments or individuals, and no one knows who's got what or what you're allowed to do there." Libraries can do this simply by providing a list of websites that they have contracts with and that the researchers have privileges to use.

Librarians know that to gain researchers' participation in their programs, they must design services that match researchers' goals and seek their feedback in the design of services. Telling researchers to share data needs to be backed up by providing the right resources for depositing and training. Libraries need to better understand reuse from the user perspective and take that into account in repository design. One librarian said, "I actually don't think we've done a very good job as a profession understanding how people use other people's data and how that influences what we do with our repository. I think we've done a much better job figuring out what depositors want, what they need, and not what end users want and need."

Data sharing that goes beyond fulfilling the federal mandate and makes it usable by someone else requires more than just handing over data to a librarian. It requires supporting the PI to learn about the other kinds of information that are necessary. One researcher said, "The librarians are doing a terrific job of trying to push people a little bit to supply enough metadata about the data sets they are archiving so that it is actually useable by someone else."

For the library to add new methodologies for preservation and access while maintaining the traditional ones is a huge task. One librarian said, "We have 200 years' worth of methodologies for preserving and accessing information that we must adhere to. We cannot get rid of any of what our past practice has been because all that stuff still exists, right? We have one foot steeped in these traditional methods that are cumbersome, but those things just have not been wiped away because we created something called the Internet. We just keep adding."

Recommendations

1. Investigate the ethics frame and consider how to engage researchers on ethics.

Given the resistance we heard to the word "ethics" from researchers who were deeply invested in this issue, it would be fruitful to better understand how that frame is understood in computer science discourse more broadly. Because ethics is often seen as outsider language, it is important to understand emic frames that may be more productive in engendering the ethical thinking that is desirable, both from outsiders and from most computer scientists. If there is a strong consensus that computer scientists need to understand "ethics" using that language, research would be needed to determine how

to bridge the gap between how computer scientists are thinking about these issues and how those outside the field are.

2. Embed ethical guidelines and practices as early as possible and in courses and projects.

Taking into account that ethical violations frequently result from lack of training or coordination rather than malicious intent, a holistic approach to teaching data ethics is needed. This would provide nuanced discussions to build on current trainings such as RCR and the standard IRB training. While the existing trainings are important for helping students to understand the background of why they have IRBs and their impact on the field, these experiences are not sufficient for preparing students for the range of challenging ethical situations they will encounter in academia and in current and future professional settings. Ethical discussions should take into account varying cultural beliefs and sensitivities and the role of information fluency in strengthening students to make responsible decisions.

It is also important to better understand what is most pedagogically viable - required ethics courses, ethics modules in existing classes, or perhaps actively designed course assignments and data sets that require ethical thinking as part of the technical work. Further research is needed to understand what would work best, and more work is necessary to develop curricula that could be broadly adopted.

For further research, surveys of students, professors, and alumni could examine the long-term influence of ethics trainings and courses on future decision-making and could help to identify best practices in teaching ethics.

3. Data Clinics provide a potentially effective model for encouraging technical researchers to discuss ethics informally.

Drop-in data clinics, inspired by statistics clinics, provide a unique opportunity at a handful of universities for students and researchers to discuss their data work in a dedicated space or via rotating office hours in different departments, such as providing data assistance to sociologists as data comes up in their work. They are described as supporting a "grab-bag" of issues such as data acquisition and cleaning, or manipulation of large-scale data sets. Right now, there are very few such programs and the incentive structures are not necessarily designed to enable hard ethical questions or trade-offs to emerge.

Providing a structure for ethics issues to be addressed alongside technical quandaries in the process of active research allows for complex issues to unfold within one-on-one or group discussions. Additionally, through the group workshops convened for this project, stakeholders showed enthusiasm for having a connector or consultant play an interlocutor role to address ethics issues by looking across formal and informal structures on campus. This led two universities to contact Data & Society about replicating the study. Data clinics play a role that is similar to these workshops. Further study of the discussions at data clinics could provide insight not only into the intersection of common ethical and

technical questions, but the issues that are not adequately addressed by existing structures, courses, and trainings. The unstructured nature of drop-in clinics could provide an opportunity to uncover unmet needs within a researcher's work to aid in university planning.

4. Clearer guidelines and better coordination are needed from funders, the academic communities, discipline-specific organizations, and/or institutional structures like IRB regarding who ultimately bears responsibility for data impacts.

Centralized knowledge structures could assist researchers in navigating the disparate resources on campus, helping to guide them to particular people or places as ethics questions come up throughout the research lifecycle. Researchers are seeking stronger guidance regarding standards for public data collection and publishing, as well as security precautions and procedures for using third party data storage providers such as Dropbox or Amazon Web Services, which would be helpful across universities. This may be a place for the research library as a central hub. Librarians have a key set of values and skills that could benefit these researchers, though additional training in data science may be necessary for some to gain comfort with helping researchers navigate grey areas.

In addition, educating the IRB on emerging issues in data science, or appointing data scientists and experts to the IRB could open up new opportunities for substantive discussions or procedural changes regarding the emerging, potential long-term harms of big data research. Furthermore, restructuring the IRB to enable non-confrontational dialogue between technical researchers and ethical overseers could go a long way to help researchers working in new arenas to reflect on ethics concerns.

5. With additional data science literacy, librarians can better support the growing data-centric work of researchers on their campuses.

Researchers need campus partners who understand the new and complex ethical issues they are navigating in their work. Librarians can assist in this new data-driven environment by experiencing the research data lifecycle themselves, becoming data savvy, embracing the data science culture, and considering the new roles for information professionals in the current environment. Libraries work with students and researchers and teach information literacy classes. Fluency in data literacy and understanding the potential impacts of technical research will benefit the campus. This includes being seen as the source of information fluency on their campus and data literacy is part of that role.

6. Expansion of current services provided by libraries could support some immediate needs.

From what we saw in our project, there are straightforward ways for libraries to increase their support for researchers. Legal use of information is sometimes complex to navigate but libraries have been providing copyright, Internet Protocol (IP), and Creative Commons assistance and resources on campuses for some time. There could be a role for libraries to be resources for navigating murkier areas such as data ownership, TOS violations, and web scraping concerns.

As needs for safe, secure, and long-lasting research repositories increase, libraries have an opportunity, which some have already taken, to host robust data repositories by partnering on campus or with a consortium of organizations to create data repositories, especially for potentially sensitive data. As catalogers of knowledge, libraries need to be creating and thinking through metadata to ensure the security and privacy of sensitive data sets and proper cataloging for future use. This metadata can help ensure any sensitive data is wrapped with the right descriptive information for future sharing.

When libraries advocate for open access, open science, and open data, they must take the next step and help support what it means to make data open and sharable – having the difficult conversations about ensuring privacy, confidentiality, and potential unintended future uses of data. If services are to increase in a meaningful way, additional staff or funding may be necessary.

Conclusion

Navigating ethical issues in big data research presents complex challenges to researchers in computer science and beyond, creating the need for an ecosystem of support that spans data management, education, training, and oversight. Promoting the discussion of ethics is challenging but the future will be even more confusing, as developments in data science raise possibilities for human harms that transcend traditional ethics regulations. It is important to equip students and faculty with the resources, policies, and support networks necessary to navigate an unknown future ethically. Overall, we found that all of the schools we talked to are actively seeking and struggling with how best to approach this, and the collective development of curricula, problem sets, relevant data sets, and best practices for privacy and security in data storage would be helpful.

Big data research creates ambiguities and questions surrounding issues of consent, adherence to Terms of Service, web scraping, and secondary use of publicly available data. Researchers receive assistance from various sources, including peers, advisors, data clinics, and the IRB. However, gaps remain and many potential ethical problems are not thoroughly addressed. In being a profession that is concerned about privacy, intellectual freedom, and the public good, libraries have a role to play as we all figure out new norms for how to handle data being collected about us, how we think about future uses of it, and where we go from here. Bridging these two worlds requires a deep commitment by universities, librarians, and faculty, which could significantly benefit everyone if designed well.

Acknowledgements

This research was made possible by a grant from the Alfred P. Sloan Foundation, with additional support from the National Science Foundation (IIS-1413864). Dr. Rachelle Hollander, Director for the Center for Engineering, Ethics, and Society at the National Academy of Engineering, served as a project advisor. We also received additional advice and intellectual support from the Council on Big Data, Ethics, and Society.

Data & Society is a research institute in New York City that is focused on social, cultural, and ethical issues arising from data-centric technological development. To provide frameworks that can help address emergent tensions, D&S is committed to identifying issues at the intersection of technology and society, providing research that can ground public debates, and building a network of researchers and practitioners that can offer insight and direction. To advance public understanding of the issues, D&S brings together diverse constituencies, hosts events, does directed research, creates policy frameworks, and builds demonstration projects that grapple with the challenges and opportunities of a data-saturated world.

Contact

Bonnie Tijerina
bonnie@datasociety.net
Data & Society Research Institute
36 West 20th Street, 11th Floor New York, NY 10011
Tel. 646-832-2038
datasociety.net