Beyond IRBs: Ethical Guidelines for Data Research

Omer Tene and Jules Polonetsky [*]

The ethical framework applying to human subject research in the biomedical and behavioral research fields dates back to the Belmont Report.[1] Drafted in 1976 and adopted by the United States government in1991 as the Common Rule,[2] the Belmont principles were geared towards a paradigmatic controlled scientific experiment with a limited population of human subjects interacting directly with researchers and manifesting their informed consent. These days, researchers in academic institutions as well as private sector businesses not subject to the Common Rule, conduct analysis of a wide array of data sources, from massive commercial or government databases to individual tweets or Facebook postings publicly available online, with little or no opportunity to directly engage human subjects to obtain their consent or even inform them of research activities. The challenge of fitting the round peg of data-focused research into the square hole of existing ethical and legal frameworks will determine whether society can reap the tremendous opportunities hidden in the data exhaust of governments and cities, health care institutions and schools, social networks and search engines, while at the same time protecting privacy, fairness, equality and the integrity of the scientific process. One commentator called this "the biggest civil rights issue of our time."[3]

These difficulties afflict the application of the Belmont Principles to even the academic research that is directly governed by the Common Rule. In many cases, the scoping definitions of the Common Rule are strained by new data-focused research paradigms. For starters, it is not clear whether research of large datasets collected from public or semi-public sources even constitutes human subject research. "Human subject" is defined in the Common Rule as "a living individual about whom an investigator (whether professional or student) conducting research obtains (1) data through intervention or interaction with the individual, or (2) identifiable private information."[4] Yet, data driven research often leaves little or no footprint on individual subjects ("intervention or interaction"), such as in the case of automated

---

[*] Jules Polonetsky is Executive Director and Omer Tene Senior Fellow at the Future of Privacy Forum.
We would like to thank Joe Jerome and Kelsey Finch for their help.

[1] NATIONAL COMM'N FOR THE PROT. OF HUMAN SUBJECTS OF BIOMEDICAL AND BEHAVIORAL RESEARCH, BELMONT REPORT: ETHICAL PRINCIPLES AND GUIDELINES FOR THE PROTECTION OF HUMAN SUBJECTS OF RESEARCH (1979), *available at* http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html.

[2] HHS, FEDERAL POLICY FOR THE PROTECTION OF HUMAN SUBJECTS ('COMMON RULE'), http://www.hhs.gov/ohrp/humansubjects/commonrule/.

[3] Alistair Croll, *Big data is our generation's civil rights issue, and we don't know it*, O'Reilly Radar, Aug. 2, 2012, http://radar.oreilly.com/2012/08/big-data-is-our-generations-civil-rights-issue-and-we-dont-know-it.html.

[4] 45 CFR 46.102(f).

testing for security flaws.[5] As Michael Zimmer notes in his paper for this symposium, "the perception of a human subject becomes diluted through increased technological mediation."[6] Arvind Narayanan and Bendet Zevenbergen explain that "the Internet is more properly understood as a sociotechnical system in which humans and technology interact."[7] Moreover, the existence—or inexistence—of identifiable private information in a dataset has become a source of great contention, with de-identification "hawks" lamenting the demise of effective anonymization[8] even as de-identification "doves" herald it as effective risk mitigation.[9]

Not only the definitional contours of the Common Rule but also the Belmont principles themselves require reexamination. The first principle, *respect for persons*, is focused on individual autonomy and its derivative application, informed consent. While obtaining individuals' informed consent may be feasible in a controlled research setting involving a well-defined group of individuals, such as a clinical trial, it is untenable for researchers experimenting on a database that contains the footprints of millions, or indeed billions, of data subjects. The second principle, *beneficence*, requires a delicate balance of risks and benefits to not only respect individuals' decisions and protect them from harm but also to secure their well-being. Difficult to deploy even in traditional research settings, such cost-benefit analysis is daunting in a data research environment where benefits could be probabilistic and incremental and the definition of harm subject to constant wrangling between minimalists who reduce privacy to pecuniary terms and maximalists who view any collection of data as a dignitary infringement.[10]

In response to these developments, the Department of Homeland Security commissioned a series of workshops in 2011-2012, leading to the publication of the *Menlo Report on Ethical Principles Guiding Information and Communication*

---

[5] *See, e.g.,* Arvind Narayanan & Bendert Zevenbergen, *No Encore for Encore? Ethical Questions for Web-Based Censorship Measurement.*

[6] Michael Zimmer, *Research Ethics in the Big Data Era: Addressing Conceptual Gaps for Researchers and IRBs.*

[7] Narayanan & Zevenbergen, *supra* note 5.

[8] *See, e.g.*, Arvind Narayanan & Ed Felten, *No Silver Bullet: De-Identification Still Doesn't Work*, July 9, 2014, http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf; Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization* , 57 UCLA L. REV. 1701 (2010).

[9] See, e.g., Daniel Barth-Jones, *The Antidote for "Anecdata": A Little Science Can Separate Data Privacy Facts from Folklore*, Nov. 21, 2014, https://blogs.law.harvard.edu/infolaw/2014/11/21/the-antidote-for-anecdata-a-little-science-can-separate-data-privacy-facts-from-folklore/; Kathleen Benitez & Bradley K. Malin, *Evaluating Re-Identification Risks With Respect to the HIPAA Privacy Rule*, 17 J. AMER. MED INFORMATICS ASSOC. 169 (2010); Khaled El Emam et al, *A Systematic Review of Re-Identification Attacks on Health Data*, 6 *PLoS One* 1, December 2011; *also see* Jules Polonetsky, Omer Tene and Kelsey Finch, *Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification* (on file with authors).

[10] Case C-362/14, Maximillian Schrems v. Data Protection Commissioner, 6 October 2015, http://curia.europa.eu/juris/document/document.jsf?docid=169195&doclang=EN; *also see* Ryan Calo, *The Boundaries of Privacy Harm*, 86 IND. L.J. 1131 (2011).

Technology Research.[11] That report remains anchored in the Belmont Principles, which it interprets to adapt them to the domain of computer science and network engineering, in addition to introducing a fourth principle, *respect for law and public interest*, to reflect the "expansive and evolving yet often varied and discordant, legal controls relevant for communication privacy and information assurance."[12] In addition, on September 8, 2015, the U.S. Department of Health and Human Services and 15 other federal agencies sought public comments to proposed revisions to the Common Rule.[13] The revisions, which address various changes in the ecosystem, include simplification of informed consent notices and exclusion of online surveys and research of publicly available information as long as individual human subjects cannot be identified or harmed.[14]

For federally funded human subject research, the responsibility for evaluating whether a research project comports with the ethical framework lies with Institutional Review Boards (IRBs). Yet, one of the defining features of the data economy is that research is increasingly taking place outside of universities and traditional academic settings. With information becoming the raw material for production of products and services, more organizations are exposed to and closely examining vast amounts of often personal data about citizens, consumers, patients and employees. This includes not only companies in industries ranging from technology and education to financial services and healthcare, but also non-profit entities, which seek to advance societal causes, and even political campaigns.[15]

Whether the proposed revisions to the Common Rule address some of the new concerns or exacerbate them is hotly debated. But whatever the final scope of the rule, it seems clear that while raising challenging ethical questions, a broad swath of academic research will remain neither covered by the rules nor subject to IRB review. Katie Shilton shows that academic researchers today have inconsistent views about how to handle these issues.[16] Currently, gatekeepers for ethical decisions range from private IRBs to journal publication standards, association guidelines and peer review. A key question for further debate is whether there is a need for new principles as well as new structures for review of academic research that is not covered by the current or expanded version of the Common Rule.

---

[11] DAVID DITTRICH & ERIN KENNEALLY, THE MENLO REPORT: ETHICAL PRINCIPLES GUIDING INFORMATION AND COMMUNICATION TECHNOLOGY RESEARCH, U.S. Dept. of Homeland Sec., (Aug. 2012), *available at* https://www.predict.org/%5CPortals%5C0%5CDocuments%5CMenlo-Report.pdf.

[12] *Ibid*, at 5.

[13] HHS, NPRM for Revisions to the Common Rule, Sept. 8 , 2015, http://www.hhs.gov/ohrp/humansubjects/regulations/nprmhome.html.

[14] *Also see* Association of Internet Researchers, Ethical Decision-Making and Internet Research Recommendations from the AoIR Ethics Working Committee (Version 2.0), 2012, http://aoir.org/reports/ethics2.pdf (original version from 2002: http://aoir.org/reports/ethics.pdf).

[15] Ira S. Rubinstein, *Voter Privacy in the Age of Big Data*, 2014 WISC. L. REV. 861. .

[16] Katie Shilton, *Emerging Ethics Norms in Social Media Research*.

In *Beyond the Common Rule: Ethical Structures for Data Research in Non-Academic Settings*, we noted that even research initiatives that are not governed by the existing ethical framework should be subject to clear principles and guidelines. Whether or not a research project is federally funded seems an arbitrary trigger for ethical review. Urs Gasser et al note "[the] larger trend of big data research conducted outside of traditional oversight mechanisms due to the limited scope of research subject to existing regulations."[17] To be sure, privacy and data protection laws provide an underlying framework governing commercial uses of data with boundaries like consent and avoidance of harms. But in many cases where informed consent is not feasible and where data uses create both benefits and risks, legal boundaries are more ambiguous and rest on vague concepts such as "unfairness"[18] or the "legitimate interests of the controller."[19] This uncertain regulatory terrain could jeopardize the value of important research that could be perceived as ethically tainted or become hidden from the public domain to prevent scrutiny.[20] Concerns over data ethics could diminish collaboration between researchers and private sector entities, restrict funding opportunities, and lock research projects in corporate coffers contributing to the development of new products without furthering generalizable knowledge.[21]

In a piece he wrote for a *Stanford Law Review Online* symposium we organized two years ago,[22] Ryan Calo foresaw the establishment of "Consumer Subject Review Boards" to address ethical questions about corporate data research.[23] Calo suggested that organizations should "take a page from biomedical and behavioral science" and create small committees with diverse expertise that could operate according to predetermined principles for ethical use of data. The idea resonated in the White House legislative initiative, the Consumer Privacy Bill of Rights Act of 2015, which requires the establishment of a Privacy Review Board to vet non-

---

[17] Urs Gasser, Alexandra Wood, David R. O'Brien, Effy Vayena, and Micah Altman, *Towards a New Ethical and Regulatory Framework for Big Data Research*.

[18] FTC Policy Statement on Unfairness, Appended to International Harvester Co., 104 F.T.C. 949, 1070 (1984). *See* 15 U.S.C. § 45(n).

[19] Article 29 Working Party, WP 217, Op. 06/2014 on the Notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, Apr. 9, 2014, http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

[20] The Common Rule's definition of "research" is "a systematic investigation, including research development, testing, and evaluation, designed to develop or contribute to *generalizable knowledge*." (Emphasis added).

[21] Jules Polonetsky, Omer Tene, & Joseph Jerome, *Beyond the Common Rule: Ethical Structures for Data Research in Non-Academic Settings*, 13 COLO. TECH. L. J. 333 (2015).

[22] Stan. L. Rev. Online Symposium Issue, *Privacy and Big Data: Making Ends Meet*, September, 2013, http://www.stanfordlawreview.org/online/privacy-and-big-data; *also see* stage setting piece, Jules Polonetsky & Omer Tene, *Privacy and Big Data: Making Ends Meet*, September 3, 2013 66 STAN. L. REV. ONLINE 25.

[23] Ryan Calo, *Consumer Subject Review Boards: A Thought Experiment*, 66 STAN. L. REV. ONLINE 97 (2013), *available at* http://www.stanfordlawreview.org/online/privacy-and-big-data/consumer-subject-review-boards.

contextual data uses.[24] In Europe, the European Data Protection Supervisor has recently announced the creation of an Advisory Group to explore the relationships between human rights, technology, markets and business models from an ethical perspective, with particular attention to the implications for the rights to privacy and data protection in the digital environment.[25]

Alas, special challenges hinder the adaptation of existing ethical frameworks, which are strained even in their traditional scope of federally funded academic research, to the fast-paced world of corporate research. For example, the categorical non-appealable decision making of an academic IRB, which is staffed by tenured professors to ensure independence, will be difficult to reproduce in a corporate setting. Yet as Curtis Naser points out in his piece, any institution whose power falls short of an "IRB sledgehammer" becomes merely advisory.[26]

■ ■ ■

To set the stage for possible adoption of IRB-like structures by corporate or non-profit entities, which are currently outside the ambit of the Common Rule, we suggest posing five wh-questions:

What would be subject to review?

Even after its impending expansion, the Common Rule will likely remain incompatible with research that is not federally funded or does not constitute "big R research", contributing to and advancing generalizable knowledge. As discussed above, it is important to extend some form of ethical review process to address such activities. At the same time, it is clear that IRBs cannot be charged with second guessing every operational business decision.

Which corporate research projects should become subject to ethical review? When does business analytics or A/B testing become "human subject research"? Should different rules apply to the same examination of the same dataset simply because the researchers has a different affiliations or motives? Such disparate treatment could risk regulatory arbitrage—academics "laundering" research through corporations or non-profits to escape the strictures of academic IRBs—or incentivize researchers to withdraw knowledge from the public sphere.

One solution would be to separate review of research projects geared toward publication from that of analytics intended for product development and improvement. Unfortunately, the line is not always clear. For example, would a

[24] CONSUMER PRIVACY BILL OF RIGHTS §103(c) (Administration Discussion Draft 2015), *available at* https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf.
[25] European Data Protection Supervisor, Ethics Advisory Group, Dec. 3, 2015, https://secure.edps.europa.eu/EDPSWEB/edps/site/mySite/Ethics.
[26] Curtis Naser, *The IRB Sledge-Hammer, Freedom and Big-Data*.

project become research if a company publishes its results on its own website or a case study in a marketing document? And if a product or service is designed to improve health or advance a technology with broad societal implications, should ethical permissions differ depending on whether the results of an experiment are confidential or published? Some leading companies, large and small, are advancing ethical review models already, but can such models be formalized to have legal consequence or be feasible for start-ups and diverse business models?

Who would review?

Should an internal corporate organ or an external body be charged with conducting an independent ethical review process? A private IRB would necessarily provide external stakeholders with less transparency about corporate processes than an external reviewer. Critics and consumer advocates may not view an internal review board as trustworthy or independent. Consequently, an internal review will necessitate mechanisms to ensure accountability, such as detailed documentation requirements and perhaps regulatory oversight and enforcement *ex post*. In addition, for it to be a meaningful gatekeeper, the composition and structure of an internal review board would have to be regulated. Moreover, a private review process would not contribute toward the creation of industry wide ethical standards and best practices.

At the same time, it would be difficult for organizations to hand over a high volume of strictly confidential business decisions, possibly exposing intellectual property, trade secrets and their pipeline of innovative projects, to an external decision making body. In addition, an external review board would lack the ability – or capacity – for ongoing monitoring of an organization's activities over time. Furthermore, with ethical decisions being made in a virtual vacuum, specific decisions may not reflect the full spectrum of risks and rewards underpinning an organization's broader operations.

An external review board would be an attractive option for an organization that lacks the resources, ability or expertise to develop methodical internal processes. Such bodies could serve multiple companies in an industry or sector thus solving the problem of small and medium size enterprises that lack scale to create an internal review board. In addition, an industry-wide review board could help develop ethical standards and best practices as well as an institutional memory that benefits the public at large.

Other questions concern the identity of members of a review board. Subject matter experts may have a better grip of the technological and business issues raised by a project but lack ethics expertise. Lawyers and ethics experts may master the legal and ethical framework but lack understanding of technical product detail or business strategy.

When would review be conducted?

When should ethical gatekeepers engage with researchers to assess their project? Garfinkel points out that existing IRB practice requires a research experiment to be designed and approved *before* real world deployment. Narayanan and Zevenbergen discuss the *retrospective* role of conference program committees, which are the arbiters of prestigious computer science research publications in conference proceedings.[27]

On the one hand, as several authors demonstrated, *ex ante* review of a research project enables a board to weigh in at the design stage, ensuring the research is ethically structured. Shilton discusses the consultative nature of review processes as well as the informal influence of peer review. Dennis Hirsch and Jonathan King draw on experience with environmental law, noting "back-end environmental management strengthened compliance by the book but stifled innovation in environmental compliance itself."[28] Early scrutiny would also ensure researchers do not waste valuable time and resources pursuing illegitimate trails. Importantly, as Narayanan and Zevenbergen note, where the putative harm of a project arises from *conducting* the research rather than its publication, a retrospective ethical review in conjunction with submission of the research for publication fails to prevent that harm.

On the other hand, *ex post*, or better yet, *continuous* review, ensures that a project and its data trail are scrutinized at the dissemination stage and potentially when information is repurposed, shared or reused. Gasser et al note that ethical oversight is currently focused on the front end, "directed at reducing risks at the study design and data collection stages and, to a much lesser extent, those that arise in later stages such as dissemination and re-use of data. As advances in big data drive sharing and re-use of data by researchers, more of their activities will be subject to limited or, in some cases, no oversight."[29] Garfinkel explains that the exploratory nature of data research, forming hypotheses only after conducting repeated analyses, simply does not fit an ethical review system that requires the procedures and scientific justification of an experiment to be vetted in advance.

Which principles would apply?

In *Beyond the Common Rule*, we demonstrated that the substantive principles of Belmont and Menlo pair well with fundamental principles of privacy law, including the FTC's unfairness doctrine in the U.S. and the Data Protection Directive's legitimate interest test in the EU. Neil Richards and Woody Hartzog suggest reviewing data research through the prism of trust doctrine, including by imposing on researchers a duty of loyalty. The Information Accountability Foundation

[27] Narayanan & Zevenbergen, *supra* note 5.
[28] Dennis D. Hirsch & Jonathan H. King, *Big Data Sustainability: An Environmental Management Systems Analogy*.
[29] Gasser et al, supra note 17.

presents a detailed framework for corporate ethics that takes into account the expected benefits of an organization's big data inquiry, the array of stakeholders for whom processing may pose risks, and the measures that can be taken to mitigate those risks.[30] Similarly, in a previous White Paper, we suggested a structured process to guide businesses in weighing potential data benefits against privacy risks.[31]

Leading corporations have also contributed to the development of new principles for data research. Merck publishes a set of ethical privacy values that includes respect for individual privacy expectations, building and preserving trust, preventing privacy harms, and compliance with the letter and spirit of privacy and data protection laws around the world.[32] Intel's white paper, *Rethink Privacy 2.0: Fair Information Practice Principles Reinterpreted*, highlights the enduring nature of the fair information practice principles and suggests new approaches to their implementation.[33] While much of this work is geared to address commercial analytics and product development, it could perhaps be replicated and extended to the arena of publishable data research.

Where does the line cross for data centered research?

The need for ethical research rules is not restricted to experimentation involving personally identifiable information. The debate over the attacks reportedly launched in the wild by Carnegie Mellon University researchers against users of Tor demonstrates that even without a focus on—or arguably collection of[34]—personal data, research can have profound implications for individual privacy and safety.[35] The *Black Hat* conference canceled a scheduled presentation of the CMU research apparently due to ethical hurdles. Narayanan and Zevenbergen discuss similar concerns with respect to the *Encore* project, a web-based censorship measurement coopting unsuspecting users into the experiment.[36] On their website, *Encore* researchers state "Our Institutional Review Board (IRB) has declined to formally

---

[30] Information Accountability Foundation, Big Data Assessment Framework and Worksheet, July 6, 2015, http://informationaccountability.org/wp-content/uploads/IAF-Big-Data-Ethics-Initiative-Part-B.pdf.
[31] Jules Polonetsky, Omer Tene & Joseph Jerome, *Benefit-Risk Analysis for Big Data Projects*, Aug. 2014, https://www.ftc.gov/system/files/documents/public_comments/2014/08/00027-92420.pdf.
[32] Merck, 2014 Corporate Responsibility Report, Global Privacy Program, http://www.merckresponsibility.com/ethics-transparency/global-privacy-program/.
[33] Intel, *Rethink Privacy 2.0: Fair Information Practice Principles Reinterpreted,*
[34] The debate about what is or is not personally identifiable information continues, for example, with respect to data points such as an IP address.
[35] Ed Felten, *Why were CERT researchers attacking Tor?*, FREEDOM TO TINKER, July 31, 2014, https://freedom-to-tinker.com/blog/felten/why-were-cert-researchers-attacking-tor.
[36] Sam Burnett & Nick Feamster, *Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests*, ACM SIGCOMM, Aug. 2015, http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p653.pdf.

review Encore because it isn't considered human subjects research."[37] While publishing their piece, the ACM SIGCOMM program committee added a strongly-worded disclaimer, stating it "found the paper controversial because some of the experiments the authors conducted raise ethical concerns" and concluding "The PC endorses neither the use of the experimental techniques this paper describes nor the experiments the authors conducted."[38]

Non-data related ethical concerns are not unique to big R research. As they develop products, companies frequently test and experiment in ways unrelated to the collection and use of personal information. They A/B test products, experiment with new drugs and closely examine the performance of new services. Some of these activities are governed specifically by a range of regulatory agencies handling safety issues, including the Food and Drug Administration, Department of Transportation, Consumer Product Safety Commission, Consumer Financial Protection Bureau and more generally, the FTC. This article focuses specifically on issues related to data-driven research, which is an area where the notion of harm is still hotly debated and both benefit and risk are typically intangible.

We suggest that regardless of whether or not personally identifiable information is used, ethical principles should extend to research affecting individuals. As the field of data ethics develops and grows, policymakers should seek to harmonize the principles and procedures governing academic research, corporate research and corporate product development using personal data, as well as research projects affecting individuals in real ways.

[37] Encore: Measure Web Censorship, https://encore.noise.gatech.edu/faq.html
[38] Burnett & Feamster, *supra* note 36.