

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 Montjoye, 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 Montjoye 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.

行善 (beneficence) 和不伤害 (nonmaleficence)

难点在于聚合数据将个体利益与社区利益交织在一起

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.

研究者与社区成员花费足够的时间在一起，以理解可能导致无意暴露的规范。与社区的直接、个人接触可以抵消数字化方法带来的疏离感。定量技术的增长可以通过定性理解加以平衡

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.

信息传播是分配正义的基本要素

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.

50 FAIRFIELD AND SHTEIN

在任何框架下，针对一个研究对象的伦理研究成本变化将与针对100万个研究对象的成本变化截然不同

等待技术创新的可能性来使责任得以更充分地实现

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). Reappreciating 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>