Towards a New Ethical and Regulatory Framework for Big Data Research

Effy Vayena, Urs Gasser, Alexandra Wood, David R. O'Brien, and Micah Altman[1]

Prepared for the Future of Privacy Forum Workshop
"Beyond IRBs: Designing Ethical Review Processes for Big Data Research"

December 10, 2015

Vast amounts of data about individuals are increasingly being created by new services such as mobile apps and through methods such as DNA sequencing. These data sources can be quite rich, containing large quantities of fine-grained data points related to human biology, behaviors, and relationships over time. Because they can enable analyses at an unprecedented level of detail, these large-scale data sources hold tremendous potential for scientific inquiry. In addition, the costs of obtaining, storing, and analyzing data from these sources are low and continuing to fall relative to the costs of conducting traditional research studies. For these reasons, big data are leading to rapid developments in research, particularly through the emergence of fields such as computational social science and biomedical big data research. As just one example, public health researchers are supplementing traditional methods of disease outbreak detection with streams of data from social networks, chat rooms, and web search queries.[2] Looking ahead, interest in the research potential of big data is expected to continue to rise as the number of large-scale data sources increases and technological capabilities for big data analysis advance.

We recognize the immense research value of big data and believe new large-scale data sources should be made available so that their full potential can be realized. Given the substantial benefits of big data research, the central question moving forward is not whether big data should be made available, but rather how their value can be captured in a way that respects fundamental principles of ethics and privacy. We

---

[1] We describe contributions to the paper using a standard taxonomy, *see* Liz Allen, Jo Scott, Amy Brand, Marjorie Hlava & Micah Altman, *Credit Where Credit Is Due*, 508 NATURE 312 (2014). EV provided the core formulation of the article's goals and aims. AW led the writing of the original manuscript. All authors (EV, UG, AW, DO, and MA) contributed to conceptualization through additional ideas, and through commentary, review, editing, and revision. This material is based upon work supported by the National Science Foundation under Grant No. 1237235.

This opinion piece summarizes, in part, joint work with other collaborators. *See* Jeffrey P. Kahn, Effy Vayena & Anna C. Mastroianni, *Learning as we go: Lessons from the publication of Facebook's social-computing research*, 111 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 13677 (2014); Micah Altman, Alexandra Wood, David R. O'Brien, Salil Vadhan & Urs Gasser, *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. __ (forthcoming); Salil Vadhan et al., Comments to the Department of Health and Human Services Re: Advance Notice of Proposed Rulemaking: Human Subjects Research Protections: Enhancing Protections for Research Subjects and Reducing Burden, Delay, and Ambiguity for Investigators, Docket No. HHS-OPHS-2011-0005 (Oct. 26, 2011); Effy Vayena, Marcel Salathé, Lawrence C. Madoff & John S. Brownstein, *Ethical Challenges of Big Data in Public Health*, 11 PLOS COMPUTATIONAL BIOLOGY (2015); David R. O'Brien et al., *Integrating Approaches to Privacy Across the Research Lifecycle: When Is Information Purely Public?*, Berkman Center Research Publication 2015-7 (2015).

[2] Effy Vayena et al., *supra* note 1.

argue that big data research presents new risks that the current regulatory framework is ill-suited to address, and we recommend updates to the oversight framework that would help enable the collection, use, and sharing of big data in line with ethical principles, research community norms, and expectations of human subjects. Achieving this balance will be critical to ensuring the trust and support of the public and, ultimately, the continued viability of big data research.

Recent illustrations of oversight issues in big data research

There have recently been a number of high-profile incidents that have illustrated gaps in the oversight of big data research. Most notably, researchers involved in a joint Facebook-Cornell University study generated controversy in 2014 when publishing the results of empirical research involving interventions with Facebook users without their knowledge.[3] The study aimed to observe changes in user behavior and mood in response to variations in the emotionally-charged content viewed via the Facebook platform. These types of interventions almost certainly would have required approval from an institutional review board (IRB) had the research been conducted under a federal grant rather than in a commercial setting. This is just one example of the types of research activities increasingly conducted beyond the reach of traditional oversight due to the limited scope of the regulations in place.

Potential oversight gaps have been discovered not only in study design and data collection but also in data release. In 2008, researchers published findings on a methodology for determining whether data about a specific individual are contained in a database of mixtures of genomic DNA collected from hundreds of people.[4] Although the genomic DNA databases were believed to be sufficiently aggregated so as to pose little risk to individual privacy, researchers were able to show that an individual's participation in a study about a specific medical condition could potentially be confirmed using the released data. The National Institutes of Health revoked public access to two DNA databases as a result of this study, and other organizations maintaining similar databases are following suit. In another study, researchers demonstrated the ability to infer the surnames of individuals in de-identified genomic databases.[5] More generally, privacy is a significant challenge for large-scale datasets, as the quantity of data associated with a given record make it highly likely for it to be unique and therefore identifiable. Techniques for learning about individuals in a data release are rapidly advancing, enabling new scientific discoveries but also exposing vulnerabilities in many commonly used measures for protecting privacy. These vulnerabilities are calling into question regulatory approaches that permit the public release of aggregated or de-identified data.

Gaps in the scope of research covered by the existing regulatory framework

Human subjects research protection frameworks developed in the late 1970s fail to address many of the oversight challenges in big data research. Broadly speaking, social and behavioral researchers have long

[3] *See* Adam D. I. Kramer, Jamie E. Guillory & Jeffrey T. Hancock, *Experimental Evidence of Massive-scale Emotional Contagion Through Social Networks*, 111 PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 8788 (2014).
[4] *See* Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-density SNP Genotyping Microarrays*, 4 PLOS GENETICS 8 (2008).
[5] *See* Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321 (2013).

argued that the regulatory framework emphasizes practices, such as obtaining informed consent and balancing the benefits of research against the risks of participation, that are out of place in non-clinical research. These gaps are especially pronounced with respect to many types of big data research. For example, when using data originally collected by a third party like Facebook, a researcher has not interacted with the subjects of the data and informed them of the risks associated with their participation. Furthermore, regulations currently emphasize risk mitigation at the study design and data collection stages of the information lifecycle and, to a much lesser extent, those that arise in later stages such as the dissemination and re-use stages. As advances in big data drive increased data sharing and re-use by researchers, more of their activities will be subject to limited or, in some cases, no oversight.

Additionally, definitions found in the federal policy for the protection of human subjects, known as the Common Rule, create gaps in oversight. What qualifies as human subjects research—and therefore triggers applicability of the Common Rule and IRB review—is rather narrowly defined. Its scope is limited to research involving ". . . a living individual about whom an investigator (whether professional or student) conducting research obtains (1) [d]ata through intervention or interaction with the individual, or (2) [i]dentifiable private information."[6] Many types of research conducted today using large-scale datasets do not fall squarely within this definition. For example, research using a pre-existing Facebook dataset arguably falls outside the scope of this definition because it does not involve an intervention or interaction between the researcher and the subjects of the research.

The second part of this definition likely excludes from oversight some research associated with non-minimal risk of harm. For instance, it permits a researcher who conducts secondary analysis using a de-identified dataset to apply for an exemption from IRB review. However, de-identification alone does not prevent all disclosures, minimize all privacy risks to subjects, or protect information in the manner most subjects would expect. A research dataset that has been de-identified can, in many cases, be re-identified easily.[7] Numerous attacks on de-identified datasets have demonstrated that it is often possible to identify individuals in data that have been stripped of direct and indirect identifiers.[8] It has been shown more generally that very few pieces of information can be used to uniquely identify an individual in a released set of data.[9] As illustrated by the genomic DNA database examples provided above, data stripped of identifiers or released in aggregate form can nevertheless be associated with significant privacy risks. Alternatives to traditional de-identification techniques, such as privacy-aware methods for producing contingency tables, synthetic data, data visualizations, interactive mechanisms, and multiparty computations can provide strong guarantees of privacy while also largely preserving the utility of the data.[10] However, rather than promoting the use of more robust approaches such as these, the

[6] 45 C.F.R. § 46.102(f) (2014).
[7] *See* Arvind Narayanan & Edward W. Felten, *No silver bullet: De-identification still doesn't work* (July 9, 2014).
[8] *See id.*
[9] *See* Yves-Alexandre de Montjoye et al., *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 SCIENCE 536 (2015); Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, 3 NATURE SCI. REP. 1376 (2013).
[10] Many of these advanced methods are also compatible with a strong, formal guarantee of privacy known as differential privacy. *See* Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMMUNICATIONS OF THE ACM 86 (2011).

Common Rule arguably encourages the wide use and sharing of data that have been de-identified using heuristic techniques and released in forms that may be associated with significant privacy risks.

The second part of the definition also exempts research using information considered to be non-private. However, the distinction between public and private information is the subject of significant debate.[11] Sensitive information is increasingly captured in big data sources scraped from the web, or observed via sensors in public spaces, and used for research, often with little or no oversight. Although the protections of the Common Rule apply only to research using personal information subjects have a reasonable expectation will not be made public, many individuals have mismatched expectations regarding secondary uses of information deemed to be public. Consequently, some commentators argue that IRBs and investigators should bear more of the burden of protecting information rather than rely on broad assumptions regarding research subjects' expectations. Compare, for instance, the approach taken by the Common Rule to that found in the European Union, where many categories of information are protected as personal data regardless of their public nature. In response to this debate, research communities have developed ethical guidelines, and the Secretary's Advisory Committee on Human Research Protections has developed draft guidance on the use of data collected from Internet sources. These resources aim to address many of the challenges associated with determining whether information collected online qualifies as public or private under the existing regulations.[12] However, what is considered to fall within these definitions is open to interpretation and will likely evolve over time, and further guidance on interpreting such standards and incorporating them into IRB policies is needed.[13]

Another sizable subset of big data research not subject to the Common Rule is research funded exclusively using private dollars. The sharp distinction between publicly and privately funded research appears to be largely arbitrary, as many privately-funded research activities carry the same types of risks as federally-funded research. In fact, identical studies conducted by two different organizations, one privately-funded and another publicly-funded, could be subject to markedly different requirements. Note, however, that institutional policies are evolving to partially address this gap. For example, many IRB policies cover certain research projects that are not federally funded, and journal policies in many cases require authors to undergo a formal ethical review before publication. A privately funded researcher may also come under the federal rules if she collaborates with a federally funded researcher. Furthermore, laws at the state level may impose additional requirements for human subjects research review that partially fill this gap.

In addition to the various regulations and policies that apply to different classes of researchers within the United States, the regulations of foreign jurisdictions may also apply if any of the collaborating researchers or research subjects are located outside the United States. In fact, many big data research initiatives are international in nature, and protections vary substantially depending on the national data protection regulation that applies. This can lead to mismatches between the safeguards used and the

[11] For a discussion of the evolving notion of public information, see David R. O'Brien et al., *supra* note 1.
[12] Secretary's Advisory Committee on Human Research Protections (SACHRP), Considerations and Recommendations of Concerning Internet Research and Human Subjects Research Regulations, with Revisions (March 2013).
[13] *See* David R. O'Brien et al., *supra* note 1.

expectations and understanding of individual participants, as they may expect that the regulations of their home country will protect their personal data when in fact the requirements of another jurisdiction could be followed once their data cross a border.[14] The difference in treatment that results from the application of regulatory requirements and varying expectations of privacy across jurisdictions creates challenges for researchers, particularly in the secondary analysis of data, as the location of each of the subjects might be uncertain. These factors contribute to uncertainty regarding which standards apply to a specific research activity as well as inconsistency in oversight.

The inadequacy of informed consent requirements

Informed consent is a cornerstone of human subjects research protection. However, an approach based solely on notice and consent has many known weaknesses. Consent forms and terms of service are lengthy, complex, and difficult to understand. Disclosures often do not inform subjects of all potential data uses and harms that could result from misuse of their personal information. In addition, subjects generally have limited opportunities to modify, retract, or withhold consent. These concerns are heightened in big data research, which is often characterized by a substantial separation between the researcher and the research subject. For example, mobile and social networking platforms often embed notice about data collection and sharing practices, including the potential research uses of data collected through the platform, in lengthy terms of service, and individuals impliedly consent to sharing their data under such terms through their use of the service. Because these details are often buried in lengthy terms of service, users are likely unaware that they are participating in human subjects research through their use of a mobile or social networking platform. More generally, the reliance on terms of service that are often vague, complex, and subject to modification without notice, leaves users with an incomplete understanding of how their personal information will be used and shared by the service. These practices arguably fall short of the informed consent requirements intended by research ethics and regulatory frameworks that were developed primarily for the clinical research context, in which extensive recruitment and informed consent processes have been established. If the research oversight framework is to be expanded to provide coverage for new categories of big data research, protections beyond the consent practices currently in place will be likely be necessary.

Recommendations for a new ethical and regulatory framework for big data research

A robust oversight framework is essential to safeguarding the interests of research subjects; ensuring trust, transparency, and accountability in the research community; and maintaining continued support for, funding of, and participation in research studies. As demonstrated by the oversight gaps discussed above, changes to the existing framework are needed to continue to advance these values in big data research. At the core of these recommendations is a recognition of the human right to participation in the production of scientific knowledge.[15] A component of the human right to science, this right refers to the obligations of

[14] For a discussion of many of the issues that may arise when collecting, using, and sharing research data about human subjects across multiple jurisdictions, see David R. O'Brien et al., *supra* note 1.

[15] For a discussion of the human right to science and its application to the regulation of citizen science, see Effy Vayena & John Tasioulas, *"We the Scientists:" A Human Right to Citizen Science*, 28 PHILOSOPHY AND TECHNOLOGY 479 (2015).

governments and other actors to protect and promote participation in science across all stages of the research lifecycle. An intervention designed to protect human subjects therefore should not prevent people who are willing to participate in a study from doing so and thereby impede the capacity of big data research to yield insights into human biology and behavior. Below, we provide a set of objectives and substantive components to consider as part of a new ethics and regulatory framework guided by these design goals. As we describe each objective of the proposed framework, we also sketch example ways in which they could be met through changes to regulations, the policies of review boards, the recommendations of an expert body, or best practice guidance.

Universal coverage. Oversight should cover the full scope of human subjects research. Changes to the existing framework are needed to address gaps in coverage for research involving many categories of information deemed to be public or non-identifiable, privately funded research, and research activities across all stages of the lifecycle including the storage, processing, analysis, release, and post-release stages.[16] Moving towards the model adopted by several European countries, in which regulations cover all research activities regardless of the institution or source of funding, is a potential way to address this gap. To limit the burden on IRBs as a result of an expanded scope of coverage, some responsibilities could be shared with emerging review bodies such as consumer review boards, participant-led review boards, and personal data cooperatives. In addition, for research subject to IRB review, regulators should consider adopting new exemptions to full review that are based in part on the risk-benefit determination described below, as well as explore emerging technological solutions for automating decisions. In addition, changes to the Common Rule could direct IRBs to implement a limited review process for all research at the proposal stage, followed by regular monitoring throughout the research lifecycle to identify research activities for which additional review is needed.

Conceptual clarity. Revised definitions and standards for privacy protection, as well as guidance on interpreting these definitions and applying appropriate safeguards, would likely help IRBs and investigators provide adequate and consistent protection for human subjects. As discussed above, the Common Rule's definition of human subjects research, particularly its reliance on a sharp binary determination based on the presence of "identifiable private information," leads to inconsistency and uncertainty in practice. To provide clarity, the regulations should establish definitions for terms such as privacy, confidentiality, security, and sensitivity, and the terminology should be used consistently. In particular, these definitions should capture important distinctions between terms such as privacy and security. Changes to the Common Rule could direct investigators to implement a combination of both security and privacy controls, where security controls can be viewed as restricting access to information and privacy controls as limiting the potential for harm once access to information is granted. The revised definitions should also be developed based on a modern understanding of privacy that is not limited to a strictly binary conception of "identifiability" or "public availability." For instance, they should cover more broadly the potential for others to learn about individuals based on the inclusion of their information in a set of data, as well as establish a privacy goal against which a technique for privacy protection can be evaluated. Regulators and review boards should consider consulting with ethics and privacy experts, or

[16] For an example framework for analyzing threats, harms, and vulnerabilities in data collection, storage, and release activities, and aligning them with appropriate interventions at each stage of the information lifecycle, see Micah Altman et al., *supra* note 1.

even establish a regularly-convening advisory committee to provide concrete recommendations, as they formulate and implement clarifying definitions, practices, methodologies, and guidelines for implementation. In particular, this expert body could be involved in developing detailed guidance, which IRBs and other review boards can reference as they incorporate revised concepts into their review processes and educational materials for researchers and subjects. Regulators should also consider establishing a clearinghouse of review board policies and decisions that would enable these bodies to learn from one another and achieve greater consistency in the application of standards for human subjects protection.

Risk-benefit assessments. Researchers and review boards should be encouraged to incorporate systematic risk-benefit assessments.[17] Such assessments should evaluate the benefits that would accrue to society as a result of a research activity, the intended uses of the information, the privacy threats and vulnerabilities associated with the research activity, and the potential harms to human subjects as a result of the inclusion of their information in the data. Results from this assessment can be used to guide the selection of protections that are appropriate given the specific benefits, intended uses, and threats and vulnerabilities associated with the research activity. Regulators, in consultation with data privacy experts, could develop detailed guidance to help review boards and researchers systematically examine the privacy threats and vulnerabilities at each stage of the information lifecycle, drawing from concepts found in the technical literature on data privacy and information security.[18] An expert body could, for example, be involved in the development of guidance on modeling the specific uses, threats, and vulnerabilities associated with a research activity. This guidance could also help researchers and review boards select privacy controls that are aligned with these factors. For example, it could provide instructions on evaluating the learning potential from personal data involved in a research activity, as well as the expected harm from uncontrolled use of the data, and matching these characteristics to privacy and security controls that are suitable for addressing these risks. Review boards could use this general guidance as a basis for developing more detailed materials that are specific to their institutional contexts. As the nature of the benefits and risks changes over time, assessments should also evolve, and therefore regulators should consider meeting regularly with an expert body to update the guidance materials that are developed.

New procedural and technological solutions. Researchers should be incentivized to select from the wide range of procedural, economic, legal, educational, and technical protections that are available, rather than rely on a narrow subset of controls such as consent and de-identification. Adoption of techniques from the full scope of available controls could be encouraged through revisions to the Common Rule to incorporate language requiring researchers to consider implementing suitable procedural, economic, legal, educational, or technical safeguards at each stage of the information lifecycle. In addition, regulatory language referring to consent and de-identification controls could be amended to acknowledge that in some cases they should be supplemented by additional controls including information security controls.

---

[17] For an introduction to the components of such a risk-benefit assessment model, see the framework for a modern privacy analysis proposed in Micah Altman et al., *supra* note 1, and the reference guide for conducting benefit-risk analyses provided in JULES POLONETSKY, OMER TENE & JOSEPH JEROME, BENEFIT-RISK ANALYSIS FOR BIG DATA PROJECTS, Future of Privacy Forum Report (Sept. 2014).

[18] For a framework for analyzing informational harms throughout the information lifecycle, see Micah Altman et al., *supra* note 1.

Regulators should also consider creating a safe harbor for researchers who use strong privacy-preserving techniques.[19] Regulators, in consultation with an expert body of privacy researchers, IRB administrators, and researchers, could compile a list of approved techniques that provide a strong guarantee of privacy protection. Examples of some of the technological controls that should be considered for inclusion in this list include privacy-aware methods for contingency tables, synthetic data, data visualizations, interactive mechanisms, and multiparty computations,[20] and controls other than technological controls should additionally be considered. Revisions to the regulations could require regulators and experts to meet regularly to update the list of approved techniques to reflect technological advances. Regulators should also consider tasking data privacy experts with drafting guidance materials on selecting among and applying specific controls, at each of stage of the information lifecycle, that are suitable based on the comprehensive risk-benefit assessment described above. Guidance materials developed by this expert body, or by review boards, should also summarize new approaches to established controls such as consent, including methods for standardizing privacy policies for ease of understanding and processes for dynamic consent that enable individuals to grant, modify, and revoke fine-grained research permissions over time.

Tailored oversight. No one-size-fits-all solution to privacy exists, and researchers should instead be encouraged to adopt procedures and safeguards that are calibrated to the intended uses of the information collected, the benefits of the research activity, and the threats, vulnerabilities, and harms associated with the activity. One way to tailor oversight is to subject different categories of research activities to oversight by different review boards, including IRBs, consumer review boards, participant-led review boards, or personal data cooperatives. For example, in cases where IRB review is not required by the Common Rule, seeking approval from an appropriate review board could be required by journal editors or institutional policies and recommended more generally as an industry best practice. Oversight can also be tailored at the data sharing stage through tiered access, which permits different levels of access to users with different characteristics and needs. Tiered access enables a data provider to finely match the data sharing mechanism with the risks of sharing such data, including factors related to the structure of the data, the sensitivity of the information and potential harms of disclosure, the level of consent obtained from subjects, the credentials of the intended recipients of the data, and the types of analyses they intend to perform.[21] For example, sharing aggregate data using one of the privacy-aware methods described above, such as statistics in the form of contingency tables generated using methods providing a formal guarantee of privacy, could be a suitable option for making data available to the public. An intermediate level could allow approved researchers with suitable credentials to analyze the data through a protected server after agreeing to the terms of a data use agreement providing the data subjects with additional legal protections from misuse. For full access to raw data, the data could be shared through a monitored data environment, such as a virtual data enclave, under the terms of a data use agreement. Similar mechanisms for aligning safeguards with intended uses can be implemented at other stages in the research lifecycle. For example,

[19] For a proposal for a safe harbor for privacy-preserving techniques under the Common Rule, see Salil Vadhan et al., *supra* note 1.
[20] For an expansive catalog of the privacy and security controls that researchers should be encouraged to consider adopting, see Micah Altman et al., *supra* note 1.
[21] For an example of a systematic framework for matching privacy interventions to the threats, harms, and vulnerabilities in a specific data release case, see Micah Altman et al., *supra* note 1.

data minimization and purpose specification principles, operationalized through computable policies, could be applied at the study design and data collection stages to ensure that only the minimum amount of information is collected from human subjects and that data uses are narrowly circumscribed to that authorized by the subjects. Regulators, in consultation with data privacy experts, could be required to establish guidance on tailoring controls at each stage of the lifecycle, and review boards could be empowered to supplement this guidance with detailed instructions that are specific to their institutional contexts.

Development of a new ethics framework with these components would likely require a multistakeholder process with involvement from researchers, institutional review board administrators, industry representatives, regulators, scholarly journal editors, and research participants. In addition to established principles of human subjects research protection, this multistakeholder group could be guided by the human right to participation in the production of scientific knowledge and seek to harmonize the latter right with other interests such as the right to privacy. One output this group could consider developing is a set of ethical norms based in part on existing best practices for research ethics. A panel of domain experts from fields such as computer science, information security, law, and ethics could be convened to develop recommendations regarding practices, methodologies, and tools that are appropriate in different contexts, which could in turn inform the multistakeholder group's assessment of existing best practices. The set of norms developed by the group might begin as general guidelines but evolve over time into more formal codes of practice.

Interfacing with existing ethics and IRB processes, as well as with emerging oversight processes, such as consumer review boards, participant-led review boards, and personal data cooperatives, would be a key component of this process. The involvement of regulators and institutional review board administrators as stakeholders in this process could evaluate the extent to which the current regulatory system is compatible with big data research, or whether changes to the Common Rule would be required. The multistakeholder group could also assess whether institutional review boards are appropriate as the primary oversight body for big data research. The group may alternatively find that technological solutions can help to automate some decisions traditionally made by IRBs, or that oversight by consumer review boards, participant-led review boards, and personal data cooperatives can provide more universal review of big data research.

Researchers should also be involved in the formulation of the framework, in recognition of the human right to participation in science across the entire lifecycle of research. Researcher input would likely help ensure that the oversight framework does not create new inefficiencies or burdens on the research process. The multistakeholder group would likely benefit from regular meetings to review and update the framework once it is in place, to ensure its flexibility and adaptability to unforeseen technological advancements, emerging study design and analytical techniques, new research questions, evolving privacy and other risks to human subjects, regulatory shifts, and changes in societal expectations of privacy.