Research Ethics in an Age of Big Data

Chris Alen Sula

EDITOR'S SUMMARY

The era of big data introduces new considerations into the traditional context of research ethics. Ethical questions may be considered in terms of accuracy, humane treatment, informed participants and the necessity and applicability of the work, but big data complicates these issues. Since social media participants reflect certain demographic features, data drawn from those sources should not be taken to represent the general population. Big data collection may be more invasive than necessary due to easy access, and consent may be nonexistent. Data that was once anonymous may become identifiable, last indefinitely and conflict with goals for publication. Ways to respect ethics in big data research include involving participants throughout the process, avoiding collecting information that should remain private, notifying participants of their inclusion and providing them options to correct or delete personal information, and using public channels to disseminate research.

KEYWORDS

big data
ethics
research methods
data collection
personal information
anonymity
privacy
information dissemination

Chris Alen Sula is assistant professor and digital humanities coordinator at Pratt Institute's School of Information. He holds a Ph.D. in philosophy from the Graduate Center of the City University of New York. His research applies visualization and network science to humanities datasets, especially those chronicling the history of philosophy. He has also published articles on the politics of technology and ethical and activist uses of visualization. He can be reached at csula <at>pratt.edu.

The ethics of big data has been making headlines in recent years – and on several fronts. Government surveillance has been a steady topic of discussion since the 2013 revelations made by former CIA contractor Edward Snowden. Stories of cybercrime and cyberterrorism are increasingly front page news. And popular and academic venues alike considered ethical questions raised by a 2014 Facebook study of emotional contagion [1, 2, 3]. That experiment manipulated 689,003 users' news feeds to study how their moods changed when they were presented with positive or negative posts. Users selected for the study were unaware, explicitly, that their feed was being altered, though Facebook added research to its data use policy four months before the study was published [4].

These events give pause for academics and researchers to reflect on the ethical, legal, social and political implications of big data. Much good work has been written in this area, and much more remains – to say little about its prospects for implementation. Here, however, I want to consider a much more limited question but one for which the outcomes are more directly in our control: How do we, as researchers, approach our work ethically where new data collection and analysis tools are concerned? How do we do ethical research in an age of big data?

Traditionally, institutional research has been constrained by two factors: (1) physical limitations, such as difficulty recruiting participants and ensuring their representativeness, and (2) ethical considerations, such as possible harms caused to participants and protections for vulnerable populations. How do these factors scale individually and jointly with advancements in technology? How might changes in the former – say, the ability to canvass large numbers of people through public social media posts at little to no cost – necessitate changes in the latter – say, principles of

informed consent? In short, how do technologies like digital archives, blogs, massive online surveys, crowdsourcing such as Mechanical Turk, social media platforms such as Twitter, Facebook, tumblr or Instagram, dating apps such as OkCupid, Tinder or Grindr and geolocation services (alone or in combination with the others) invite fresh consideration of questions in research ethics?

An Ethical Framework

To address these questions, I want to borrow from "The Ethics of Fieldwork," a teaching module developed by the Program for Ethnographic Research & Community Studies (PERCS) at Elon University [5]. This module covers all phases of the research lifecycle – design, implementation and dissemination – and organizes ethical questions into four broad categories of value: accuracy, humane treatment, informed participants and the necessity and applicability of the research (see Table 1). Together, PERCS' framework organizes 31 familiar ethical questions that arise in research, such as prediction of possible harms, leading questions and the availability of raw materials to other researchers.

Though many of these questions apply to any kind of research involving persons, I want to highlight here some particular concerns introduced by big data. Following that, I'll suggest a few general strategies for conducting ethical research.

New Questions Raised by Big Data

Research using big data raises a number of new ethical questions in the areas of participant selection, invasiveness, informed consent, privacy/anonymity, exploratory research, algorithmic methods, dissemination channels and data publication.

Participant Selection. The availability of online data (for example, blogs, fora, social media posts) makes them tempting targets for research; never before has so much content been publicly available for inspection and from so many people. But even though large numbers of people contribute content online, not all spaces are representative or even appropriate for a given study. Each social media platform, for example, is known to have different demographic characteristics. Twitter saw a significant rise between 2013 and 2014 in online adults who are male, white, ages 65 and older, live in households with an annual household income of $50,000 or more, college graduates and urbanites, unlike Instagram users, who were more likely to be young adults, women,

TABLE 1. The ethics of fieldwork (PERCS)

	Formation: Developing the Proposal	Conduct: Behavior in the Field	Communication: Making the Work Public
Accuracy	1. Basic topic of the study 2. Self-fulfilling study 3. Sampling and participant selection	13 Leading questions 14. Biased researcher 15. Biased informants	24. Truthfulness and veracity 25. Meeting audiences' expectations
Humane Treatment	4. Prediction of possible harms 5. Selection of methods 6. Obligation to informants	16. Establishing rapport 17. Learning local norms of conduct 18. Negotiation of defined harms—learning local concerns 19. Participants as exemplified or exotic	26. Will participants be represented in ways they can understand? 27. Embarrassing revelations
Informed Participants	7. Degree of anonymity or confidentiality 8. Representation of researcher identity 9. Sampling and participant selection 10. Self-assessment of ability to conduct the work	20. Power differentials in fieldwork 21. What and how much can we promise?	28. Participants changing their minds after the study 29. Power differentials in writing
Necessity & Applicability	11. Motivations for doing the work 12. Possible applicability of the work for the participants	22. Learning local knowledge needs 23. Learning locally desired applications or service	30. Publication and distribution channels 31. Availability of raw materials to other researchers

Hispanics and African-Americans, and those who live in urban or suburban environments [6]. Any study using social media data should note these limitations, although many use such data uncritically to make claims about the general population, simply because such data is readily available. Employing data from any online community without engaging in routine questions of sample reliability and representativeness will compromise research – and its ethical merit.

Invasiveness. Ethical research makes the least possible impact on subjects, asking or collecting only as much as is needed to answer its questions. (The applied *results* of research, on the other hand, are hopefully beneficial, widespread and accrue to participants as well as others.) More invasive forms of research may cause direct harm to participants or otherwise violate ethical norms, such as appropriating value or violating privacy constraints. Big data, though prevalent and easy to obtain, are not necessarily the least invasive, since they often contain much more information than is strictly necessary to carry out a particular study. Collecting an entire profile of social media posts to investigate one topic, for example, may inadvertently reveal other information about the authors, information that could be damaging or used for other purposes. When combined with timestamps and geolocation information now embedded in many posts and images, the information may be enough to compromise anonymity and personally identify some individuals. Ethical researchers are careful to collect and analyze only what is necessary to their research design, either refusing to collect additional data in the first place or perhaps removing such data immediately from large datasets they obtain before taking further steps with that dataset.

Informed Consent. Informed consent is a cornerstone of research ethics, seldom overridable and then only when informing participants of specific information would fundamentally compromise the design of the research (for example, placebo effects). Even in drug trials, however, participants are told that control and experimental groups exist (just not the specific group to which they belong), and consent is obtained before the study begins. Big data collection, by contrast, often proceeds without even informing "participants" that a study is underway, much less asking for their consent. Researchers may point to terms of service or the public nature of posts to justify this behavior, but it is far from clear that these arguments pass ethical scrutiny. In some cases, it may be impossible to obtain consent in advance of the data being created, for instance, in a study that examines older posts. In such cases, we may need to develop an alternative model of consent, one that moves from pre-research activities to in- or post-research activities and allows users to opt out of studies and correct or remove their data after the fact. I'll say more about this model in suggestions.

Privacy/Anonymity. The digital format of data makes it less likely that datasets will disappear gracefully over time (as is the case with most analog datasets) and more likely that they will be interoperable, allowing comparison and merger between datasets collected for entirely different purposes. Information that a user later deletes online may still remain in a dataset collected years before – and conceivably remain there for generations. What is anonymous today may become personally identifiable tomorrow based on integration with new datasets and the introduction of new analytical methods. For example, a study of social movements that makes visible some population may also make members of that population vulnerable years or even decades later. The longevity of data and its unanticipated uses call into question researchers' ability to guarantee privacy and anonymity to subjects in the present – if such conversations even occur. Careful attention must be given to what variables are collected and disseminated and whether those variables are likely, in time, to integrate with other datasets in ways that compromise privacy and anonymity. To the extent possible, any necessary but possibly identifying values should be masked to prevent future identification.

Exploratory Research. In the case of big data (and especially the information visualization that supports it), researchers may not know exactly what they are looking for, instead using data for exploratory purposes. In these cases, it seems nearly impossible to inform participants of all anticipated harms and benefits in advance. Researchers, presumably, have some sense of what they hope to learn from the study; otherwise, they

run afoul of ethical concerns about its necessity and its impact on participants. If they know enough to predict these outcomes, they should know enough to predict a reasonable range of possible effects. To the extent that they cannot predict harms and benefits, they should make that clear to participants.

Algorithmic Methods. Researchers have extensive control over how their procedures and results are described and how participants are represented. Given the complexity of new algorithmic methods for examining big data, it is far from clear that most participants will understand how their data is being used, or perhaps even what the results mean in the context of the research questions. Consider a study that refines sentiment analysis using a large collection of tweets. How could a machine classifier be described in plain language? How might researchers describe nuances of irony, metaphor and ambiguity that complicate their results, while still presenting those results as rigorous and reliable in the context of the field? These and related questions invite us to reexamine the connections between researchers and participants, academic and public. Ethical researchers will strive to describe their procedures in ways that participants can understand, even when those methods are complex and technical.

Dissemination Channels and Participant Response. Ideally, research results should be made available to participants as well as fellow researchers, and participants should be afforded an opportunity to respond to the research in which they have been included. In the case of big data, many of the same platforms used to collect data can also be used to disseminate information and facilitate participant response. These steps complement the measures described in the previous section, which include describing procedures and results in plain language and representing participants in ways they can understand.

Data publication. Increasingly, federal and funding agencies require the publication of datasets along with research results. In addition to aiding public accountability and furthering research, data publication also brings ethical benefits for participants. By sharing their existing datasets, researchers minimize the need to collect additional ones (sometimes the same ones), making their work less invasive and potentially less harmful to new research population. Data publication, however, may be in tension with concerns raised earlier about privacy and anonymity, and researchers must weigh these ethical concerns against each other before deciding to release their data publicly.

Strategies for Ethical Research with Big Data

Strategies for ethical research with big data include the following:

(1) Involve participants more fully in the research process.

A number of the concerns raised above can be attributed to the increased distance between researchers and "participants" in big data studies. Researchers can gather data without even contacting participants, much less securing their explicit consent. Researchers are also remote with respect to the complex tools they use to analyze these large datasets, tools participants may have difficulty understanding. The venues in which researchers publish their findings may be inaccessible to participants because of paywalls, language or other barriers. To address these concerns, researchers might involve participants more fully at all stages of the research process. This involvement could be accomplished through a participatory research design or through a multi-stage approach, in which likely participants might be asked about their attitudes toward potential studies, any concerns they might have and what they would like to learn or receive from it – all in advance of finalizing the research design. These steps would help discover (rather than assume) what potential participants think about the study, especially its potential harms and benefits.

(2) Don't collect any information you don't think should be made public.

Given the longevity of datasets and their interoperability, researchers should adopt a general policy of avoiding data with personally identifiable information or information that could later be used to identify participants in connection with other datasets (for example, screenname). The potentials for data loss, theft and unintended consequences are high – but entirely mitigated when no personally identifiable information is collected in the first place. To the extent that researchers are unable to predict whether they can guarantee privacy and anonymity, they should make that known to participants.

(3) Inform participants of their status and provide them with opportunities to correct or remove data about themselves.

In most cases, researchers with the technical skills to analyze big data also have the ability to develop simple procedures and tools for notifying users of their inclusion in a study and for providing simple, user-friendly mechanisms for correcting or removing data about themselves. For example, a Twitter study might use the platform to alert users whose tweets have been included in the study and provide a link to remove their data before results are finalized or datasets, published. This post-hoc version of informed consent would empower users to refrain from participating in the final study and, at the very least, make them aware that and how their data is being used.

(4) Communicate research broadly through relevant channels.

Many of the platforms from which big data originates are also places where results and even data might be shared. By moving beyond traditional, academic venues, researchers extend the reach of their findings, helping participants understand what role their data played in the findings. In some cases, participants may be able to respond directly to research, affording them a degree of agency not found in, say, academic journal publications. These more public venues may require shifts in the ways researchers present their work (particularly with technical language), but these changes further the ethical principle that participants should be represented in research in ways they can understand.

Conclusion

Big data introduces big challenges for research ethics, but none that seem to go beyond the typical concerns raised by traditional research: participants should be well selected, informed of their status and consenting. Researchers should make every attempt to protect against harm and share their work accessibly for the benefit of participants and others. Though tools and methods have changed dramatically in an age of big data, this is an occasion for us to expand, not avoid, ethical conduct in research.

Acknowledgements

I am grateful to my co-presenters, Tara Conley and Jaime Riccio, and to attendees at the Technology & Culture Working Group session at the Cultural Studies Association 2015 Conference, who offered helpful feedback on these ideas. ■

Resources Mentioned in the Article

[1] Albergotti, R. & Dwoskin, E. (June 30, 2014). Facebook study sparks soul-searching and ethical questions. *The Wall Street Journal*. Retrieved from www.wsj.com/articles/facebook-study-sparks-ethical-questions-1404172292.

[2] Martin, M. (host), Junco, R. (interviewee) & LaFrance, A. (interviewee). (July 1, 2014). Facebook's newsfeed study: Was it ethical or a violation of privacy? *NPR*. Retrieved from www.npr.org/2014/07/01/327248369/facebooks-newsfeed-study-was-it-ethical-or-a-violation-of-privacy.

[3] Voosen, P. (December 15, 2014). Big-data scientists face ethical challenges after Facebook study. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Big-Data-Scientists-Face/150871.

[4] Kashmir, H. (June 30, 2014). Facebook added "research" to user agreement 4 months after emotion manipulation study. *Forbes*. Retrieved from www.forbes.com/sites/kashmirhill/2014/06/30/facebook-only-got-permission-to-do-research-on-users-after-emotion-manipulation-study.

[5] Elon University. Program for Ethnographic Research & Community Studies. The ethics of fieldwork module. Retrieved from www.elon.edu/e-web/org/percs/EthicsHumans.xhtml.

[6] Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (January 9, 2015). Social media update 2014: Demographics of key social networking platforms. *Pew Internet*. Retrieved from www.pewinternet.org/2015/01/09/demographics-of-key-social-networking-platforms.