

Curtis M. Kularski
7 December 2016
Topics in Philosophy: Big Data Ethics
Final Paper

Big Data Social Research Ethics and the Belmont Report

The Belmont Report is the key guiding document for ethical research involving human subjects in the United States (Metcalf and Crawford). The scope of the Belmont Report and the research procedures that were inspired by it is not restricted to experimental research, but also extends into social science research. In the social sciences the Belmont Report guidelines protect subjects from unintentional emotional or social harm rather than the physical types of harm that are possible in biomedical experiments. One of the new tools to social science research is big data. Big data uses large data sets and algorithmic processing of information to seek results to research questions. Most big data is considered to be anonymized or detached from the subjectivity of the person that is described in each data point. Big data relies on data for virtually¹ an entire population rather than a selected sample. Through big data researchers have access to more information than is possible through conventional methods of conducting research. This is an advantage if used correctly, but it also presents new challenges for research ethics. Historically a research participant has had the capability to decline participation or even leave a study that was in progress. Big data collects data on a variety of aspects of a person's daily life without their informed consent. In academic research big data is often

¹ Realistically not all persons are included in most datasets, but the number of cases is so relatively high that excluded cases are often considered insignificant

classed as secondary research and is therefore exempt from some research guidelines under the rules of the institutional review board (IRB) at many institutions. Big data is functionally different than traditional research on many characteristics and compliance with Belmont Report guidelines is impractical in most cases. Under consideration in this paper are questions related to the ethical nature of big data research. First, The Belmont Report specifies the responsibilities of researchers to human subjects, should big data be an exception to those expectations? Second, is big data research actually human subjects research? If not, then is it possible to conduct social science research using big data? Finally, the Belmont Report is almost 40 years old, is research ethics in need of an updated core document that reflects the changing nature of human subjects research, social or otherwise?

The Belmont Report

The Belmont Report provides specific recommendations as well as general principles for the ethical conduct of research. The guidelines of the report are designed to ensure the preservation of the autonomy, agency and safety of participants that are involved in research during their participation and beyond. The three principles of the report are respect, beneficence and justice. The report also specifies requirements for informed consent, risk/benefit assessment and fair selection of subjects to meet the objectives of the principles (Ryan, Brady and Cooke).

Respect for persons, the first of the principles, requires that human subjects be treated as autonomous agents. This includes ensuring the protection of persons with

diminished autonomy. The primary way that researchers can accomplish this principle is through informed consent before the research begins. Informed consent involves presenting as much information about the research being conducted as is practical, disclosure of any potential harms, ensuring that the subject has comprehended the information and reinforcing the voluntariness of the participation in the research (Ryan, Brady and Cooke). From time to time researchers will alter some elements, such as the disclosure about research information, because it would interfere with the results. When information is withheld before a subject is involved in research it is expected that the researchers will debrief the subject at the conclusion of research and expose any deceptions used (Ryan, Brady and Cooke). It is expected that information about the research will be presented in plain language that is easy for most human subjects to understand without assistance (Wirshing, Wirshing and Marder 1508).

Beneficence, the second principle, requires that research participants not be intentionally harmed and that the researchers maximize benefits while minimizing risks for the subject. This principle calls for an assessment of the potential risks and benefits. Where possible those benefits and risks are to be weighted by how likely they are to occur. This requirement specifically bans “brutal or inhumane treatment of human subjects” (Ryan, Brady and Cooke).

The final principle, justice, requires that all people have the chance to benefit or risk equally. This does not necessarily mean that all members of a population should have the chance to participate, but instead it directs researchers to not single out one group for participation and do not disqualify a group from participation unless there is some relevant reason (Ryan, Brady and Cooke).

In 1991 the Belmont Report was codified into the Federal Policy for the Protection of Human Subjects, also known as Common Rule (Office for Human Research Protections 1). This paper will refer to the original Belmont Report as it contains insight into the moral frameworks used to construct the subsequent regulations.

Big Data

Big data is a collection of datasets that are generated in a systematic way over the course of daily life. The data is collected automatically without intervention, consent or, in many cases, the awareness of the subject. Control of the datasets is usually decentralized. Often organizations will sell the data they collect or trade it with other organizations. Data is collected through computer use, Internet searches, online shopping, credit card utilization, academic administrative databases, cameras mounted on cars, sensors mounted on streetlights, microphones in personal electronics, smartphone GPS use and a variety of other sources (Anagnostopoulos, Zeadally and Exposito 1497). A current trend in human-computer interaction research is the automatic recognition and classification of human emotional responses (Kularski 15). Any type of data that can be encoded in a machine-usable format is subject to inclusion in a dataset that can later be aggregated with other datasets to develop more complex simulations of individuals (Clarke 77).

The defining characteristics of big data are volume, velocity, variety and veracity. Volume indicates the quantity of data collected, which is only limited by the availability of storage. Velocity refers to how quickly the data can be collected and analyzed. Most data

is collected automatically, eliminating the delays that are encountered in older data collection techniques. Algorithmic analysis and specialized big data analysis techniques speed the process of making conclusions based on the data. Variety is the characteristic given to big data by the nature that it does not have to be limited by traditional data structures or the confines of a purely relational database environment (Clarke 83). Veracity is the certainty or uncertainty of the data (IBM Corporation 1; Berti-Equille and Ba 3). The four properties are used to describe a wide variety of datasets and combine them under the common label of big data.

In most cases datasets are assumed to be anonymized, in the sense that the datasets are disassociated from an individual's name, however if enough data is present it is possible to reconstruct the person's identity (Metcalf and Crawford 3). Big data is not typically reviewed for individual subjects, but is instead used to find patterns (Mutzel 2-3). Data analytics algorithms are used to construct models from the data points. Many algorithms are written for commercial use and are closed-source, obstructing the operation of the algorithm from view of the user (Neyland 51). Lack of transparency in algorithms limits the ability to explain exactly what they are doing with the data and limits the ability to diagnose results methodologically.

Big Data Social Science

The use of big data and big data techniques in the social sciences began with the digital humanities movement. Large dataset research initially consisted of applying machine learning and natural language processing to textual data analysis tasks (Mutzel

2-3). The attraction for social scientists is the presence of a ready-made dataset that has been collected, presumably cleaned, is sufficiently large for most tasks and can be analyzed efficiently (Millington and Millington 145-146). Big data social science is purely a quantitative area of study. Due to the massive amount of data it is only possible to represent the dataset in mathematical and statistical terms. Qualitative analysis, which relies on researchers selectively analyzing a limited set of cases to look for explanatory features, is not possible. Qualitative data, such as written responses, can be analyzed with big data analytics methods, but it is reinterpreted in a quantitative framework and must be explained with a quantitative narrative. In sociology the long standing debate between the benefits of quantitative over qualitative methods, which momentarily resolved through mixed methods, has become more divisive with the introduction of big data methods (Ruppert, *Rethinking Empirical Social Sciences* 270).

Uninformed Non-consent

In big data, informed consent does not exist. Data is usually collected without the subject's knowledge, often in ways that could be classified as surveillance. Any action taken in an electronic, public or private space is subject to being observed and recorded. In some instances, the subject may be presented with a document, describing what information is collected and how that information is to be used, perhaps with an option for the subject to consent or decline to the terms. Even in cases where information is provided the document is not written in language that most subjects can understand, preventing the consent from being obtained under informed circumstances. The penalty

for declining consent is usually denial of service or exclusion from some benefit that is unrelated to the data that the provider wishes to collect.

Unlike experimental, behavioral or social research, there is often no mechanism to withdraw consent or to request removal from a big dataset. In the age of big data, there is also little recourse for a person who does not wish to be surveilled. The Belmont Report states that “respect for persons demands that subjects enter into the research voluntarily and with adequate information” (Ryan, Brady and Cooke). The methods used to collect information in big data do not meet either of those requirements. Assuming that informed consent was given at the time the data was collected there would still be a challenge to this principle as it is not possible for the subject to be made aware of the full extent that their data would be used.

The importance of the consent aspect of the Belmont Report is related to the concept of privacy in big data. In essence, the concept of privacy as it applies to big data is the right of a person to exercise control over their own information. Infringing upon privacy undermines individual autonomy (Taylor 589-590). Privacy is an expectation of American society, which affords the opportunity to live without supervision or undue judgement. Legal due process is founded on the principle of privacy as a default state. The erosion of basic privacy threatens the capacity for a person to exercise full autonomy without fear of social judgement or unexpected penalty from entities that disapprove of an activity. Psychological research moves beyond the societal expectations of privacy to a cognitive expectation of privacy, which is essential to the management and maintenance of self-identity (Burkell 18). The nature of big data suggests that individuals will be lost in the data set among thousands, millions or billions of other subjects or

reduced to the output of a summary algorithm. While this is the way that big data is intended to operate there are no technical limitations to an individual entity being singled out and then all data points related to that entity being assembled into a detailed individual profile. In essence, a perfectly large dataset could allow for the assembly of a person's life history including age, gender, purchasing habits, criminal history, credit record, medical records and academic history. This extreme scenario may not be presently possible due to the lack of connectedness between data sources, but it demonstrates the potential threat to privacy in big data.

Big data sets are inherently exploitative due to the nature of how they are collected. The multitude of organizations (institutions) and agencies that record data in a way that contributes to big data reduces the choice for being excluded to the option of removing one's self from society. In addition to the almost compulsory participation in the dataset there is an exploitative element in the commoditization of the datasets themselves. Datasets are created by automated means from the unpaid labor of the subjects. Data brokers and other commercial entities benefit directly and indirectly from the subject's participation. Subjects are rarely compensated for their participation in the dataset.

Potential for Harm in Big Data

Minimizing harm is discussed at length in the Belmont Report. On the surface big data does not seem to cause harm as the data is anonymized and there are never any risky

experimental conditions applied to the subjects. However, there are many risks and harms associated with big data.

Data sets can be breached and subjects can be individually identified (Tene and Polonestsky 462). Some datasets are subsets of other data, such as medical or credit records, which are de-identified for analysis. The de-identification relies on an intentional separation of the identifying data from the transactional data. Improper data handling puts all of the subjects at risk. The other major risk of individual identification is in re-identification through the combining of data sources. Metcalf and Crawford describe the case of a British artist using the alias “Banksy” who wished to remain unidentified and took precautions to protect his identity. Researchers were able to combine multiple public data sources to reveal the artist’s name, physical address and schools he had attended. The most disturbing aspect of the case is that the researchers did not feel that there was any invasion of privacy in tracking and identifying the artist simply because they used public datasets to locate and identify him (Metcalf and Crawford 2).

The data collection process can be flawed and create bad data, resulting in flawed conclusions in research (Ioannidis 40). The truth claims surrounding big data are potentially harmful to societies, especially if used in policy decision making. Some supporters of the big data movement suggest that existing methods of conducting social, psychological and other science are obsolete (Clarke 76). One of the difficulties with accepting this perspective is that it is difficult, if not impossible, to replicate the findings of big data research. This limits the ability of the academic process to utilize safeguards that are designed to eliminate procedural bias and discover data collection flaws.

If there is benefit to big data, then groups that are excluded from big data are also excluded from its benefits. As much as big data and its collection are omnipresent in daily life for most, there are people who are outside of its reach (Lerman 58). The principle of justice in the Belmont Report would classify this as harm through injustice. People who are not included in the dataset are not included in the research. Similarly, people who are included in the dataset are at risk of being excluded. Social science research has always been directed at locating large correlations or attempting to explain things through cases that appear in the center of a bell curve, but big data's massive size pushes researchers even further from their cases (Metcalf and Crawford 3). In traditional research, cases that do not match the hypothesis are considered outliers and can be eliminated statistically. In some instances, the outliers get more attention from researchers because they are anomalies worth looking at and explaining. Big data does not afford that type of qualitative evaluation of data points that are disregarded.

Depending upon the perspective taken it may be said that the accumulation of data is in itself harm as it infringes upon human dignity (Tene and Polonestsky 461).

Metcalf and Crawford point out that the most dangerous risks in big data research are the risks that are not yet known. While this is likely true in almost every new research technique or procedure, big data is intentionally too large and too complex to be fully understood. Humans are creators of information and over time large amounts of data have been created with a particular understanding of privacy and individual information freedom. Now the existing ideas of information must be retrofit with the new concept of a big data reality.

Potential for Benefit in Big Data

Big data offers some benefit to social sciences. Large, pre-existing datasets provide a target for testing theoretical ideas without the need to collect data and then analyze it through traditional methods. Datasets can be obtained either publically or at less expense than ordinary sampling and participant-based research options. Big data also has the advantage that larger scopes can be considered, making research more efficient if a large-scale answer is sought. If used correctly there are also advantages to research accuracy from the excessively inclusive datasets.

For participants the benefit is that there is no inconvenience of needing to answer a survey or participate in a study, everything is collected from activities they are already doing or it is collected passively.

Balancing Benefits and Rights of Subjects

The predictive capability of performing data analysis on a population is of great benefit if the results are accurate. Public policy and others areas with an interest in resolving social problems can benefit from the power of advanced analytics and the backing of a truly representative sample of the population. The difficulty is in balancing those benefits with the privacy and autonomy of subjects. For big data to be at its most effective, as many subjects as possible must be included in the dataset. Privacy of subjects is therefore counter to the objectives of big data and may “undermine the public good” (Ioannidis 40).

Guidance of the Belmont Report on the concept of beneficence requires that the subject, either directly or indirectly through their community, should benefit from the research in which they are involved. While there are obvious potential benefits in social research using big data for large communities there are many fewer direct benefits because the research tends to be on a larger scale or implemented in a more general way. John Ioannidis questions the ethical implications of research that is conducted blindly in a big dataset where researchers peruse personal data in the hope of finding an answer to a problem rather than positing a functional hypothesis and then using the data to answer (Ioannidis 40).

A Question of Ontology: Is it still social science?

One of the challenges to considering the protection of human subjects in the context of big data is determining if there are human subjects in big data (Metcalf and Crawford 1). Data points describing the behaviors of a human subject are reflective of that subject and may be able to be used to simulate behavior of that subject, but it is also reasonable to consider whether the data is an independent entity. One interpretation is that the data is the person, or an extension of their personhood. The other interpretation is that the data is no more the person than footprints left behind, a mere indication of the person. Ruppert (The Governmental Topologies of Database Devices 119) suggests that the complexity of the data collected is so detached from the subject that it is not predictive of the subject but only capable of producing a set of probabilistic outcomes.

Common rule defines a human subject as “a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through an intervention or interaction with the individual or (2) identifiable private information” (Tene and Polonestsky 459). Under the first part of this definition, the initial collection of information would be considered an interaction and therefore does constitute human subjects research. The application of this definition does not apply to secondary uses of data as the researcher is not having any interaction with the individual, only their data, which places the burden on the originally collector of the data, which may not be subject to Common Rule requirements of research. The second component of the definition, the use of identifiable information, is a debated point as in the most direct sense the data does not name a specific person.

If the data is detached of the person and exists as its own entity, how can the data be relevant to social science research? The existence of the data for data sake would remain in the realm of computer scientists and not those who seek to understand the functioning of societies and communities.

The detached approach could also be problematic for some aspects of public policy decision-making based in big data. The detached perspective allows only for analyzing problems that appear in data, not problems that exist in grounded reality. Establishing problems in terms of certain social metrics contained in the data can be used to detect patterns that are likely to cause those problems in the data, but the underlying social problem may not be predicted by the metrics or patterns that have been used to advise on policy to correct the problem.

Big data is likely both a reflection of the subject and an independent entity in itself. Information contained in big data sets is inherently personal as it is a detailed collection of information about a person's activities and history. A person and their data cannot be unlinked. The complexity of the data sets and the process of algorithmic analysis embody a new entity. A continuously evolving dataset reveals patterns and provides insights into collective behavior that cannot be discovered by considering individual cases. The relationships between data are the difference in these perspectives.

Need for New Research Guidance

Tene and Polonestky describe applying the Belmont Report guidelines to big data research as “the challenge of fitting the round peg of data-focused research into the square hole of existing ethical and legal frameworks” (Tene and Polonestsky 459). Overall, the principles of respect for persons, beneficence and justice do not pose a problem for social science research utilizing big data. The codified requirements, specifically informed consent, directly conflict with the way that big data is collected and managed. Most datasets used by social science researchers are considered secondary use data and therefore are exempt from research regulations as they are assumed to present only minimal risk to the subjects (Metcalf and Crawford 1). This assumption does not account for the advanced computational capabilities or data analytics techniques that are now available to make use of previously innocuous public data (Metcalf and Crawford 2).

Researchers in the social sciences are not data experts in the way of computer scientists. Computer scientists lack the understanding of human subjects that social

science researchers possess. Ignorance on how the concepts of the human subject and big data interact leaves the human subjects vulnerable. Clear guidelines as to what types of research are appropriate in big data are needed. It may be the case that big data research can be conducted ethically on entire datasets without need for the burden of individual consent, but attempts to re-identify individuals constitutes an ethics violation. In this case, the change to research ethics policy would need to move away from being concerned with the data collection process but instead with the data handling and analysis process.

Conclusion

Big data and existing research protections under the Belmont Report/Common Rule are incompatible. Big data is a new and emerging field with unknown risks to subjects. In some areas, such as informed consent, the Belmont Report may be too restrictive and in violation of its own principles of beneficence. The gaps in Common Rule that allow public datasets to be utilized without detailed review of the research project by an institutional review board place additional risk on subjects in the dataset. The current state of big data research is established in a social environment where privacy is valued by many people and it cannot be assumed that there is a social agreement that the possible risks of big data are worth the potential benefits. Academic researchers, who tend to be early adopters of new technologies and new techniques, are embarking on a new type of research to which existing ethical frameworks were not prepared to adapt.

References

- Anagnostopoulos, I., S. Zeadally and E. Exposito. "Handling big data: research challenges and future." *Journal of Supercomputing* (2016): 1494-1516.
- Berti-Equille, Laure and Mouhamadou Lamine Ba. "Veracity of Big Data: Challenges of Cross-Modal Truth Discovery." *Journal of Data and Information Quality* (2016).
- Burkell, Jacquelyn Ann. "Remembering Me: Big Data, Individual Identity and the Psychological Necessity of Forgetting." *Ethics of Information Technology* (2016): 17-23.
- Clarke, Roger. "Big Data, Big Risk." *Information Systems Journal* (2016): 77-90.
- IBM Corporation. *Big Data & Analytics Hub*. 2013.
<http://www.ibmbigdatahub.com/infographic/four-vs-big-data>. 30 November 2016.
- Ioannidis, John P. A. "Informed Consent, Big Data, and the Oxymoron of Research That Is Not Research." *American Journal of Bioethics* (2013): 40-42.
- Kularski, Curtis M. "Emotions in the Social Aspects of Human-Computer Interaction." *Unpublished Manuscript* (2016).
- Lerman, Jonas. "Big Data and Its Exclusions." *Standord Law Review* (2013): 55-64.
- Metcalf, Jacob and Kate Crawford. "Where are human subjects in Big Data research? The emerging ethics divide." *Big Data & Society* (2016).
- Millington, Brad and Rob Millington. "The Datafication of Everything: Toward a Sociology of Sport and Big Data." *Sociology of Sport Journal* (2015): 140-160.
- Mutzel, Sophie. "Facing Big Data: Making sociology." *Big Data & Society* (2015): 1-4.
- Neyland, Daniel. "Bearing Account-able Witness to the Ethical Algorithmic System." *Science, Technology & Human Values* (2016): 50-76.

Office for Human Research Protections. *Common Rule*. 16 March 2016.

<http://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/>. 7

November 2016.

Ruppert, Evelyn. "Rethinking Empirical Social Sciences." *Dialogs in Human Geography* (2013): 268-273.

—. "The Governmental Topologies of Database Devices." *Theory, Culture & Society* (2012): 116-136.

Ryan, Kenneth John, et al. *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. Washington, D.C.: National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979.

Taylor, James Stacey. "Privacy and Autonomy: A Reappraisal." *Southern Journal of Philosophy* (2002).

Tene, Omar and Jules Polonestsky. "Beyond IRBs: Ethical Guidelines for Data Research." *Washington & Lee Law Review* (2016).

Wirshing, Donna A, et al. "Informed consent: Assessment of comprehension." *American Journal of Psychiatry* (1998): 1508-1511.