Routledge
Taylor & Francis Group

Journal of Mass Media Ethics

Exploring Questions of Media Morality

ISSN: 0890-0523 (Print) 1532-7728 (Online) Journal homepage: http://www.tandfonline.com/loi/hmme20

Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism

Joshua Fairfield & Hannah Shtein

Journal of Mass Media Ethics, 29:38–51, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 0890-0523 print/1532-7728 online
DOI: 10.1080/08900523.2014.863126

Routledge
Taylor & Francis Group

Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism

Joshua Fairfield and Hannah Shtein
Washington and Lee University

As big data techniques become widespread in journalism, both as the subject of reporting and as newsgathering tools, the ethics of data science must inform and be informed by media ethics. This article explores emerging problems in ethical research using big data techniques. It does so using the duty-based framework advanced by W.D. Ross, who has significantly influenced both research science and media ethics. A successful framework must provide stability and flexibility. Without stability, ethical precommitments will vanish as technology rapidly shifts costs. Without flexibility, traditional approaches will rapidly become obsolete in the face of technological change. The article concludes that Ross's duty-based approach both provides stability in the face of rapid technological change and flexibility to innovate to achieve the original purpose of basic ethical principles.

Big data, as a single technique, is big news. The Prism program, revealed by *The Guardian* and *Washington Post*, has focused public attention on the nature of mass market consumer datamining. Data flows into these systems from social networks, Internet hubs, and smartphone applications. The data are then aggregated and parsed for purposes of targeted advertising. These databases are at the disposal of government actors, often with only policy protections to prevent misuse.

At the same time that big data is dominating the news, journalists themselves have embraced big data social science techniques. Emily Bell (2012) notes: "The truth is, those streams of numbers are going to be as big a transformation for journalism as [the] rise of the social web. Newsrooms will rise and fall on the documentation of real-time information and the ability to gather and share it." Using these tools will require new skills. "Yet while social media demands skills of conversation and dissemination familiar to most journalists, the ability to work with data is a much less central skill in most newsrooms."

While journalists are embracing social science techniques, social scientists are undergoing a fundamental shift in the ethical structure that has defined the moral use of these techniques.

Manuscript submitted November 8, 2012; revision accepted October 9, 2013.

Correspondence should be sent to Joshua Fairfield, Washington and Lee University, School of Law, 4007 Sydney Lewis Hall, Lexington, VA 24450. E-mail: fairfieldj@wlu.edu

Much of social science ethics focuses on rights and responsibilities toward the individual human participant. Big data as a technique does not accommodate this well. There can be millions of research subjects, yet none of them has given traditional informed consent. The traditional focus of social science has been on physical, rather than informational harms, and on not harming individuals, but big data impacts communities as much (or more) than individuals. Yet the notion that information is not a cognizable harm is not supportable in the context of an information-based society.

This technological shift requires a rethinking of how ethical principles are carried out. The article proposes a duty-based framework based on the work of W.D. Ross because it maintains the notion of a core duty while allowing for flexibility to move at the pace of emerging technologies. Ross also serves as a useful bridge between media and social science ethics, since his framework has been influential in both fields. Several of the principles in Ross's work are already present as foundational principles of scientific research involving human subjects: autonomy, beneficence, nonmaleficence, and justice. The influential *Belmont Report* codifies these principles into guidelines for social science research. There are significant differences, such as the *Belmont Report's* combination of Ross's duties of beneficence and nonmaleficence into one principle, but the two frameworks stand on similar theoretical ground. These core principles serve as an organizing scheme to develop guidelines for the ethical use of big data in the media.

BACKGROUND AND ETHICAL FRAMEWORK

This part explores the rise of big data in journalism and describes some commonalities between social science research ethics and journalistic ethics. The goal is to provide insights from the changes currently taking place in social science ethics to serve as a starting point for the necessary development of an ethics governing the use of social science data techniques by journalists. The article does so by finding common ground in the source and development of both social science and media ethics, through the lens of some of the philosophical roots of both disciplines.

Social Science and Journalism

Journalists increasingly study social science methods. Data science techniques have long been used in financial journalism but have since spread to almost every other area. There have been broad calls in the journalistic community for the application of social science techniques to journalists' use of big data in journalism. For example, the Philip Meyer Awards are given for quality journalism projects that use computer-assisted reporting and social science techniques. Fred Schulte, Joe Eaton, David Donald, and Gordon Witkin of the Center for Public Integrity received first place in the 2012 awards for their project "Cracking the Codes," which "uncovered the vast scale of Medicare billing errors and abuses that have padded the incomes of thousands of medical professionals to the tune of more than $11 billion over the past decade" (Schulte, Donald, & Witkin, 2012). They did so by plotting Medicare payment codes over a period of years, noting that up-coding resulted in overcharging over a multiyear period.

As journalists learn social science techniques, they should also become familiar with social science ethics. Big data does not come without its problems. For example, danah boyd and Kate Crawford (2012) note that big datasets are not objective. boyd and Crawford, as well as others in the community, suggest applying social science methods as a means to temper the subjectivity of algorithms and datasets. Mark Hansen, director of Columbia University's Brown Center for Media Innovation, notes that data "have something to say about us and how we live. But they aren't neutral, and neither are the algorithms we rely on to interpret them. The stories they tell are often incomplete, uncertain, and open-ended" (Bell, 2012, p. 3). This necessitates in turn a journalistic check on the use of big data in the media. "Without journalists thinking in data," Hansen states, "who will help us to distinguish between good stories and bad? We need journalists to create entirely new kinds of stories, new hybrid forms that engage with the essential stuff of data" (Bell).

Ethical Framework

Because of this article's hybrid nature, it will first explore the common philosophical roots from which both social science ethics and journalistic ethics have drawn. This will lay the groundwork for the following discussions of emerging issues in social science ethics as applied to big data. The hope is that once the common roots of ethical inquiry are established, the emerging issues will serve as a useful guide for framing the use of big data in journalism.

The core principles of social science research are, as they are for all human subjects research, autonomy, beneficence, and justice (Childress, Meslin, & Shapiro, 2005). These principles are found in the flagship *Belmont Report*, a government document outlining guidelines for the ethical treatment of human subjects in research. The autonomy principle denotes respect for persons. Individuals are to be treated as autonomous agents. In practice, the autonomy principle translates to the need to obtain informed consent. Beneficence requires that the researcher minimize harm where possible and relate it proportionally to the potential benefit of the study. As the *Belmont Report* notes, "beneficence is understood as an obligation." Two complementary rules have been recognized as falling under the beneficence category: "(1) do not harm and (2) maximize possible benefits and minimize possible harms." The principle of justice is described in the *Belmont Report* as fairness in distribution of benefits (here, the benefits of research). "An injustice occurs when some benefit to which a person is entitled is denied without good reason or when some burden is imposed unduly."

While journalism is not human subjects research, journalists have already drawn from similar principles in an attempt to structure ethical standards. With the increasing use of big datasets in journalism, there is a corresponding need to expand the ethical framework to include the new discipline. There is already common ground. Journalistic ethics has profitably drawn from the same sources as did the principles in the *Belmont Report.* The philosophy of Ross (1930/2002) had a marked influence on the *Belmont Report* principles, and his approach has also been influential in media ethics. Ross therefore serves as a bridge and a way to translate ethical discourse between the two disciplines. One caveat is appropriate here: This article does not claim that Ross is the only or most prominent influence on the development of these ethical systems. Rather, the purpose is to establish that there are common roots between the basic principles of the ethics of social science research and those of media ethics, such that the following discussion of emerging issues in big data studies of social media will be profitable to both disciplines.

We start with an analysis of each basic principle, followed by a discussion of the commonalities and applicability to the use of big data in journalism. According to the authors of the *Belmont Report*, the principles of that report were influenced by Ross's philosophy (Childress et al., 2005, p. 14). Ross advanced an intuitionist and deontological approach to ethics. Although he did not reject consequentialism entirely, it did not drive his approach. Duties were to be performed because they were discernibly right, not because of their consequences.

This stands in significant contrast to utilitarianism, which has also significantly influenced media ethics (Christians, 2007). From a utilitarian perspective, if an action becomes too costly to perform in relation to the benefit it provides, it ceases to be moral obligation. If performance of the duty is, on aggregate, more costly than the overall benefit it provides, the action creates disutility. Ross's critique is that an action does not cease to be a duty merely because it becomes costly. Ross (1930/2002) holds that "a great part of duty consists in an observance of the rights and a furtherance of the interests of others, whatever the cost to ourselves may be" (p. 16). Ross argues that the very nature of a duty is that it is binding upon agents, even if there is a personal cost attached. The avoidance of aggregate harm is itself a "Rossian" duty. However, unlike in utilitarianism, where this is the primary consideration, this duty is one that must be weighed against several equally important duties. One may decide, in a particular situation, that certain duties and not others predominate. But each duty remains, even though it may be for the moment outweighed by other obligations.

This matters for social science and mass media ethics in the study of social media. The single greatest changing feature in this technological landscape is cost. As datasets get larger and algorithms get better, the cost to intrude on privacy has radically lessened. At the same time, the cost of securing informed consent from subjects has increased. Under a utilitarian framework, these radical cost shifts could upend the moral calculus. A deontological framework such as that proposed by Ross may perform better. It recognizes that one duty may in certain contexts carry more weight than another, but the duties remain intact. This approach provides useful insight into how ethics will change in the face of radical cost shifts in information processing.

Ross held that certain prima facie duties existed without *a priori* justification and were discernible by mature moral sense. Ross believed that these duties should then be weighted as part of a reflective process seeking to produce right action. Note that Ross indicated that his list was open, and subsequent thinkers have added principles to Ross' or have adapted extant principles. It is to these prima facie duties mentioned by Ross or those further developing his framework that we first turn as we consider potential roots of the Belmont Report principles.

The *Belmont Report* principle of autonomy can be seen as a Rossian prima facie duty (Meyers, 2003, 2011). While not originally included in Ross' list of duties, other ethical theorists have viewed autonomy as a principle that can be included within the Rossian framework. This principle is most commonly described as enshrining respect for persons. In the social science context, the autonomy principle requires strong informed consent for human subjects research. Subjects must be informed of the nature of the experiment and told about any potential risks. If the subject is a minor, the parents must be informed and must assent. These basic precautions are necessary because participants are instrumental in an experiment; they provide significant value. It is important to ensure that they are valued as individuals, rather than purely for the benefit they provide within the social science context. This is a Ross-oriented viewpoint: The duty is binding upon agents even if there is a personal cost attached. Thus, the duty to obtain informed consent cannot be avoided purely as a matter of convenience or efficiency.

In the search for common ground with media ethics, we note that prominent voices share the same analysis of the autonomy principle as applied to journalism. Meyers (2011), in his article "*Re*appreciating W.D. Ross," addresses the autonomy principle as a Ross prima facie duty (p. 326). Meyers claims that the autonomy principle requires that moral actors deal with others as "free beings whose moral autonomy must be honored" (p. 326). This is the same description one finds in the *Belmont Report*. Meyers argues that reporters have an obligation not to use people as mere means but admits that there is an instrumental element: "The key qualifier is 'mere,' as we regularly use one another for mutual benefit, with the distinguishing standard being *consent* ... " (p. 327).

Informed consent in the journalism context does of course differ from informed consent to human subjects research. But the principle should act as a familiar touchstone for journalists who encounter some of the difficulties in obtaining informed consent in the mass media context. For example, there are no informed consent forms in mass market big data databases. It is not possible to contact the contributors of the data; there are too many. A new approach, which accommodates the nature of big data projects and is consistent with the underlying principle of autonomy, must be developed.

The beneficence principle as established in the *Belmont Report* includes two components: an exhortation to not harm and the requirement to maximize benefits while minimizing harms. These two components fit precisely with Ross's prima facie duties of nonmaleficence and beneficence. Ross (1930/2002) defined beneficence as a set of duties that "rest on the mere fact that there are beings in the world whose condition we can make better in respect of virtue, or of intelligence, or of pleasure" (p. 21). For Ross, nonmaleficence duties are those "that may be summed up under the title of 'not injuring others.' No doubt to injure others is incidentally to fail to do them good; but it seems to me clear that nonmaleficence is apprehended as a duty distinct from that of beneficence, and as a duty of a more stringent character" (p. 21).

This line of analysis also appears in media ethics scholarship. Meyers (2011) notes that journalists should "do what [they] reasonably can to improve the situation of others. Included under beneficence is the journalistic (role-based) duty to take positive measures to try, via good reporting, to *prevent* harms caused by others or by natural events" (p. 327). Christians (2007) notes: "For reporters who arrive at the scene of an accident, their primary moral duty of beneficence is to treat other human beings with dignity and care; their second duty is providing information services to the reading or viewing public" (p. 122). This is a duty not just to the individual, but to the community. "A duty-oriented social ethics emphasizes democratization; the politics of community must be reconstituted ... " (p. 122). Similarly, the duty of nonmaleficence captures the idea that an actor should not by her own action cause harm. Meyers (2011) considers whether the personal data of key participants "are relevant to determining whether, for example, reporting on someone's actions would significantly undermine their ability for a flourishing life ... " (p. 325).

Again, there is significant common ground between social science and media ethics. Consider the problem of informational harms. On a cost-benefit analysis, leaking someone's data as part of a big dataset may not be catastrophic. In dollar figures, the cost to a consumer of being part of a data spill may be low. But the cost to the dignity and personhood of an individual whose entire search history has been exposed to the world can be significant, if not easily measured in cost-benefit terms. When social scientists or journalists study an online community, the very fact of observation may damage the sense of community and safety that the group has worked to create. It is not enough merely to get consent from one person as an entrée to the community.

The final *Belmont Report* principle is justice. Ross (2002) defined duties of justice as those that "rest on the fact or possibility of a distribution of pleasure or happiness (or of the means thereto) which is not in accordance with the merit of the persons concerned; in such cases there arises a duty to upset or prevent such a distribution" (p. 21). Thus, Ross's justice means that one must "give to persons what they have legitimately earned, and apply corresponding social structures ... in an unbiased manner ... " (Meyers, 2011, p. 327). The *Belmont Report* takes this a step further based on Rawls's (1971) theory of distributive justice. The question is who ought to bear the risks and share the benefits of research. "Individual justice in the selection of subjects would require that researchers exhibit fairness: thus, they should not offer potentially beneficial research only to some patients who are in their favor or select only 'undesirable' persons for risky research."

An example tying both social science and journalism together with regard to justice might be the impact of big data research on minority populations. Big data analysis can be used to generate statistics that are at best misleading, and at worst actively harmful. Consider the statistics on "black on black" crime. These statistics can be easily compiled from Justice Department datasets using data science tools. However, they do not tell the whole story: Most demographic groups commit more crimes against members of their own demographic. "In fact, all races share similar ratios ... the myth and associated fear of 'black on black' crime is sold as a legitimate, mainstream descriptive and becomes American status quo" (Williams, 2012, p. 1). The use of "black on black" crime statistics is therefore misleading and disproportionately burdens one segment of the population.

PROBLEMS

This section seeks to articulate the problem in greater depth by crystalizing the stresses that large-set datamining bring to bear on the principles of autonomy, beneficence, and justice. Size, aggregation, and informational harm all present emerging challenges to social scientists and journalists working with big data. Some of these problems bear on all three foundational ethical principles, and others affect only one or two.

Size

The raw quantity of data bears on each of the foundational principles of ethical research in a separate way. According to some estimates, Google processes more than 20 petabytes of data a day (Scott & Bracetti, 2013). Facebook has 1.10 billion active users; a single one of Facebook's data clusters contains 100 petabytes of data (Metz, 2013). Each petabyte is equal to 250 billion pages of text (Vance, 2012). When this much data are contributed by this many people, the nature of the relationship between researcher and researched must necessarily change. The question is how it can do so while still remaining true to foundational principles.

Raw number of participants makes informed consent problematic. Indeed, it is already problematic to call people with data stored in big datasets "participants." They were not asked, have not consented, and do not know most of the time that their data are being used. This creates a basic tension for ethical researchers. As boyd and Crawford (2012) note: "It may be unreasonable to ask researchers to obtain consent from every person who posts a tweet, but it is problematic for researchers to justify their actions as ethical simply because the data is accessible" (p. 672). The problem of quantity does not absolve researchers or journalists of the shared ethical obligation to respect autonomy. boyd and Crawford further state that "in order to act ethically, it is important that researchers reflect on the importance of accountability: both to the field of research and to the research subjects" (p. 672). The problems caused by the size of datasets put more, not less, of an onus on the researcher. This is in line with Ross's intuitionist and nonconsequentialist approach. Under Ross's duty-oriented framework, a duty may be weighed and evaluated, but it should not be abandoned merely because it becomes more difficult to perform.

Social scientists who use secondary datasets—datasets gathered by someone else—do not have to get informed consent a second time. Yet for the commercial datasets that are the majority of big data datasets, consent to research use was never expressly granted. Thus social scientists often rely on vague and hidden language in commercial clickthrough End User License Agreements to support research use of secondary datasets. This is not a tenable or ethically sound strategy. The informed consent that a researcher must receive from a subject is of a different kind and quality than the sort of consent granted by a clickthrough commercial online license agreement.

There are possible solutions. End User License Agreements do not currently inform users that data might be used for research purposes, but these agreements could be written to do so. Companies that use data for undisclosed purposes should be exposed. A further step relies on data granularity. Often data analysis techniques parse large datasets but yield individual results. Consider, for example, the AOL data spill in 2006. Millions of search profiles were released unintentionally to the public. News stories about the event, however, tended to focus on de-anonymizing specific search profiles, and connecting someone to the search history. It is possible, and indeed necessary, to get consent at that point. Just because big datasets are unwieldy from the autonomy standpoint does not mean that their results are. Data reported out may be limited enough that seeking traditional informed consent again becomes possible.

The size of datasets bears on the beneficence principle as well. The harm that a mistake can cause grows as a function of the amount of data one has. Consider a simple case of a data leak. In the past, if a reporter left some papers in a cab, then a source or subject might be damaged. But if a reporter leaves a laptop with a database behind in a cab, the potential for harm is far greater. Moreover, large datasets gather not only more people's data but also more of any one person's data. Again, consider papers left behind in a cab. There is a marked difference between leaving one of a subject's letters behind and leaving all email correspondence that the person has had for a period of years. The size of the dataset directly drives the severity of the harm.

At first blush, the size characteristic of datasets does not seem to impact the foundational principle of justice. The baseline of ubiquitous and continuous surveillance might mean that everyone is under an equal amount of surveillance. In short, although there is harm done, it is general, indistinct harm, not visited unequally or unfairly upon any given group. But upon closer examination, the quantity problem bears very much on the research principle of justice. The results of machine learning algorithms used to mine big datasets reduce the costs of searching out and detecting patterns. Thus, for example, such algorithms are reportedly used to read email texts, whether for Google's advertising, or the government's other purposes. There is still a cost to action, however. Datamining might detect any behavior, but businesses have limited advertising dollars and governments have limited enforcement dollars. Assume person A is a Muslim and person B is an evangelical Christian. Assume low transaction costs to parse data and limited enforcement dollars. What becomes apparent is that it is much more likely that the lens of deep datamining, as well as the business end of enforcement will be and are disproportionately directed at Person A (Chen, Reid, Sinai, Silke, & Ganor, 2008). The apparent egalitarianism of ubiquitous surveillance is anything but. Ubiquitous surveillance disproportionately impacts minorities. Political, religious, or other minority groups are much more likely to be the focus of limited enforcement dollars (Cesari, 2010). The deciding principle then becomes the exercise of enforcement discretion rather than either a function of the rule of law or an outcome derived from ethical first principles.

Aggregation

Aggregation causes tension between individuals and communities. All group members' data are captured. In social media settings this can mean that a database captures entire communities. This can be invaluable in researching a particular subject, but the difficulty is in respecting the rights of other community members who may not be the subject of research or who have not given consent.

Like size, aggregation bears on autonomy, beneficence, and justice. Aggregation affects autonomy because the admixture of data from multiple subjects complicates the question of who can consent to use of the community's information. Researchers may take individually granted consent too far. For research conducted on subjects using social media, the barrier between what is included in the study and what is not is rarely so clear (Nissenbaum, 2004). Consider research on an online community, where the participants are part of the community. Studying their conversations reveals information about other community members. The subject may consent, but other community members have not.

Moreover, researchers may become removed from their research subjects because they are only looking at data. It is difficult to feel attached to the potential concerns of any one subject when the dataset consists of thousands of subjects aggregated together. Researchers may justify their actions based on the aggregated good to the broader community. Consider boyd and Crawford's (2012) example of a Harvard research group that gathered Facebook data from more than 1,000 college students to study "how their interests and friendships changed over time" (p. 671). That the goal of the study is a useful or beneficial one does not override the fact that personal information was gathered from the students without their consent.

The tension between individual and community also impacts beneficence. What is good for the individual participants in the study is not always good for the community, and vice versa. It is easy to imagine that information about the individual would be beneficial to the community, while revelation of that information would harm the individual. A current trend in social science far too strongly favors individual benefits over community. It is even controversial, as a matter of social science ethics, whether researchers should consider harms beyond their individual research participants at all. Yet social media research can have an obvious and direct negative impact on communities. A researcher observing an online community might entirely unintentionally cause dissension and significant harm due to eroded privacy, despite taking serious precautions to obtain individual consent (Reid, 1996).

In aggregated data, there is not just tension between individual and community but also between subgroups within the community. This yields insight for the principle of justice. Out of any dataset, some people will be disproportionately affected by the data. Certain subgroups within the group may have more power while others are disadvantaged. Consider a study of Facebook "like" patterns (Kosinski, Stillwell, & Graepel, 2013). Such patterns can reveal political affiliation or sexual orientation, even for users who have not chosen to reveal that information. The disadvantaged group may rely on privacy as a measure against discrimination. Through analysis of aggregated data, researchers can create tools that make disadvantaged subgroups more vulnerable.

Another aspect of justice is equality of access to aggregated data. Big datasets require large computer arrays to process, large amounts of storage capacity to maintain, and often special relationships with industry to obtain in the first place. At a fundamental level, there are inequalities between researchers with respect to who has access to and the ability to effectively analyze large datasets. "This produces considerable unevenness in the system: those with money—or those inside the company—can produce a different type of research than those outside. Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access" (boyd & Crawford, 2012, p. 674).

INFORMATIONAL HARM

The trend in social science research has been to improve encryption and storage standards for research material, but to deemphasize informational harm in the ethics procedures surrounding approval of research projects. Indeed, "research ethics guidelines were originally meant to protect subjects from bodily harm rather than informational harm. Issues related to data release, data sharing, and unanticipated findings are still the subject of much debate and may not be dealt with adequately in current guidelines" (Szego, Buchanan, & Scherer, 2013, p. 1209). A proposed modification of the rules for human subjects research is underway in the form of the advance notice of proposed rulemaking (ANPRM). The ANPRM previews potential modifications to the Federal Common Rule, which is the regulatory framework drawn from the *Belmont Report*, and which codifies the rules of human subjects research. Changes to the Common Rule change the operative ethical framework under which science is conducted. As noted above, issues related to informational harm are a relatively recent development and therefore subject to change and contention. For instance, some social science organizations have submitted comments noting the need to amend the ANPRM's informational harm provision. However, the present discussion focuses on the ANPRM because the Federal Common Rule is still the primary source of guidelines on issues of human subjects research.

The ANPRM proposes that informational harm should not be a harm considered by institutional review boards under the revised rules. The problem has its roots in the biomedical beginnings of human subjects research. The rules were originally conceived for medical studies. Harm in the biomedical context is direct, physical harm. While the ANPRM's focus on physical harm makes sense in the biomedical context, it is problematic for the study of social media. There are three false premises: (1) informational harm is a distinct category of harm; (2) datasets are objective and do not harm merely by describing what exists; and (3) data in big datasets are anonymized, and thus information revealed through the dataset does not harm individual people.

Each of these premises is demonstrably wrong. In an information society, informational harm is not a separate category of harm; it is straightforward harm. Information impacts the real world. Even if the cause is information, the resulting harm is in no way purely informational. Leaks of location-based information pose a direct threat to the personal well-being of the subject, for example (Hirsch, 2006). The assumed objectivity of datasets is also incorrect. Datasets can harm merely by describing, and algorithms are not value-neutral. "What [big data] quantifies does not necessarily have a closer claim on objective truth—particularly when considering messages from social media sites" (boyd & Crawford, 2012, p. 667). Finally, anonymization is a failed protection. Big datasets permit de-anonymization. A subject's search history, social media messages, or contacts can identify her as easily as her name. In the Facebook study mentioned above, "what other researchers quickly discovered was that it was possible to de-anonymize parts of the dataset: compromising the privacy of students, none of whom were aware their data were being collected" (p. 672).

Although it is true that large datasets are often anonymized, the effectiveness of current anonymization procedures is likely to decrease as technology continues to advance. This already appears to be occurring. For example, while "a simply anonymized dataset does not contain name, home address, or other obvious identifier [,] ... if an individual's patterns are unique enough, outside information can be used to link the data back to an individual" (de Montijoye, Hidalgo, Verleysen, & Blondel, 2013, p. 1). In one famous example, for instance, a database of medical records was combined with a voter list to find the health records of the then-Governor of Massachusetts (Sweeney, 2002).

Additionally, mobility data is now "broadly available," thanks to "the advent of smartphones and other means of data collection" (de Montijoye et al., 2013, p. 1). Further,

> the uniqueness of human mobility traces is high [,]...mean[ing] that little outside information is needed to re-identify the trace of a targeted individual even in a sparse, large-scale, and coarse mobility dataset. Given the amount of information that can be inferred from mobility data, as well as the potentially large number of simply anonymized mobility datasets available, this is a growing concern. (p. 4)

Journalism may be able to require accountability from social scientists on this point. "Accountability requires rigorous thinking about the ramifications of big data ... " (boyd & Crawford, 2012, p. 673). From a media perspective, one cannot accept that harm should be dismissed merely because it was caused by the dissemination of information, rather than involving physical harm. Journalism ethics is substantially devoted to minimizing harm caused by information dissemination. Accordingly, social science ethics would benefit by incorporating the journalistic view that informational harm is just as real and important as physical harm. Because the big data technique is not objective and makes us especially susceptible to these types of harms, a developed ethics of journalism regarding dissemination of insights generated by big data is particularly necessary.

There are too many problems of accuracy and accountability when researchers draw from large, aggregated datasets without meaningful involvement with or investigation of the communities from which the data are drawn. Some research has begun to respond to this criticism. Mixed-methods research uses qualitative methods such as participant observation to ensure that there is a sufficient connection between researcher and research subjects to enable the minimization of harm. It is worth noting that these qualitative research methods are a close approach to some methods of journalism. If mixed-methods research can help with some of the problems of big data analysis, journalists may be able to do the same. Speaking with members of the community, hearing their concerns, and understanding some of the potential side effects of the release of information cannot be optional. Without it, unrestricted information dissemination risks to do much harm.

EXTENDING THE ETHICS OF BIG DATA

In the study of big data, it is easy to lose sight of an ethical framework because of shifting technology and costs. If cost or efficiency drives morality under a utility maximization principle, moral positions will need to change as quickly as the technology itself. On the other hand, the problem with deontological precommitments is flexibility. In a deontological framework, it is easy to lose sight of the flexibility necessary to reweight priorities in the face of technological change.

Ross's philosophy is attractive here. His thinking has influenced both human subjects and media ethics and provides a useful touchpoint between the disciplines. Ross's prima facie duties represent firm pre-commitments that resist erosion as a function of rapid cost shifts. However, his list of prima facie duties is also open-ended, and thinkers have added to or subtracted from the list as they have developed his theory. This means that Ross may represent a deontology that is compatible with rapid technological change.

Ross's philosophy combines duty with innovation. As Meyers (2003, 2011) describes, a duty may give way to another that is weighted more heavily, but the duty itself does not go away. Innovation is required to satisfy the original duty, if it is not to be completely subsumed. A few examples follow, one for each foundational principle discussed in this article. This list is not exhaustive. Rather, it is meant to demonstrate that it is better to hold to a principle in the face of particularly rapid technological change and to innovate to meet the goals of the original purpose of the principle than it is to eliminate the principle based on a re-evaluation of its costs.

With respect to autonomy, the size of databases has shifted the nature of informed consent. Most consumers have not meaningfully consented to external research use of their social data. Technology has made it possible to aggregate the data, but researchers may feel as if securing consent from each person who has contributed to the dataset is impractical for cost reasons. How, then, can the ethical principle of autonomy be developed? The cost of obtaining consent should not erode the principle. Yet it is also not possible to make an absolutely inflexible precommitment to traditional informed consent—getting consent from 1.10 billion Facebook users for a Facebook study will not work. What is required is a philosophical framework that combines the flexibility to innovate with a precommitment to the ethical obligation. To do that, the underlying duty must remain intact, such that innovation is focused on new ways of satisfying it. For example, it is relatively simple to obtain active consent from a subset of a social media population who wish to participate in a study, by selecting the sample to study and obtaining their informed consent via a Facebook ad campaign or the like, by ensuring that the website has noncoercive opt-outs, so that users can choose to opt out of research use of their data, or by ensuring at a minimum that the End User License Agreement includes explicit consent for research use. It is too simple to merely focus on the high cost of consent. In doing so, one might erode or eliminate the concept of autonomy, rather than innovate to satisfy it.

The next example focuses on beneficence and nonmaleficence. Here the difficulty is that aggregation tangles the interests of individuals with those of communities. It is hard to maximize benefit and minimize harm when doing so for an individual damages the community or vice versa. Utilitarianism does not provide much help. The "greatest good for the greatest number" conflicts with the direct obligation to not harm the person with whom the researcher is working. Innovation may again help restore the principles of beneficence and nonmaleficence in their new technological context. One innovation that has gained steam in the science context is the use of participant observation methods to ensure that someone has spent enough time with the community to understand the norms that may be inadvertent exposure. Direct, personal engagement with the community can counterbalance the detachment of numbers. The growth of quantitative technique may be counterbalanced by qualitative understanding.

A third example relates to the principle of justice, the fair distribution of the benefits of inquiry. Here again, the technology of big data has altered costs such that it might at first blush appear too costly to maintain foundational principles. In the case of access to research, the challenge is that big data requires big computers and relationships with the companies that create the databases. Equality of benefits and equality of access become costly foundational principles to maintain. A precommitment to basic duties helps maintain the integrity of the principle through innovation. There are indeed numerous innovations on the question of access to data and access to tools. Open access data and open source tools may be directions for the future. Free and open source machine learning toolkits are being developed to provide the tools of big data analysis to everyone. The open access movement has begun to nibble at the edges of academic publishing—it may be that databases as well as published papers one day will be included in open source requirements. In the fair dissemination of the benefits of research, media ethics will play an important part in establishing controls over corporate and academic information hoarding. The benefits of research cannot be shared if they are not disseminated. Here, dissemination of information is the essential element of distributive justice. Media ethics, which has a more developed history of the ethics of information dissemination, should provide a useful check on research ethics, which has focused most heavily on the ethics of knowledge generation.

CONCLUSION

At the same time that journalists have embraced data science techniques, approaches to scientific ethics have had to adapt in the face of rapid technological change. The nature of big data technology fundamentally challenges traditional methods of framing ethical principles. The question is whether the ability to find ever smaller needles in ever larger haystacks has ethical implications for the study of social media by both social scientists and journalists. This article argues that it does.

This technological shift requires some thinking about ethical paradigms. The three elements listed here are database size, data aggregation, and informational harm. These are merely three examples of problems that pose a challenge for traditional approaches to carrying out ethical obligations by shifting costs. Big data technologies increase costs of compliance with traditional ethical values, while steeply lowering costs of invasion of privacy. These shifts can strain cost-based views of ethical behavior. An ethical framework must be flexible enough to accommodate these changes in cost such that core principles are preserved; otherwise, costly ethical precommitments will fall away. The framework must therefore evolve to ensure that these core, essential values are still honored in the face of technological change.

There is common ground here between journalism and science, as there must be if scientific methods and results are to be responsibly used and disseminated in the media. The philosophy of Ross and those who have interpreted his work serves as a touchpoint to start a discourse between the traditions. His duty-based approach avoids erosion of principles in the face of shifting costs. Ross's open-ended paradigm permits technological innovation and expansion. The basic expressions of principles permit ethicists to examine the core principles with a view toward expressing them in a new technological medium.

The function of technology is to shift costs and open possibilities. A serious question for online ethics is whether the first function, cost-shifting, should be allowed to impair the second, the opening of new possibilities. As big data methods permit sifting through data from millions of people, the relationship between researcher and research subjects must change as a function of cost but also must be maintained through the adoption of innovative new possibilities. The choice of ethical framework matters. Given rapid technological change, under any framework the shift in costs of ethical research with regard to one subject will be very different from ethical research of 1 million subjects. The selection of ethical framework matters, because some frameworks may better guide the reintroduction of ethical considerations in the new one-to-many ethical research context.

Here there is a stark difference between cost-based and duty-based approaches. Cost responds directly to cost shifts. If the cost of an action is too high in comparison with its current benefit, the action drops out in a utilitarian framework. On the other hand, duties may be reweighted with shifting costs, but the duties do not themselves become defunct. Duties remain part of the ethical calculus, waiting for the possibility that technological innovation will enable the duty to be more fully realized. The duty may be underwater, but in a duty-based approach it does not dissolve. The duty continues to exert a gravitational pull on the ethical framework, pulling it back into alignment, and requires actors to seek uses of the same technology to fix the ethical problems implicated in its use. Rapid technological flux makes it more necessary than ever for social scientists and journalists who use social science tools and report results to have flexible but firm ethical commitments. Without them, basic ethical obligations will dissolve in the wash of data.

REFERENCES

Bell, E. (2012, September 5). Journalism by numbers. *Columbia Journalism Review*. Retrieved from http://www.cjr.org/cover_story/journalism_by_numbers.php?page=all boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication, & Society*, *15*(5), 662–679.

Cesari, J. (Ed.). (2010). *Muslims in the West after 9/11: Religion, politics and law*. New York, NY: Routledge.

Chen, H., Reid, E., Sinai, J., Silke, A., & Ganor, B. (Eds.). (2008). *Terrorism informatics: Knowledge management and data mining for homeland security*. New York, NY: Springer.

Childress, J., Meslin, E., & Shapiro, H. (2005). *Belmont revisited: Ethical principles for research with human subjects*. Washington, DC: Georgetown University Press.

Christians, C. (2007). Utilitarianism in media ethics and its discontents. *Journal of Mass Media Ethics, 22*(2&3), 113–131.

Hirsch, D. (2006). Protecting the inner environment: What privacy regulation can learn from environmental law. *Georgia Law Review, 41*, 1.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805.

Metz, C. (2013, February 4). Meet the data brains behind the rise of Facebook. *Wired*. Retrieved from http://www.wired.com/wiredenterprise/2013/02/facebook-data-team/

Meyers, C. (2003). Appreciating W.D. Ross: On duties and consequences. *Journal of Mass Media Ethics, 18*(2), 81–97.

Meyers, C. (2011). *Re*appreciating W.D. Ross: Naturalizing prima facie duties and a proposed method. *Journal of Mass Media Ethics, 26*, 316–331.

Montijoye, Y. de, Hidalgo, C., Verleysen, M., & Blondel, V. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*, 1–5.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*, *U.S. Department of Health and Human Services*. Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*, 119.

Office of the Secretary of the Department of Health and Human Services. (2011). Human subjects research protections: Enhancing protections for research subjects and reducing burden, delay, and ambiguity for investigators. *Federal Register, 76*(44), 512.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Belknap Press of Harvard University Press.

Reid, E. (1996). Informed consent in the study of on-line communities: A reflection on the effects of computer-mediated social research. *The Informational Society Journal, 12*, 169.

Ross, W. D. (1930/2002). *The right and the good*. Oxford, England: Oxford University Press.

Schulte, F., Donald, D., & Witkin, G. (2012, September 15). Cracking the codes: How doctors and hospitals have collected billions in questionable Medicare fees. The Center for Public Integrity. Retrieved from http://www.publicintegrity.org/2012/09/15/10810/how-doctors-and-hospitals-have-collected-billions-questionable-medicare-fees

Scott, D., & Bracetti, A. (2013, February 22). 50 things you didn't know about Google. *Complex*. Retrieved from http://www.complex.com/tech/2013/02/50-things-you-didnt-know-about-google/20-petabytes

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

Szego, M., Buchanan, J., & Scherer, S. (2013). Building trust in 21st century genomics. *G3*, *3*(8), 1209–1211.

Vance, A. (2012, August 23). Facebook's is bigger than yours. *Bloomberg Businessweek.* Retrieved from http://www.businessweek.com/articles/2012-08-23/facebooks-is-bigger-than-yours

Williams, E. (2012, April 10). Don't white people kill each other, too?. *The Root*. Retrieved from http://www.theroot.com/views/why-don-t-we-talk-about-white-white-crime