Human-Subjects Protections and Big Data: Open Questions and Changing Landscapes

DRAFT VERSION

by Jake Metcalf / April 22, 2015
Produced for Council for Big Data, Ethics, and Society[1]

Framing

In previous Council conversations, the question of how human subjects protections relate to big data research techniques has risen at several points. This has theoretical and practical bearing on what regulatory regimes are available for deploying the Council's goal of expanding the reach and scope of data ethics. As members of the Council have noted multiple times, developing a robust data ethics agenda will involve placing the technological and mathematical side of data science in conversation with the long running debates about ethical regulation of social science and the humanities. This document contextualizes data ethics within the major changes in human-subjects protections already on the horizon and tracks several major themes of ethical regulations in social and behavioral science relevant to data ethics. At the April 2015 Council meeting we will discuss how the Council might engage existing and emergent regulatory policies and practices protecting human subjects from informational risk.

Introduction

As a nascent field, data ethics is in a peculiar position in relation to human-subjects protections. Discussions of human subject protections in big data research are necessarily discussions of how data ethics will relate to already established norms and institutions that have not yet grappled with the ways in which data research impacts human subjects. The Council's conversations have already identified many ways in which it is an odd fit for those institutions. Putting data science in conversation with theories and practices of human subjects protections generates quite a few discontinuities. Historically, the predecessors of data science—statistics, computer science, software engineering—have had little contact with the infrastructures of human subjects protections at universities and funding agencies. However, insofar as big data promises a new model of knowledge production that pulls knowledge out of vast and otherwise disconnected datasets that often contain human data, it also draws those technological and mathematical disciplines into much closer contact with human subjects and resultant ethical obligations.

[1] Funding for this Council was provided by the National Science Foundation (#IIS-1413864).

In that light, we suggest data science practitioners should be positioned in continuity with (though not identical to) a long-running conversation in the humanities and social sciences about researchers' responsibilities toward human subjects. boyd and Crawford (2012) write:

> *"As computational scientists have started engaging in acts of social science, there is a tendency to claim their work as the business of facts and not interpretation. A model may be mathematically sound, an experiment may seem valid, but as soon as a researcher seeks to understand what it means, the process of interpretation has begun. This is not to say that all interpretations are created equal, but rather that not all numbers are neutral."*

As the reach of data science grows, bringing computer sciences into closer contact with human sciences, the assumption of neutral computing has become less viable and a procedure for addressing research ethics has become a pressing concern.

It is reasonable that the predecessors of data science have been largely exempt from human subjects regulations (excluding medical research). Typically, human subjects protections do not apply to data that cannot be readily associated with an individual who bears risk of harm in their everyday life. We expect human subjects protections to apply when a human subject is involved, but in many cases technical measures have added a degree of protection to human subjects by creating substantial distance between their everyday lives and their data (Dwork & Mulligan, 2013; Dwork, 2011). Although such protections are widely accepted by Institutional Review Boards (IRB's) and other bodies tasked with human subjects protections, scholars have demonstrated how those technical protections are often inadequate by re-identifying sensitive data across distinct databases (see for example Sweeney, 2002; Malin & Sweeney, 2004). When considered in conjunction with the multiple, complex reasons for sharing, re-using and circulating research data (Borgman, 2011), the types of harms that may befall human subjects are challenging to predict.

It would be a mistake to respond to the discontinuities between data ethics and standard models of human subjects protection by attempting to strictly define what would count as a human subject in data science and what protections they require/deserve. From the earliest biomedical research ethics documents and policies, scientists, physicians, ethicists and patients have contested who exactly is the proper holder of human-subject status and the protections it entails. The regulatory definition of human-subjects, and exactly what protections they require, has always been contested (Shea, 2000; Annas, 1992), and therefore data ethics practitioners should keep in mind that they are already engaging with a moving target.

In that spirit, this document is intended to spark questions about what sorts of challenges big data research techniques pose for extant human-subjects protections, particularly if we work from an assumption of continuity between social and behavioral science research ethics and data science research ethics. How do we understand human subjects research in terms of big data? Do data ethicists need to be asking different questions about human-subjects protection? How well is

the regulatory concept of "human-subjects" formulated for data-intensive research? Do major conceptual constellations at play in research regulation align with data science methods?

Changing landscape of human subjects data and protections in the US

In the US, human subjects research protections are governed by Title 45 Section 46 of the Code of Federal Regulations, known as "the Common Rule." The current regulations went into force in 1981, and were last revised substantially in 1991. This federal regulation applies only to research that receives federal funding, although most institutions that regularly receive such funding require all researchers to have their proposals pass an IRB even if they aren't using federal funds. In 2011, the U.S. Department of Health and Human Services (HHS) and the Office of Science and Technology Policy (OSTP) published an advanced notice of proposed rulemaking (ANPRM), "Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay and Ambiguity for Investigators," to request input on major changes to the Common Rule. Many of the areas in which they requested input have direct bearing on big data research techniques (for example, resolving conflicts between the Common Rule and HIPAA regulations about protected medical data). The sheer volume of data available about human behavior, and the new techniques for handling that data, has introduced challenges for protecting human subjects not foreseen in the decades-old federal regulations. Additionally, HHS is attempting to respond to meta-studies that have identified burdensome and inconsistent applications of human-subjects regulations as a threat to the utility of human-subject protections (Abbott & Grady, 2011). The most substantial response to the ANPRM has been the 2014 National Academies report addressing regulation of social, behavioral and economic (SBE) research, "Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences."

Among the major changes proposed by the report's authors is the creation of new categories of research largely reserved for SBE research that would result in such research having little to no regulatory oversight. Although not explicitly stated in the report, these new categories would also appear to apply to much of the big data basic science funded by federal agencies.

Currently, the Common Rule defines human subjects as:

> (f) Human subject means a living individual about whom an investigator (whether professional or student) conducting research obtains
>
> (1) Data through intervention or interaction with the individual, or
>
> (2) Identifiable private information.

IRB's are thus currently tasked with reviewing any research that risks harm to an individual person which the researcher is interacting or intervening with in order to collect data. The scope of this mandate has always posed a problem for SBE researchers, especially because the regulations do not make a distinction between SBE and biomedical research that carry substantially different types of risk. SBE researchers vociferously contested the first drafts of the Common Rule because it applied the same level of scrutiny to medical experiments on humans as

sociologists' interviews of humans. The version adopted in 1981 established the category of "Exempt" research, which allows most types of SBE research to have very minimal oversight. "Exempt" is a bit of a misnomer, however, because researchers must submit applications to the IRB that identify the possible risks to subjects, include a consent form and describe a data protection plan. The National Academies report cites much confusion about, and inconsistent application of, the Exempt criteria by IRB's across universities and disciplines.

The report's authors are largely targeting the inappropriate and inconsistent application of research regulations to social science by IRB's. The report claims that these problems result substantially from the lack of guidance from the federal government about how to empirically (rather than intuitively) measure the risks and benefits of SBE research. Thus the bulk of the report consists of detailed guidance about how to effectively classify SBE research according to risks and benefits, ultimately enabling IRB's to consistently apply a set of reduced regulations.

The changes proposed by the National Academies would apply to all funded research, but would dramatically change the regulations of SBE research while largely leaving biomedical research regulations unmodified. They propose to drop the Exempt category and divide research into a "**not human-subjects research**" category and a "**human-subjects research**" category. The latter has three subcategories along an axis of informational risk.

Perhaps counter-intuitively, the "**not human-subjects**" category covers much research that makes use of data about humans. It applies when an investigator only uses information already in the public domain (including data that can be bought) and/or information that can be observed in public contexts. The National Academy report states:

> *"New forms of large-scale data should be included as not human-subjects research if all information is publicly available to anyone (including for purchase), if persons providing or producing the information have no reasonable belief that their private behaviors or interactions are revealed by the data, and if investigators have no interaction or intervention with individuals. Investigators must observe the ethical standards for handling such information that guide research in their fields and in the particular research context." (National Academies Press, 2014: 4)*

In other words, data about humans would not be considered "human-subjects" data if contextual factors outside of the control of the investigator had already made that data adequately public and/or benign. Even publicly identifiable information would be permitted as long as the subjects had no reasonable expectation of privacy. The primary trigger to make research reviewable appears to be direct intervention of the investigator in an individual subject's private behaviors. If it is the investigator's activities that generate the data and resultant risk then the research is considered to be about human "subjects"; if the subjects can be construed to have provided the data prior to the investigator's activities then the research does not trigger review. The report relies on the definition of "publicly available data" created by the Inter-University Consortium for Political and Social Research, including data repositories, de-identified data, pre-existing data, publicly available information, public-use data files and restricted access data (NAP, 2014: 41).

If the research is categorized as **human-subjects** then it would fall into one of three categories arrayed by increasing levels of informational risk: **excused research**, **expedited review** and **full review**.

The National Academies report states that the **excused** research category "is intended to cover research involving only informational risk either (a) where the risk of disclosure and the potential harm from it involve no risk or no greater than minimal risk or (b) where data protection plans and risk reduction mechanisms reduce the risk of disclosure to no greater than a minimal level." This category covers the study of already existing data that contains some private identification and/or studies that produce new data using benign and familiar techniques, such as interviews and surveys. The report lists examples such as, "a study of learning and distraction in which adult volunteers are asked to memorize nonsense syllables while being distracted by, for example, having to flag particular words among a string of words rapidly presented over earphones," and "a group of college students are given an anonymous survey about their mental health history and beliefs and attitudes toward school health policies" (NAP, 2014: 51-53). Human-subjects research that is "excused" would still need to be registered with an IRB office, but would not require permission from that office to be conducted. The report's authors call for some degree of random, non-onerous auditing to ensure that basic data protection practices are followed, but the excused category is largely intended to move low informational risk human-subjects SBE research out of the purview of IRB's.

Expedited review applies to projects with a slightly higher degree of physiological or psychological risk to human-subjects that can be made minimal with reasonable modifications to the research plan. The report recommends that "minimal risk" be defined in terms of physical or informational risks subjects encounter in daily life, such as routine medical examinations or educational tests. The report lists as examples of studies eligible for expedited review, "a study recruiting street-drug users in public spaces that has the potential to alert local police to prospective participants' illegal behaviors," and "a focus group study on parenting styles that asks for specific examples of physical discipline that may elicit reports meeting criteria of child abuse that an investigator is required by law to report" (NAP, 2014: 80). This category is considered "expedited" because the nature of the risks should keep the review period to under two weeks and would not require review from a full IRB panel. The report encourages HHS to to facilitate expedited review by supplying clear guidance to IRB's about available empirical (rather than intuitive) measures of harm and plausible methods to reduce risk. The report's authors recommend that expedited be considered the default status for all SBE research that is not excused.

A research project merits a **full** review if the project's risks would rise above the minimal physical and informational risks encountered in ordinary life, and those risks cannot be mitigated by obvious risk-minimizing research practices. Most biomedical research would continue to fall under the full review category, and the vast majority of SBE research would not.

If we understand data science research ethics as continuous with the history of social and behavioral science research ethics, the changing landscape of SBE research regulation immediately poses significant points of concern. For example, should research regulations consider data collected by social networks as publicly available or private? Would that determination depend on the specific network's rules of use? Is Twitter more public than Facebook because the former has much less privacy control than the latter? Or because the latter infamously has a "real names" policy? Does tweaking an algorithm in A/B testing rise to the level of an "intervention" that would categorize a project as human-subjects? Whose data privacy expectations can be considered "reasonable" when there are widely divergent and evolving behaviors? Furthermore, what exactly constitutes "daily risk" when people use many online tools daily but do not receive daily (or even annual) reminders their data may be used in scientific research without notice?

In the face of data-intensive business and research, we are still sorting out how to rigorously assess and communicate risk, calibrate expectations of privacy and identify responsible parties in the shifting alliances of universities and business. The value-added activities in big data research and commerce come from data analytics that pull together disparate databases. The same "publicly available" database that meets the proposed "not human-subjects" criteria may have radically different consequences for the subject whose data is stored therein when multiple public databases are analyzed together. For example, correlating an individual's multiple social media feeds and running a linguistic/semiotic analysis could reveal potentially damaging information—such as political views, sexual preferences, immigration status, etc.—that is not "public" in a colloquial because the subject has chosen to represent themselves partially and differently in various online spaces (boyd, 2008; Neuhaus & Webmoor, 2012). How terms of service define publicness can be very different from how actual human subjects conduct publicness in practice (Dwork, 2011).

What is "public" and "private" is therefore not cleanly answerable by looking at the conditions of the database, but the National Academy's proposed changes to the Common Rule would appear to eliminate any formal point at which these questions could be asked. If adopted, these new regulations could risk prematurely settling a number of significant, open questions about data ethics even as they address longstanding problems with the regulation of SBE research. The report does recognize that risk profiles are rapidly changing with data-intensive research techniques, and suggests establishing an independent body capable of providing continuing advice to IRB's about how to measure and mitigate such risk (NAP 2014: 112-115). More accurate assessments of harms and risks would be critical to ensure accurately and consistently assigning projects to the correct regulatory categories. A concrete step the Council could take moving forward would be to recommend to HHS some due process procedures for recognizing emergent ethical conditions in data analytics.

Social science as a misfit for human subjects protections

Although we contend that data science research ethics should be framed as continuous with human science research ethics, we do not intend to imply such continuity will result in an easy fit

between data science and existing research regulations. Indeed, SBE researchers have long struggled with the one-size-fits-all approach of IRB's. As noted above, human subjects regulations were largely designed with biomedical research in mind and the statutes make no distinction between types of research. For that reason, SBE researchers strongly opposed original drafts of the Common Rule and fought successfully for exempt and expedited categories of research regulation.

One early complaint about social science research regulation concerned the presumed moral and political neutrality of the concepts "human subjects protection" and "informed consent." Duster et al. (1979) contend that human subjects protections intended for vulnerable populations can inadvertently result in reinforcing political disparities. They cite a field study of racial housing discrimination as an example of research that could not be conducted if consent from all parties were required. Receiving informed consent would result in alerting racist landlords and realtors, thus altering the outcomes of the study and hiding the discrimination under a veneer of neutral care for all human subjects. Duster et al. were writing prior to the 1981 publication of the Common Rule that included the Exempt category, which in all likelihood would have enabled the research to proceed as a field study. However, it serves as an effective reminder that the individuals whose well-being is protected by human subjects protections is not a neutral or resolved matter.

Kelman (1982) argues that human subjects protection in social sciences should be indexed to the particular methods being used in any study. Methods that involve deception or highly structured laboratory experiments that manipulate a person's identity necessitate substantial oversight, but surveys and field studies largely carry much less risk to individuals and therefore require less oversight. Other social scientists have argued *any* oversight of social science, especially ethnographic studies, is inappropriate. Librett and Perrone (2010) claim that standard human subjects protections and ethnography operate at ethical and epistemic odds with one another, and common human subjects protections in university settings undermine ethnographic knowledge. Ethnography aims to protect subjects by intervening in the field as little as possible and providing subjects with maximal anonymity, whereas IRB's and consent procedures ultimately require interventions in the field that increase disruption and risk confidentiality. Similarly, Dingwall (2008) argues that preemptive regulation of research in the social sciences and humanities (other than some invasive psychological procedures) ultimately causes more harm than good. Citing cases of absurd over-regulation in the US and UK [see for an extended example Shea (2000)], Dingwall argues that regulation of ethnographic and social science research according to the mores and disciplinary norms of biomedicine undermines the disciplinary self-governance of social scientists and therefore ultimately reduces our ability to understand society.

Social scientists involved in research closer to big data have made similar claims (Walther, 2002; Keller, 2003). Basset and O'Riordan (2002) argue that human subjects protections do not effectively apply to many forms of Internet-based social science research. Whereas human subjects protections are geared toward blunting interventions in individuals' lives, they contend that much of the Internet is best understood as the cultural production of texts and therefore largely exempt from research regulation. They claim that metaphors representing the Internet as a

social space has led to misplaced efforts to regulate what is ultimately textual research. Neuhaus and Webmoor (2012) similarly contend that much data-intensive, or "massified," social science research is a poor fit for human subjects protections because it has reconfigured the internal relations between researcher, subject, and data such that the regulations focussed on a researcher's interventions with individuals no longer hold. They instead suggest that researchers adopt a model of "agile ethics" that emphasizes open and transparent commitments on the part of individual researchers. Such commitments could be maintained within a peer network with whom researchers "contract" where the traditional model of contracts between researchers and subjects is no longer possible. This produces a parity of exposure in which the researcher bears some risk of misbehavior being exposed. Privacy scholars have made similar claims about the shifting concept of individual privacy, noting that the liberal subject for whom privacy protections were designed is no longer a relevant construct (Cohen, 2013), that privacy should be reconfigured as a well-managed flow of information rather than a restriction (Nissenbaum 2004, 2009), and that maintaining effective social scientific knowledge production ultimately requires a model of privacy built on trust rather than restrictive data-access controls (Daries et al., 2014).

Social scientists clearly continue to chafe against human subjects protections, particularly as they seek to bolster or innovate social science's epistemic commitments. Indeed, in some cases social scientists have defined their commitments against the conceptual frameworks that undergird human subjects protection. As big data techniques proliferate, it is plausible that we will see similar co-emergent reconfigurations of human subjects and human data that require different ethical frameworks.

Using consent to mediate risks

Informed consent plays multiple roles in research ethics. Foremost in the ethics literature, it is understood as a critical point for recognizing the autonomy and dignity of individual subjects. It also has historically acted as a bulwark against paternalism toward patients and research subjects on the part of physicians and researchers. Informed consent offers a point at which individuals can establish their own criteria for weighing the risks and benefits of participation in a study. Researchers are obligated to never undertake any research that has individual risks grossly exceeding the distributed benefits of the research, and research subjects are empowered to decide how that risk/benefit calculus applies in their own lives.

Achieving informed consent requires a substantial investment in ethics infrastructures. In his discussion of the history and ethics of experimentation on dying patients, Annas (1992) notes that the rules adopted following the Nazi medical atrocities—the Nuremberg Code and Helsinki Declarations—were immediately objects of contention from the medical and scientific establishment, particularly around issues of consent and the rights of research subjects to weigh risks and benefits for themselves. The Nuremberg Code places direct responsibility for ascertaining the quality of consent procedures on the individual researchers: "The duty and responsibility for ascertaining the quality of the consent rests upon each individual who initiates, directs or engages in the experiment. It is a personal duty and responsibility which may not be delegated to another with impunity." However, the subsequent Helsinki Declaration and its revisions sought to carve out space for peer review committees to calibrate degrees of paternalism

that can supersede informed consent, particularly in developing countries that lack clinical infrastructures. As Annas notes, the language of the Nuremberg Code invokes natural law and inalienable rights, whereas the Helsinki Declarations leaned more heavily on the positive good of research as a justification for accepting different forms of consent. That medical researchers and clinicians would seek to carve out exceptions to consent procedures in the developing world illustrates the extent to which formal ethical obligations are inextricable from the presence of institutional or bureaucratic structures and norms (Benatar, 2002; Angell, 1997; Petryna, 2006).

It is reasonable to postulate that we are more likely to see consent devolving from an individual right to an informal community standard where there is little or no infrastructure to enable formalization of norms like informed consent. As Council discussions have noted, computer scientists have rarely had access to, or fallen under the regulation of, the ethics infrastructures built to regulate biomedical and behavioral sciences. The long-standing push and pull between formal and informal models of consent is certainly echoed in the debates about the equivalency of informed consent and end-user license agreements (Chee et al., 2012). Prominent examples of less-formal models of consent replacing traditional informed consent include the Facebook emotional contagion research debacle and Apple's recently released Research Kit. Moving in the the other direction, researchers at the Personal Genome Project have operated outside of the typical ethics infrastructures to avoid certain NIH restrictions they feel are onerous and arguably unethical, yet they have developed their own model of informed consent that is substantially more rigorous than what is typically used in biomedical research (Ball et al., 2014).

Still, ethics infrastructures alone are problematic proxies for the individual researcher's responsibility to ascertain whether research subjects have adequately consented to the research. Medical ethicists have noted that the emphasis on patient consent in medical practice both empowered individuals to more vocally express their preferences and burdened them with the responsibility for balancing complex measures of harm and benefit (Grodin, 1992). Given that substantial responsibility placed on the research subjects, physicians risk treating the patient consent procedure as an end in itself rather than as one part of an obligation to engage with the subject. Formality can supplant engagement, and genuine attempts to scale risks and benefits can be undermined by reliance on that formalized infrastructure and lack of attention to the inherently social context of consent procedures (Corrigan, 2003).

Paul Ohm (2013) notes a similar dynamic with regards to big data boosterism and the ease with which supposed benefits can be cited to wave off discussion of risk. Ohm cites the Google Flu Index as an example of the supposed benefits of big data tools trumping the need for careful analysis of privacy risks and transparent engagement with users. Ohm writes:

> *"While Google's users likely would have acquiesced had Google asked them to add 'help avoid pandemics' or 'save lives' to the list of accepted uses, they never had the chance for a public conversation. Instead, the privacy debate was held—if at all—within the walls of Google alone. By breaching the public's trust, Google has expanded researchers' ability to examine our search queries and given them a motive to focus in particular on some of the most sensitive information about us, our medical symptoms."*

If, as the Nuremberg Code exhorts us, individual researchers are ultimately responsible for calibrating consent procedures to plausible risks and benefits, how should data scientists cope with big data boosterism that so skews risk/benefit management in the manner noted by Ohm? Thus far, the most notable public scandals in big data research are cases of corporate researchers algorithmically manipulating users' lives for benefits that largely amount to improved consumer experiences. If we are to understand data science as part of a trajectory of human research practices, then ultimately there need to be languages and mechanisms that can carefully engage in such calibrations and facilitate individuals' ability to deliberate on their own non/participation.

Property, vulnerability & subjectivity

Who properly has access to data, and whether there are any human subjects whose well-being needs protection, are questions inextricably linked with complex questions of property, access, vulnerability and group membership (Levine, 2004).

For example, Reardon and Tallbear (2012) examine the sometimes surreal claim made by geneticists that scientists should have rather unfettered access to aboriginal biological materials because aboriginal people represent a primitive state of *all* humankind. By virtue of being a descendent of primitive people(s), Euro-American scientists claimed that they should have access to extant "primitive" peoples' biological material/data. In particular, population geneticist Spencer Wells would attempt to convince Australian Aboriginal people to donate their biological material by referencing their obligation to help Euro-Americans establish their own "songlines" that tell history through science. Reardon and Tallbear write, "If indigenous people represent modern humans at an earlier point in evolution, then indigenous DNA is part of modern humans' inheritance and, thus, property. This implies the further right to study that DNA." Tallbear and Reardon note that similar reasoning can be identified in the Havasupai scandal and lawsuit at Arizona State University, one of the most significant bioethics scandals of the last decade. In that case, researchers used diabetes research—arguably helpful to a tribal population with high rates of diabetes—as cover to research the genetics of schizophrenia and inbreeding—a topic that is both potentially embarrassing and not consented to by the tribal members. Reardon and Tallbear diagnose these serious ethical lapses as a symptom of scientists establishing assumed property interests in any data or property simply because science is 'good for everyone.' They write, "When genome scientists view their science as neutral—that is, in the interest of all (including groups such as the Havasupai)—they miss this assumed property interest." In such cases, researchers take advantage of slippery distinctions between the interests of individuals, groups and all of humanity, functionally avoiding responsibility for properly gauging human subjects protections. This highlights the importance of interdisciplinary training in data research ethics, especially in disciplines unfamiliar with human subjects protections.

Similarly, Radin and Kowal (2015; see also Radin 2014) detail the unanticipated ethical challenges created by frozen biological materials. Starting in the mid-20th Century, public health officials and biological anthropologists invented and standardized cryopreservation techniques to store biological samples in liquid nitrogen for decades. These samples are lively once again after a 50-year deep freeze because the explosion of data-driven biology has made it possible to

interrogate biological material in many new ways. These freezers can be thought of as hard drives storing data indefinitely before scientists even knew what data might be present. Playing off of Foucault's *biopolitics* (the power to create life and let die), Radin and Kowal propose the concept of *cryopolitics* (the power to not let die) to theorize how this power to freeze and thaw alters human subjectivity. Especially pertinent here is the problem of biological samples collected under colonialist ethical regimes that are now considered highly troublesome and would not pass current standards. How should such samples be allowed to mingle with samples vetted by more rigorous contemporary human subjects protections? Should biological material collected in the past under out-dated ethical regimes be repatriated, and should any data collected thus far be destroyed? How should harm and benefit be gauged when property relations vary so widely between colonialist and colonized cultures?

The troubled history of biologists and anthropologists appropriating indigenous biological samples and cultures carries a different tenor than questions of, say, using large publicly available databases in social media research. However, it should serve as a precaution that how we assemble notions of property and consent is imbricated with unsettled notions of race and group identities. There is no stable, universalizable understanding of who owns what and who shares which interests. Furthermore, it should not be assumed that the present ethical regime, nor the discourses of property and membership, will be stable over the lifespan of the data. Indeed, big data research will inevitably introduce unforeseeable ethical challenges derived from the indefinite lifespan of digital data.

Conclusion

Big data research techniques will place new ethical burdens on fields unfamiliar with research ethics regulations. Bringing those fields into conversation with long-running debates about the ethical regulation of social and behavioral sciences how the core conceptual constellations of research regulation—informed consent, risk, harm, ownership, etc.—are stretched by big data techniques.

The current regulatory regime under the Common Rule offers few places to formally require data scientists to check their research against these issues. The proposed changes to the Common Rule will offer even fewer. While there are meaningful doubts about the appropriateness or utility of bringing big data research (or social science in general) under such regulation, the Council may have the opportunity to influence how that regulation will occur.

Works Cited

Abbott, L., and C. Grady. 2011. "A Systematic Review of the Empirical Literature Evaluating IRBs: What We Know and What We Still Need to Learn." *J Empir Res Hum Res Ethics* 6(1): 3–19. doi: 10.1525/jer.2011.6.1.3

Angell, Marcia. 1997. "The ethics of clinical research in the Third World." *New England Journal of Medicine* 337: 847-848.

Annas, George J. 1992. "The changing landscape of human experimentation: Nuremberg, Helsinki and Beyond. *Health Matrix: Journal of Law-Medicine*. Vol. 2 Issue 2: 119.

Ball, Madeleine P., et al. 2014. "Harvard Personal Genome Project: lessons from participatory public research." *Genome Med* 6.2: 10.

Bassett, E., and K. O'Riordan. 2002. "Ethics of Internet research: Contesting the human subjects research model." *Ethics and Information Technology* (4)3: 233-247.

Beecher, Henry. 1966. "Ethics and Clinical Research." *New England Journal of Medicine* 274: 24: 1354-1360.

Benatar, Solomon R. 2002. "Reflections and recommendations on research ethics in developing countries." *Social Science & Medicine* 54.7: 1131-1141.

Borgman, C. 2012. "The conundrum of sharing research data." *Journal of the American Society for Information Science and Technology*. 10.1002/asi.22634

boyd, danah, and Kate Crawford. 2012."Critical Questions for Big Data." *Information, Communication & Society* 15:5: 662-679. DOI: 10.1080/1369118X.2012.678878.

boyd, d. 2008. "Social network sites: the role of networked publics in teenage social life." In *Youth, Identity, and Digital Media*, ed. Buckingham, David. Cambridge, MA: The MIT Press 119–142.

Chee, Florence M., Nicholas T. Taylor, and Suzanne de Castell. 2012. "Re-mediating research ethics: End-user license agreements in online games." *Bulletin of Science, Technology & Society* 0270467612469074.

Cohen, Julie E. 2012. "What privacy is for." *Harv. L. Rev* 126: 1904.

Corrigan, Oonagh. 2003. "Empty ethics: the problem with informed consent." *Sociology of Health & Illness* 25.7: 768-792.

Daries, Jon, Justin Reich, Jim Waldo, Elise M. Young, Jonathan Whittinghill, Andrew Dean Ho, Daniel Thomas Seaton, and Isaac Chuang. 2014. "Privacy, Anonymity, and Big Data in the Social Sciences." *Communications of the ACM* 57(9): 56-63. doi: 10.1145/2643132.

Dingwall, R. 2008. "The ethical case against ethical regulation in humanities and social science research." *21st Century Society* 3(1).

Duster, T., D. Matza, and D. Wellman. 1979. "Field Work and the Protection of Human Subjects." *The American Sociologist* Vol. 14, No. 3, 136-142.

Dwork, C. 2011. "A firm foundation for private data analysis." *Communications of the ACM* 54.1: 86-95.

Dwork, C., and D. K. Mulligan. 2013. "It's not privacy, and it's not fair." *Stanford Law Review Online* 66: 35.

Grady, C. 2010. "Do IRBs Protect Human Research Participants?" *JAMA* 304(10):1122-1123. doi:10.1001/jama.2010.1304.

Grodin, Michael A. 1993. "The evolution of informed consent: Beyond an ethics of care." *Women's Health Issues* 3.1: 11-13.

Haggerty, K. 2004. "Ethics Creep: Governing Social Science Research in the Name of Ethics." *Qualitative Sociology* Vol. 27, No. 4.

Katz, J. 1993. "Ethics and Clinical Research Revisited: A Tribute to Henry K. Beecher." *The Hastings Center Report* Vol. 23, No. 5: 31-39.

Keller, H. E., and S. Lee. 2003. "Ethical issues surrounding human participants research using the Internet." *Ethics & Behavior* 13(3): 211-9.

Kelman, H. 1982. "Ethical Issues in Different Social Science Methods." In T. L. Beauchamp, R. R. Faden, R. J. Walters (eds.), *Ethical Issues in social science research,* 40-98.

Kowal, Emma, Joanna Radin, and Jenny Reardon. 2013. "Indigenous body parts, mutating temporalities, and the half-lives of postcolonial technoscience." *Social Studies of Science* 43.4: 465-483.

Levine, Carol, et al. 2004. "The limitations of 'vulnerability' as a protection for human research participants." *The American Journal of Bioethics* 4.3: 44-49.

Librett, M., and D. Perrone. 2010. "Apples and Oranges: Ethnography and the IRB." *Qualitative Research* 10(6): 729–747.

Malin, Bradley, and Latanya Sweeney. 2004. "How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems." *Journal of Biomedical Informatics* 37.3: 179-192.

Marshall, Patricia Loomis. 2003. "Human subjects protections, institutional review boards, and cultural anthropological research." *Anthropological Quarterly* 76.2: 269-285.

Neuhaus, F., and T. Webmoor. 2012. "Agile ethics for massified research and visualization." *Information, Communication & Society* 15(1).

Nissenbaum, Helen. 2004. "Privacy as contextual integrity." *Washington Law Review* 79.1.

Nissenbaum, Helen. 2009. "Privacy in context: Technology, policy, and the integrity of social life." *Stanford University Press*.

Noah, B. A. 2004. "Bioethical Malpractice: Risk and Responsibility in Human Research." *Health Care L. & Policy* 7 J.

Ohm, P. 2013. "The underwhelming benefits of big data." *University of Pennsylvania Law Review Online* vol 161: 339.

Petryna, Adriana. 2006. "Globalizing human subjects research." *Global Pharmaceuticals: Ethics, Markets, Practices* 33-60.

Shea, C. 2000. "Don't Talk to the Humans: The Crackdown on Social Science Research." *Lingua Franca* 10(6).

Steinbrook, R. 2002. "Improving Protection for Research Subjects." *New England Journal of Medicine* 346:1425-1430. DOI: 10.1056/NEJM200205023461828

Sweeney, L. 2002. "k-anonymity: a model for protecting privacy." *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10(5): 557-570.

Walther, J. B. 2002. "Research ethics in Internet-enabled research: human subjects issues and methodological myopia." *Ethics of Information Technology* 4: 205-16.