Ethics of
Data Science (Big Data)

Ethics of Big Data

**

COUNCIL FOR BIG DATA, ETHICS, AND SOCIETY

Council for Big Data, Ethics, and Society

In collaboration with the National Science Foundation, the Council for Big Data, Ethics, and Society was started in 2014 to provide critical social and cultural perspectives on big data initiatives. The Council brings together researchers from diverse disciplines — from anthropology and philosophy to economics and law — to address issues such as security, privacy, equality, and access in order to help guard against the repetition of known mistakes and inadequate preparation. Through public commentary, events, white papers, and direct engagement with data analytics projects, the Council will develop frameworks to help researchers, practitioners, and the public understand the social, ethical, legal, and policy issues that underpin the big data phenomenon.

The Council is directed by danah boyd, Geoffrey Bowker, Kate Crawford, and Helen Nissenbaum.

Ethics of Big Data

Data&Society

Perspectives on Big Data, Ethics, and Society

Principal Investigators:

danah boyd, Data & Society / Microsoft Research (co-PI)

Kate Crawford, Microsoft Research / New York University (co-PI)

Geoffrey C. Bowker, University of California, Irvine (co-PI)

Helen Nissenbaum, New York University (co-PI)

Council Members:

Alessandro Acquisti,
Heinz College, Carnegie Mellon University

Mark Andrejevic,
Pomona College

Solon Barocas,
Princeton University

Edward Felten,
Princeton University

Alyssa Goodman,
Harvard University

Rachelle Hollander,
National Academy of Engineering

Barbara Koenig,
University of California, San Francisco

Eric Meslin,
Indiana University Center for Bioethics

Arvind Narayanan,
Princeton University

Alondra Nelson,
Columbia University

Paul Ohm,
University of Colorado Law School

Frank Pasquale,
University of Maryland

Seeta Peña Gangadharan,
London School of Economics and Political Science/ Data & Society

Latanya Sweeney,
Harvard University

Sharon Traweek,
University of California at Los Angeles

Matt Zook,
University of Kentucky

Report Prepared By: Jacob Metcalf, Emily F. Keller, and danah boyd

2016

Ethics of Big Data

Data&Society

The Council for Big Data, Ethics, and Society

Policy Recommendations

- **Ensure the Common Rule and other relevant research ethics regulation clearly address regulation of data science.**
  Due to historical quirks in ethics regulation, data science occupies a space between 'research' and 'not research' that leads to substantial confusion and hampers effective and consistent ethics review. Future regulations should directly address this gap.
- **Seek ways to facilitate new approaches to ethics review inside academia and industry.**
  Big data research and industry has the potential to innovate improved models of ethics review, and policy-makers should facilitate this opportunity where possible.
- **Develop mechanisms of ethical assessment tailored to big data research methods and industry practices.**
  There is a notable lack of empirical research measuring potential harms of big data analytics to human subjects.

Ethics of Big Data

Data&Society

The Council for Big Data, Ethics, and Society

Pedagogical Interventions

- **Create and distribute high quality data ethics case studies that address difficulties faced by data scientists and practitioners.**
  There is a notable lack of data science case studies available for public use, particularly in the National Online Ethics Center.
- **Develop and support data science curricula with integrative approaches to ethics education.**
  Science and engineering ethics education works best when integrated across a curriculum, rather than as stand-alone units.
- **Train librarians to achieve and promulgate data science and data ethics literacy.**
  University libraries are increasingly the campus hub of data repositories and instructors of data-sharing best practices.
- **Strengthen ethics-oriented activities within professional associations.**
  Computing and data science organizations can play a significant role in setting norms for research and practice. They should update their ethics codes to reflect the specific challenges of ubiquitous data analytics.

Ethics of Big Data

Data&Society

The Council for Big Data, Ethics, and Society

Develop cultures of ethics engagement in data science/industry and encourage cross-disciplinary networking

- **Engagement requires hybrid spaces.** Advancements in data ethics will require formal and informal spaces where people with wide ranges of expertise can network and collaborate.
- **Build models of internal and external ethics regulation bodies in industry.** Industry that utilizes big data analytics faces unprecedented challenges and requires input from both internal and external bodies.
- **Set standards for responsible cross-sector data sharing.** Sharing data between industry and academics is a particularly fraught endeavor, but can carry significant upside.

Ethics of Big Data

Data&Society

The Council for Big Data, Ethics, and Society

Areas for further research

- Should human data science be regarded as human subjects research?
- What are the quantifiable risks posed by correlative and/or predictive data research?
- Similarly, how should we account for the risk of sharing datasets when we cannot know what auxiliary datasets they will be combined (munged) with in the future? Does the risk differ with public datasets?
- How should data privacy and security scientists approach illicitly gained, publicly-available data?
- How can ethical issues be integrated into core technical research?
- What motivates data scientists and their colleagues in industry to pursue ethics processes?
- What is the proper purview of "research ethics" as a topic in the age of big data?

Research Using Big Data

Example "Big Data" Research Controversies

by danah boyd and Jacob Metcalf / November 10, 2014
Produced for Council for Big Data, Ethics, and Society[1]

Facebook's Newsfeed/Emotional Contagion Study. Facebook's 'Newsfeed' is algorithmically produced, thereby determining what users see from their 'friends' based on a variety of input, most of which is not publicly known. The introduction of this function was itself controversial, and the research experiment that underpinned the emotional contagion study (conducted by both academics and corporate data scientists) was dependent on Facebook's ability to manipulate the Newsfeed. This erupted into a large-scale controversy with issues as varied as the role of an IRB, the framing of the study itself, the collaboration between industry actors and academics, the ethics of manipulating users for research, and much more. Commentary.

Research Using Big Data

Example "Big Data" Research Controversies

by danah boyd and Jacob Metcalf / November 10, 2014
Produced for Council for Big Data, Ethics, and Society[1]

- **Twitter's 'Trending' function.** Twitter's researchers designed the "trending topics" feature to help understand what topics were gaining the most traction on the site. Although an internal tool at first, they were given clearance to turn it into a product with the idea that this type of analytics may be of broader interest. This, in turn, prompted users to game the system in the hopes of shaping the trending topics, which forced the designers to alter their algorithm. Because the deployment of the feature suggests that these are the most popular topics on Twitter, there is constant outrage over the manipulation of the system. At a most basic level, few people realize that the goal was to present second order changes (a.k.a. spikes over slow builds) and so there is constant outrage when a topic that slowly gained traction never hits the Trending Topics even if it is one of the most discussed topics on Twitter. While the mechanism is reasonable from a research or feature perspective, the algorithm produces certain cultural effects that prompt others to be critical of the social impact of the company's research activities. Controversies have emerged when political topics like elections or Ebola do not trend or when trends have prompted cross-cultural conflict. Link

Ethics of Big Data

Human-Subjects Protections and Big Data: Open Questions and Changing Landscapes

by Jake Metcalf / April 22, 2015
Produced for Council for Big Data, Ethics, and Society[1]

"New forms of large-scale data should be included as not human-subjects research if all information is publicly available to anyone (including for purchase), if persons providing or producing the information have no reasonable belief that their private behaviors or interactions are revealed by the data, and if investigators have no interaction or intervention with individuals. Investigators must observe the ethical standards for handling such information that guide research in their fields and in the particular research context." (National Academies Press, 2014: 4)

Ethics of Big Data

Human-Subjects Protections and Big Data: Open Questions and Changing Landscapes

by Jake Metcalf / April 22, 2015
Produced for Council for Big Data, Ethics, and Society[1]

Using consent to mediate risks

Paul Ohm (2013) notes a similar dynamic with regards to big data boosterism and the ease with which supposed benefits can be cited to wave off discussion of risk. Ohm cites the Google Flu Index as an example of the supposed benefits of big data tools trumping the need for careful analysis of privacy risks and transparent engagement with users. Ohm writes:

> *"While Google's users likely would have acquiesced had Google asked them to add 'help avoid pandemics' or 'save lives' to the list of accepted uses, they never had the chance for a public conversation. Instead, the privacy debate was held—if at all—within the walls of Google alone. By breaching the public's trust, Google has expanded researchers' ability to examine our search queries and given them a motive to focus in particular on some of the most sensitive information about us, our medical symptoms."*

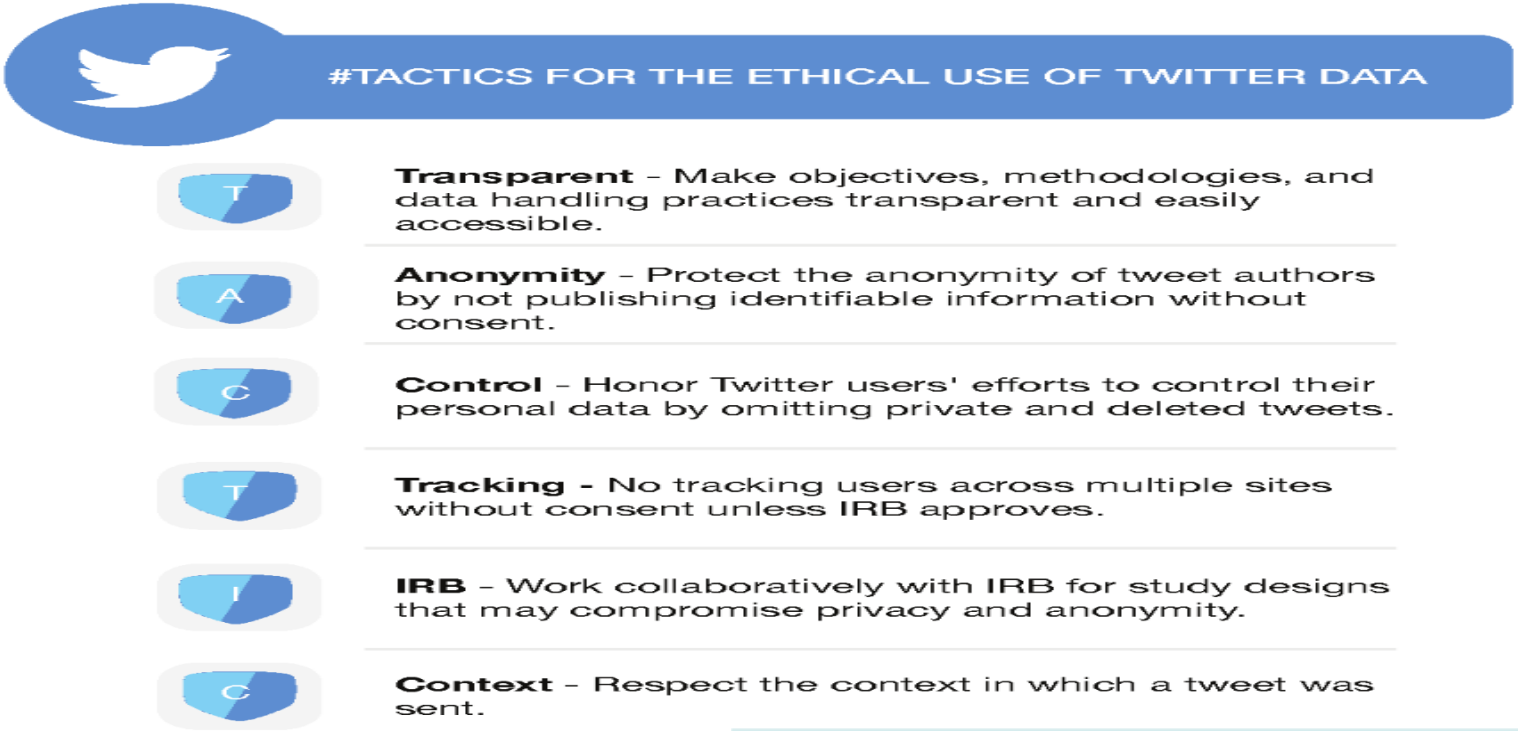
Research using Big Data

REVISED **Ethical research standards in a world of big data [version 2; referees: 3 approved with reservations]**

Caitlin M. Rivers, Bryan L. Lewis

Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech., Blacksburg, VA, 24060, USA

Figure 1. Proposed guidelines for the ethical use of Twitter for research.



F1000Research 2014, 3:38 Last updated: 25 DEC 2016

Research using Big Data

REVISED **Ethical research standards in a world of big data [version 2; referees: 3 approved with reservations]**

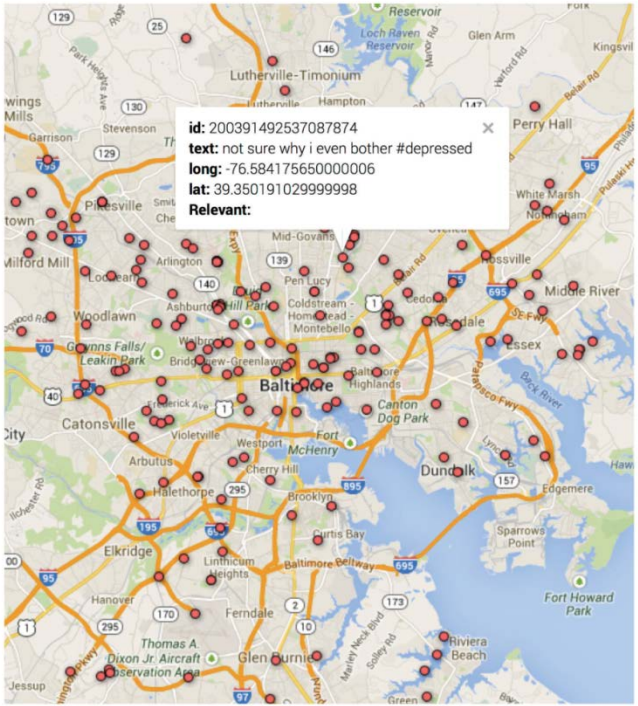
Caitlin M. Rivers, Bryan L. Lewis
Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech., Blacksburg, VA, 24060, USA **

As an example of the potential privacy pitfalls of digital research, suppose investigators were interested in the social networks of adoles
like th
words

According to the US Department of Health and Human Services Policy for Protection of Human Research Subjects, data that are publicly available are exempt from requiring IRB approval[14]. Because Twitter data are public, they technically fall under this exemption. Furthermore, Twitter's privacy policy makes no secret of the fact that user data are indexed by search engines, archived within the US Library of Congress, and are available through an API[15]. However, it is unlikely that many users follow the link to read the lengthy and complex document. One study found that it would take 244 hours a year for an average internet user to read every privacy policy of the unique sites they visit[16].

Tw

the same information were collected through surveys or other traditional means, Institutional Review Board (IRB) approval would be needed.



Each dot is a geolocated tweet collected through the Twitter API. The example tweet displayed is fabricated.

Research using Big Data

REVISED **Ethical research standards in a world of big data [version 2; referees: 3 approved with reservations]**

Caitlin M. Rivers, Bryan L. Lewis
Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech., Blacksburg, VA, 24060, USA

Proposed guidelines for the ethical use of Twitter data

The objectives, methodologies, and data handling practices of the project are transparent and easily accessible

This information should be published in manuscripts, published on the web for the public to access, and provided to IRB (when relevant). Going forward, collaboration between the research community and Twitter to provide information to users about ongoing research and relevant results may also be beneficial. Transparency regarding uses of Internet data for research purposes is needed for fostering 'privacy literacy' so that the users can make informed decisions about participating in Twitter.

Study design and analyses respect the context in which a tweet was sent

Twitter participants can reasonably expect to rely on some anonymity of the crowd to manage privacy. A tweet author discussing his mental health does not do so with the intention of sharing that data with researchers; he does it to communicate with his digital community[25].

Research using Big Data

REVISED **Ethical research standards in a world of big data [version 2; referees: 3 approved with reservations]**

Caitlin M. Rivers, Bryan L. Lewis
Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech., Blacksburg, VA, 24060, USA

The anonymity of tweet authors is protected, ensuring that subjects should not be identifiable in any way

To preserve source anonymity, direct quotes or screen names are not publishable, nor are any details that could be used to identify a subject. Any and all information that could be entered into a search engine to trace back to a human source should be protected. A composite of multiple example tweets may instead be used for illustrative purpose. Geolocations in particular should be scaled to a larger geographic area in order to avoid violating the privacy of those tweet authors. The Title 13 of the Data Protection and Privacy Policy, the federal law under which the Census Bureau is regulated, expressly forbids publishing GPS coordinates[33]; researchers should adhere to this guideline as well.

Research using Big Data

REVISED **Ethical research standards in a world of big data [version 2; referees: 3 approved with reservations]**

Caitlin M. Rivers, Bryan L. Lewis
Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech., Blacksburg, VA, 24060, USA

Tweet data are not used to harvest additional information from other sources

Focused collection is also important for preserving anonymity. It is possible to use data collected from Twitter to discern the identities of tweet authors, which can then be used to find and collect additional information from additional sources. For example an author's username, identifying details provided in tweet texts, or geolocations could all be used to collect data about that individual from other sources like Facebook, LinkedIn, Flickr, or public records. This methodology should not be pursued without consent or IRB approval.

Twitter users' efforts to control their personal data are honored

Researchers may not follow a user on Twitter in order to gain access to a protected account. Doing so would violate that user's efforts to control his or her personal data.

Research using Big Data

REVISED **Ethical research standards in a world of big data [version 2; referees: 3 approved with reservations]**

Caitlin M. Rivers, Bryan L. Lewis
Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech., Blacksburg, VA, 24060, USA

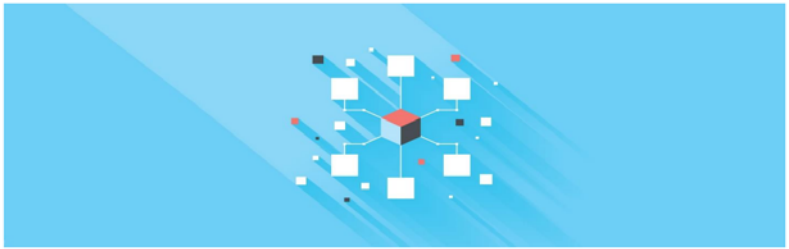
Researchers work collaboratively with IRB just as they would for any other human subject data collection

There is not currently an expectation that researchers engaging in research using Twitter will interface with their IRB. As discussed above, studies that could be conceived as individual-based should require IRB approval, whereas research designs that use data in aggregate (e.g. counts of keywords) may proceed without explicit consent. In turn, review boards should keep abreast of social network mining methodologies and corresponding ethical considerations in order provide informed guidance to researchers.

IRB & Data Research

Beyond IRBs: Designing Ethical Review Processes for Big Data Research

Conference Proceedings



Thursday, December 10, 2015 • Future of Privacy Forum • Washington, DC

FUTURE OF PRIVACY FORUM

WcL LAW

This material is based upon work supported by the National Science Foundation under Grant No. 1547506 and by the Alfred P. Sloan Foundation under Award No. 2015-14138.

Alfred P. Sloan FOUNDATION

NSF

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Alfred P. Sloan Foundation or its trustees, officers, or staff.

Contents

**

IRB and Data Research

Research Ethics in the Big Data Era: Addressing Conceptual Gaps for Researchers and IRBs

Michael Zimmer, PhD School of Information Studies
University of Wisconsin – Milwaukee **

Conceptual Gaps: Privacy, Personally Identifiable Information, Consent, and Harm

When considered through the lens of the regulatory definition of "private information," social media postings are typically considered public, especially when users take no steps to restrict access, and are thus not deserving of particular privacy consideration. For example, researchers in the Harvard "Tastes, Ties, and Time" research project (Lewis, et al, 2008) – where an entire cohort of college students had their Facebook profiles scraped annually for for years – argued that subjects do not have a reasonable expectation of privacy with their Facebook information, noting "We have not accessed any information not otherwise available on Facebook," and equating their collecting of the profile data with "sitting in a public square, observing individuals and taking notes on their behavior" (comment at Zimmer, 2008b). Similarly, much of the discussion surrounding the appropriateness of harvesting Twitter activity centers on the basic fact that a public Twitter stream is purposefully visible to anyone, thus no privacy expectations exist (see, for example, discussions at 2008b).

https://bigdata.fpf.org/papers/research-ethics-in-the-big-data-era-addressing-conceptual-gaps-for-researchers-and-irbs/

IRB and Data Research

Michael Zimmer, PhD School of Information Studies
University of Wisconsin – Milwaukee **

Conceptual Gaps: Privacy, Personally Identifiable Information, Consent, and Harm

Gross, 2006). Thus, it remains unclear whether Internet users truly understand if and when their online activity is regularly monitored and tracked, and what kind of reasonable expectations truly exist. This uncertainty in the intent and expectations of users of social media and internet-based platforms—often fueled by the design of the platforms themselves—create a conceptual gap in our ability to apply the definition of "private information" to ensure subject privacy is properly addressed and forces us to reconsider the justification "we have not accessed any information not otherwise available" in order to alleviate potential privacy concerns.

https://bigdata.fpf.org/papers/research-ethics-in-the-big-data-era-addressing-conceptual-gaps-for-researchers-and-irbs/

Research Ethics in the Big Data Era: Addressing Conceptual Gaps for Researchers and IRBs

IRB and Data Research

Michael Zimmer, PhD School of Information Studies
University of Wisconsin – Milwaukee **

Conceptual Gaps: Privacy, Personally Identifiable Information, Consent, and Harm

Similar conceptual gaps emerge when we consider the traditional definitions of "personally identifiable information" in the context of big data research, where there are increased pressures to release datasets, as well as increased opportunities to access and combine databases from various sources. Increasingly, datasets considered "anonymized" have been re-identified, often with relative ease, relying on information not covered under the regulatory definition of "personally identifiable." For example, researchers have been able to re-identify individuals by analyzing and comparing such datasets, using data-fields as benign as one's ZIP code (Sweeney, 2002), random Web search queries (Barbaro & Zeller Jr, 2006), or movie ratings (Narayanan & Shmatikov, 2009) as the vital key for reidentification of a presumed anonymous user. Prior to widespread Internet-based data collection and processing, few would have considered one's movie ratings or ZIP code as personally identifiable. Yet, merely stripping

https://bigdata.fpf.org/papers/research-ethics-in-the-big-data-era-addressing-conceptual-gaps-for-researchers-and-irbs/

IRB and Data Research

Research Ethics in the Big Data Era: Addressing Conceptual Gaps for Researchers and IRBs

Michael Zimmer, PhD School of Information Studies
University of Wisconsin – Milwaukee **

Conceptual Gaps: Privacy, Personally Identifiable Information, Consent, and Harm

of archiving public Twitter streams centers on whether tweets are public utterances by human subjects, and thus requiring ethical review, or merely the equivalent of published texts, thus exempted from any ethical concern (see discussion at Zimmer 2010b). Similarly, researchers studying large datasets or communication network traffic, for example, frequently perceive their studies as outside the purview of IRBs, since, in their view, the IRB review process is "used more in medical and psychology research at our university" (as quoted in Soghoian, 2012) or perceive IRBs as bothersome barriers to achieving important research outcomes (Garfinkel, 2008).

https://bigdata.fpf.org/papers/research-ethics-in-the-big-data-era-addressing-conceptual-gaps-for-researchers-and-irbs/

IRB and Data Research

Michael Zimmer, PhD School of Information Studies
University of Wisconsin – Milwaukee

Conceptual Gaps: Privacy, Personally Identifiable Information, Consent, and Harm

transaction logs. As a result, the perception of a human subject becomes diluted through increased technological mediation. To compensate, Carpenter and Dittrich (2011) encourage ethical review boards to transition "from an informed consent driven review to a risk analysis review that addresses potential harms stemming from research in which a researcher does not directly interact with the at-risk individuals" (4), and to ultimately "transition our idea of research protection from 'human subjects research' to 'human harming research'" (14). In doing so, researchers who might otherwise (even if incorrectly) feel no human is directly involved in the research study would be compelled to address the ethical implications of any harm to broader populations outside the immediate research project.

https://bigdata.fpf.org/papers/research-ethics-in-the-big-data-era-addressing-conceptual-gaps-for-researchers-and-irbs/

IRB & Data Research

6-7-2016

Beyond IRBs: Ethical Guidelines for Data Research

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 Technology Research*.[11] That report remains anchored in the Belmont Principles, adapting 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 fifteen 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]

Omer Tene & Jules Polonetsky, *Beyond IRBs: Ethical Guidelines for Data Research*, 72 WASH. & LEE L. REV. ONLINE 458 (2016), http://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/7

IRB & Data Research

6-7-2016

Beyond IRBs: Ethical Guidelines for Data Research

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

Omer Tene & Jules Polonetsky, *Beyond IRBs: Ethical Guidelines for Data Research*, 72 WASH. & LEE L. REV. ONLINE 458 (2016), http://scholarlycommons.law.wlu.edu/wlulr-online/vol72/iss3/7

Regulations Related to the Use of Data

Table 1. Consumer Privacy Bill of Rights[32].

Transparency	Consumers have a right to easily understandable and accessible information about privacy and security practices.
Respect for context	Consumers have a right to expect that companies will collect, use, and disclose personal data in ways that are consistent with the context in which consumers provide the data.
Security	Consumers have a right to secure and responsible handling of personal data.
Focused collection	Consumers have a right to reasonable limits on the personal data that companies collect and retain.
Accountability	Consumers have a right to have personal data handled by companies with appropriate measures in place to assure they adhere to the Consumer Privacy Bill of Rights.
Individual control	Consumers have a right to exercise control over what personal data companies collect from them.
Access and accuracy	Consumers have a right to access and correct personal data in usable formats.

Ethical Principles of Big Data

What's Up With Big Data Ethics?

Insights from a business executive and law professor

**

From our perspective, we believe that any organizational conversation about big data ethics should relate to four basic principles that can lead to the establishment of big data norms:

- ***Privacy isn't dead***; it's just another word for information rules. Private doesn't always mean secret. Ensuring privacy of data is a matter of defining and enforcing information rules — not just rules about data collection, but about data use and retention. People should have the ability to manage the flow of their private information across massive, third-party analytical systems.
- ***Shared private information can still remain confidential.*** It's not realistic to think of information as either secret or shared, completely public or completely private. For many reasons, some of them quite good, data (and metadata) is shared or generated by design with services we trust (e.g. address books, pictures, GPS, cell tower, and WiFi location tracking of our cell phones). But just because we share and generate information, it doesn't follow that anything goes, whether we're talking medical data, financial data, address book data, location data, reading data, or anything else.

http://radar.oreilly.com/2014/03/whats-up-with-big-data-ethics.html

Ethical Principles of Big Data

What's Up With Big Data Ethics?

Insights from a business executive and law professor

**

- ***Big data requires transparency.*** Big data is powerful when secondary uses of data sets produce new predictions and inferences. Of course, this leads to data being a business, with people such as data brokers, collecting massive amounts of data about us, often without our knowledge or consent, and shared in ways that we don't want or expect. For big data to work in ethical terms, the data owners (the people whose data we are handling) need to have a transparent view of how our data is being used – or sold.
- ***Big Data can compromise identity.*** Privacy protections aren't enough any more. Big data analytics can compromise identity by allowing institutional surveillance to moderate and even determine who we are before we make up our own minds. We need to begin to think about the kind of big data predictions and inferences that we will allow, and the ones that we should not.

http://radar.oreilly.com/2014/03/whats-up-with-big-data-ethics.html

Principles of Data Ethics

Principles for Data Ethics

Data science professionals and practitioners should strive to perpetuate these principles:

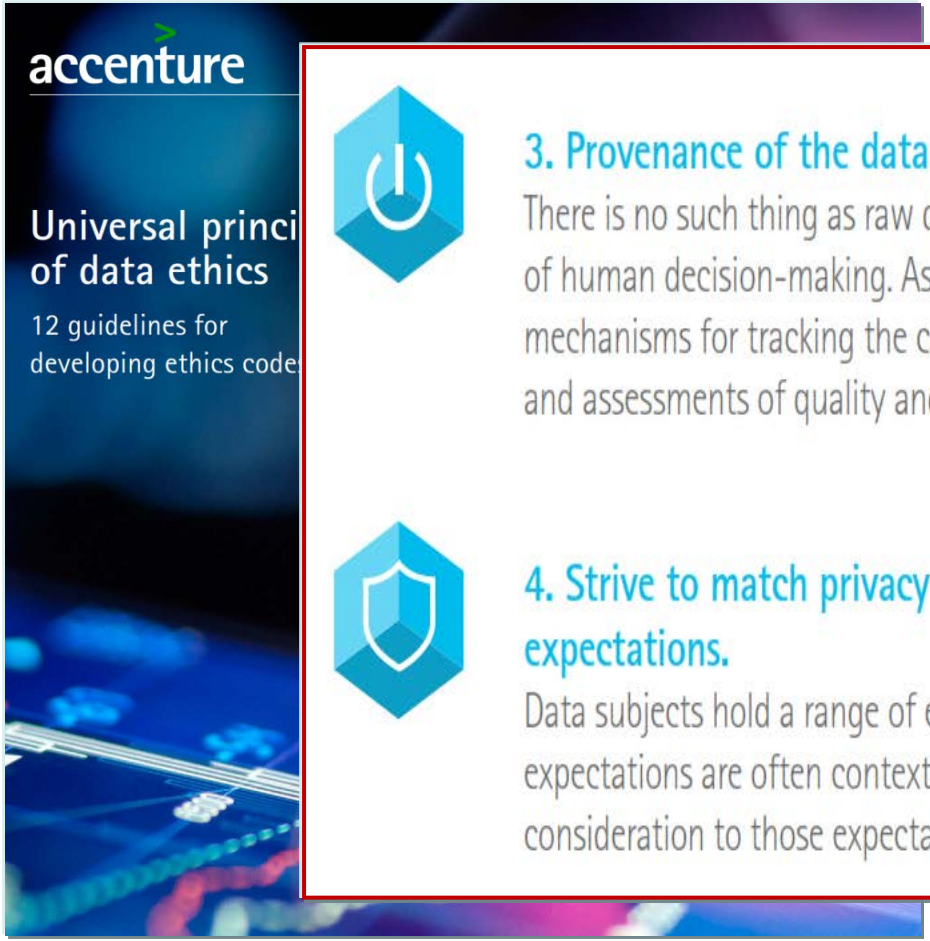
1. The highest priority is to respect the persons behind the data.

When insights derived from data could impact the human condition, the potential harm to individuals and communities should be the paramount consideration. Big data can produce compelling insights about populations, but those same insights can be used to unfairly limit an individual's possibilities.

2. Attend to the downstream uses of datasets.

Data professionals should strive to use data in ways that are consistent with the intentions and understanding of the disclosing party. Many regulations govern datasets on the basis of the status of the data, such as "public," "private" or "proprietary." However, what is *done with* datasets is ultimately more consequential to subjects/users than the type of data or the context in which it is collected. Correlative uses of repurposed data in research and industry represents both the greatest promise and the greatest risk posed by data analytics.

Principles of Data Ethics



3. Provenance of the data and analytical tools shapes the consequences of their use.

There is no such thing as raw data—all datasets and accompanying analytic tools carry a history of human decision-making. As much as possible, that history should be auditable, including mechanisms for tracking the context of collection, methods of consent, the chain of responsibility, and assessments of quality and accuracy of the data.

4. Strive to match privacy and security safeguards with privacy and security expectations.

Data subjects hold a range of expectations about the privacy and security of their data and those expectations are often context-dependent. Designers and data professionals should give due consideration to those expectations and align safeguards and expectations as much as possible.

Principles of Data Ethics

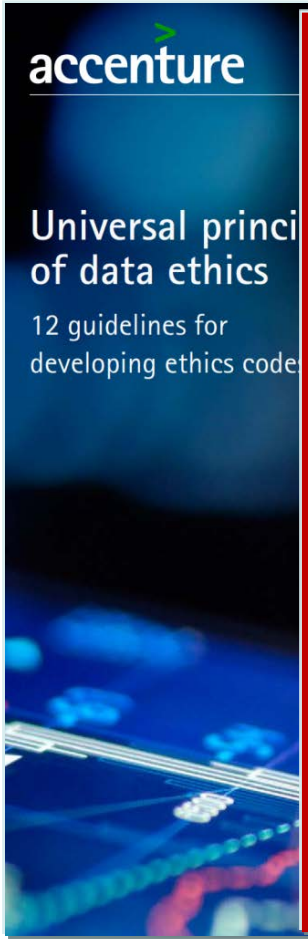
5. Always follow the law, but understand that the law is often a minimum bar.

As digital transformations have become a standard evolutionary path for businesses, governments and laws have largely failed to keep up with the pace of digital innovation and existing regulations are often mis-calibrated to present risks. In this context, compliance means complacency. To excel in data ethics, leaders must define their own compliance frameworks that outperform legislated requirements.

6. Be wary of collecting data just for the sake of more data.

The power and peril of data analytics is that data collected today will be useful for unpredictable purposes in the future. Give due consideration to the possibility that less data may result in both better analysis and less risk.

Principles of Data Ethics



7. Data can be a tool of inclusion and exclusion.

While everyone deserves the social and economic benefits of data, not everyone is equally impacted by the processes of data collection, correlation, and prediction. Data professionals should strive to mitigate the disparate impacts of their products and listen to the concerns of affected communities.

8. As much as possible, explain methods for analysis and marketing to data disclosers.

Maximizing transparency at the point of data collection can minimize more significant risks as data travels through the data supply chain.

9. Data scientists and practitioners should accurately represent their qualifications, limits to their expertise, adhere to professional standards, and strive for peer accountability.

The long-term success of the field depends on public and client trust. Data professionals should develop practices for holding themselves and peers accountable to shared standards.

Principles of Data Ethics

10. Aspire to design practices that incorporate transparency, configurability, accountability, and auditability.

Not all ethical dilemmas have design solutions, but being aware of design practices can break down many of the practical barriers that stand in the way of shared, robust ethical standards. Data ethics is an engineering challenge worthy of the best minds in the field.

11. Products and research practices should be subject to internal, and potentially external ethical review.

Organizations should prioritize establishing consistent, efficient, and actionable ethics review practices for new products, services, and research programs. Internal peer-review practices can mitigate risk, and an external review board can contribute significantly to public trust.

12. Governance practices should be robust, known to all team members and reviewed regularly.

Data ethics poses organizational challenges that cannot be resolved by familiar compliance regimes alone. Because the regulatory, social, and engineering terrains are so unsettled, organizations engaged in data analytics require collaborative, routine and transparent practices for ethical governance.

6 Questions about Big Data

• **How does the organization use Big Data, and to what extent is it integrated into strategic planning?** Clearly identifying the purpose for which data will be used helps to identify critical issues that may arise. How does that particular use benefit the customer or wider public? For data use to benefit your organization and its stakeholders, it has to be accurate and trustworthy. How do you ensure the quality and veracity of your data?

• **Does the organization send a privacy notice when personal data are collected?** s it written in clear and accessible language that allows users to give truly informed consent? For example, social media platforms ask users to agree to terms and conditions when they register. However, research shows this does not necessarily correlate to informed consent as many users do not read through lengthy, complicated documents, but simply sign them to quickly open their accounts.

• **Does the organization assess the risks linked to the specific type of data the organization uses?** Identifying any potential negative effect that the use of data might pose to particular groups, and what might happen if the data became public, is one way of increasing awareness of the damage a potential data breach would cause. In some cases, a privacy impact assessment may be advisable. The risk of misuse of the company's information by employees should not be underestimated.

http://www.journalofaccountancy.com/issues/2016/oct/big-data-ethical-questions.html Journal of Accountancy October 1, 2016

6 Questions about Big Data

- **Does the organization have safeguards to mitigate these risks?** Communicating the established preventive measures to bolster data security is an effective way to promote trust. These might include controls on data access and harsh penalties for its misuse.

- **Does the organization make sure that the tools to manage these risks are effective and measure outcomes?** Audit has a key role to play in helping companies deal with these issues.

- **Does the organization conduct appropriate due diligence when sharing or acquiring data from third parties?** When buying information from third parties, due-diligence procedures must apply as they would to other purchases. Do the suppliers uphold similar ethical standards and guarantee the accountability and transparency of these practices?

http://www.journalofaccountancy.com/issues/2016/oct/big-data-ethical-questions.html Journal of Accountancy October 1, 2016

Discussion ….

Consent issues **outside** of the HIPAA Framework

- Should consent requirements attach to those uses and disclosures that are outside of what should reasonably be expected given the context?
  - Do people understand enough of this context?
- Data that relates to health is an ever-expanding list – poses challenges to placing more stringent requirements on sharing of “health” data (e.g., GIS, Telephone traffic, Population movement,…)