CHANGING PRACTICES IN BIG DATA RESEARCH ETHICS

Katie Shilton
Associate Professor
College of Information Studies
University of Maryland, College Park

Data Cultures

- Data cultures (sciences, social sciences, medicine) where some norms around data collection are well-established
  - Although disrupted and debated
- Emergent data cultures where availability of data about people is with a newer phenomenon
  - Little to no history of dealing with logistics such as data curation, human subjects concerns such as privacy and justice
- What norms will a data culture (an occupation, an industry) adopt?
- Who will be critical voices in developing those norms?
- How will the occupation or industry decide on both ethical and business best practices?

Work to date: Social Computing

Belmont Report & Common Rule

Belmont Report published in 1979 in response to studies in psychology & medicine

- Further articulated in Common Rule legislation

Three main principles:

1. *Respect for persons*: participants know about and consent to research
2. *Beneficence*: do no harm; maximize benefits, minimize risks
3. *Justice*: fair distribution of costs/benefits to *all potential* participants

IRBs and Social Computing Research

- Belmont Principles still viable

- But their *implementation by IRBs* is not always a great fit
  - Long history of mismatch between social science methods & biomedically-focused IRB review

- Social computing brings social science, statistics, and computing into closer relationships

Research Questions

1. What ethical challenges are faced by social computing researchers?
2. What are the research ethics practices of researchers using online datasets?
3. What do researchers using online datasets believe constitutes ethical research?
4. How do these practices and beliefs vary among social computing researchers?

What ethical challenges are faced by social computing researchers?

- Semi-structured interviews with 20 researchers with PhDs in information technology, information systems, information studies, communication, business, and computer science.
- Faculty at US and European academic institutions, or researchers in consulting or industrial research labs.
- All self-identified as using open online datasets for research.

Shilton, K., & Sayles, S. (2016). “We aren’t all going to be on the same page about ethics:” Ethical practices and challenges in research on digital and social media. In *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS 2016)*. Kauai, HI: IEEE.

Challenges presented by social computing research

- Consent and privacy
  - Do people know data is being used for research?
  - Is it feasible to collect informed consent?
  - Can anonymous participants be re-identified?
  - Does research feel intrusive to participants?
- Justice and fairness
  - Lack of visibility into what online data says about individuals
  - Access and accessibility of online participation
- Risk
  - Do we really understand the risk to participants?

Understanding Emerging Norms

- Survey of 263 online data researchers

- Document ***beliefs and practices*** around which social computing researchers are converging, as well as areas of ***ongoing disagreement***.

Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2016)*. San Francisco, CA: ACM.

Method: Identifying population

- Focused on researchers in social computing-related fields where online data analysis is common.

- Used databases to identify articles published since 2011 at eight conferences: CSCW, CHI, ICWSM, iConference, WWW, Ubicomp, CKIM, and KDD.
  - Search terms included: “trace ethnography,” “big data,” “twitter,” “forums,” “text mining,” “logs,” “activity traces,” and “social network”

- Also posted survey invite to relevant listservs, including AoIR, AIS, CITASA, AIS ICA, STS, and NCA

Method: Survey Instrument

Items captured:

- Data sources, methods, and analyses employed
- Data collection, analysis and sharing practices
- Attitudes toward appropriateness of various methods and practices
- Items capturing researchers “personal codes of ethics”
- Demographics

Variable		Mean (SD)/N (%)
Sex	Male	159 (60.5%)
	Female	93 (35.4%)
Education	Bachelor’s	15 (5.7%)
	Master’s	61 (23.2%)
	PhD	180 (68.4%)
Current Location	United States	162 (61.6%)
	UK	21 (8.0%)
	Canada	14 (5.3%)
	Germany	12 (4.6%)
	23 other countries (<10 participants)	47 (17.8%)
Degree In	Communication & Media	33 (12.5%)
	Computer Science/Engineering	100 (38%)
	HCI	10 (3.8%)
	Information	50 (19%)
	Social Sciences	37 (14.1%)
	Other	27 (10.3%)
Current Field of Work	Academia (Research Focus)	207 (78.7%)
	Industry	35 (13.3%)
	Policy/Government/Non-Profit	12 (4.6%)

Sample Demographics (N=263)

Findings: Researcher Code of Ethics

Analyzed two ways:

1. Qualitative, iterative coding by authors of open-ended responses to question, “How would you describe your personal code of ethics regarding online data?”
2. EFA of 35 survey items relating to research attitudes and practices

Code	Definition	Example Statements
Public Data	Only using public data / public data being okay to collect and analyze	*In general, I feel that what is posted online is a matter of public record, though every case needs to be looked at individually in order to evaluate the ethical risks.*
Do No Harm	Comments related to Golden Rule	*Golden rule, do to others what you'd have them do to you.*
Informed Consent	Always get informed consent / stressing importance of informed consent	*I think at this point for any new study I started using online data, I would try to get informed consent when collecting identifiable information (e.g. usernames).*
Greater Good	Data collection should have a social benefit	*The work I do should address larger social challenges, and not just offer incremental improvements for companies to deploy.*
Established Guidelines	Including Belmont Report, IRBs Terms of Service, legal frameworks, community norms	*I generally follow the ethical guidelines for human subjects research as reflected in the Belmont Report and codified in 45.CFR.46 when collecting online data.*
Risks vs. Benefits	Discussion of weighing potential harms and benefits or gains	*I think I focus on potential harm, and all the ethical procedures I put in place work towards minimizing potential harm.*
Protect Participants	data aggregation, deleting PII, anonymizing / obfuscating data	*I aggregate unique cases into larger categories rather than removing them from the data set.*
Data Judgments	Efforts to not make inferences or judge participants or data	*Do not expose users to the outside world by inferring features that they have not personally disclosed.*
Transparency	Contact with participants or methods of informing participants about research	*I prefer to engage individual participants in the data collection process, and to provide them with explicit information about data collection practices.*

Emergent themes from qualitative responses: researchers' personal code of ethics

	Low variance (<.8)	High variance (> 1.2)
Agreement (>3.5)	• Remove subjects from datasets upon request[1] • Ask (1) colleagues or (2) IRBs about research[1] • Share results with participants[1] • Think about edge cases/ outliers[1]	• Use non-representative samples[2] • Remove unique individuals before sharing[1] • Researchers can't collect large-scale online data if consent is required
Disagreement (<3)	No items.	• Ignore ToS when necessary[2] • Deceive participants[2] • Share raw data with key stakeholders[1] • It's possible to obtain informed consent with large-scale studies

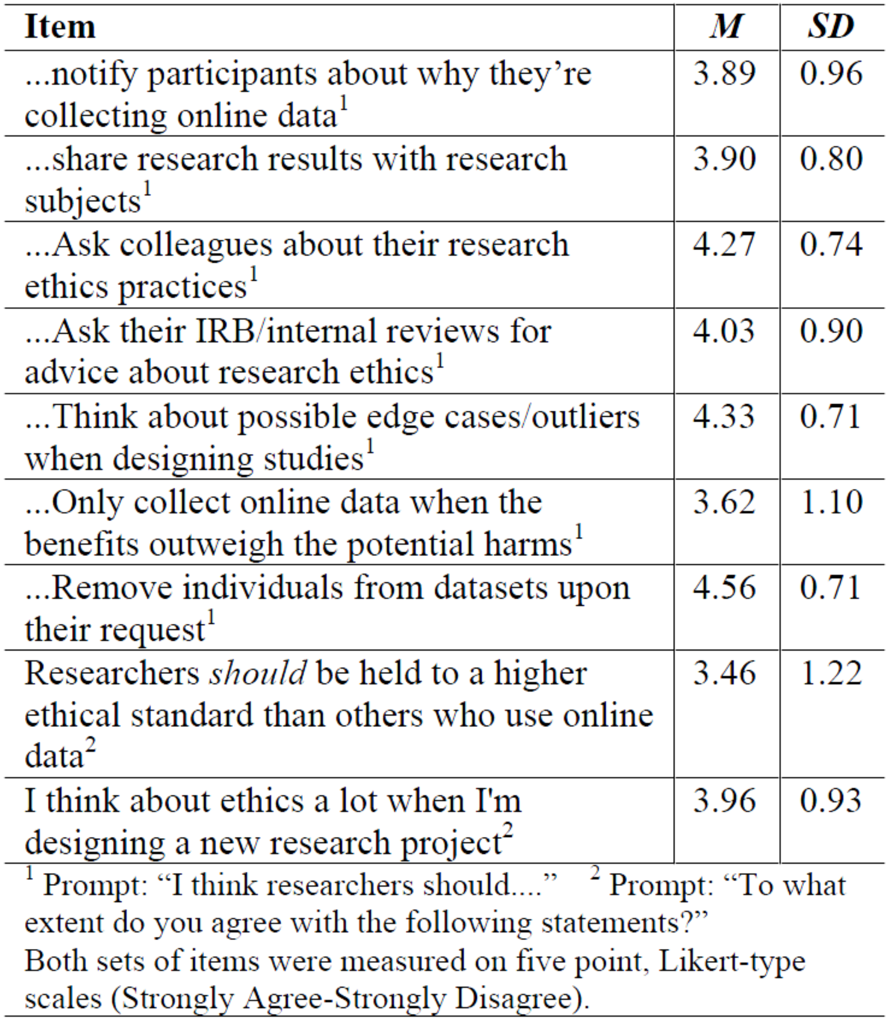
Notes:

[1] "I think researchers should…"

[2] "It's permissible for researchers to…"

Understanding Emerging Norms

Understanding Emerging Norms



Item	*M*	*SD*
...notify participants about why they're collecting online data[1]	3.89	0.96
...share research results with research subjects[1]	3.90	0.80
...Ask colleagues about their research ethics practices[1]	4.27	0.74
...Ask their IRB/internal reviews for advice about research ethics[1]	4.03	0.90
...Think about possible edge cases/outliers when designing studies[1]	4.33	0.71
...Only collect online data when the benefits outweigh the potential harms[1]	3.62	1.10
...Remove individuals from datasets upon their request[1]	4.56	0.71
Researchers *should* be held to a higher ethical standard than others who use online data[2]	3.46	1.22
I think about ethics a lot when I'm designing a new research project[2]	3.96	0.93

[1] Prompt: "I think researchers should...." [2] Prompt: "To what extent do you agree with the following statements?"
Both sets of items were measured on five point, Likert-type scales (Strongly Agree-Strongly Disagree).

Table 5. Codification of Ethical Attitudes Measure

Going Beyond Belmont

1. Transparency with participants
    - Openness about data collection
    - Sharing results with community leaders or research subjects
2. Data minimization
    - Collecting only what you need to answer an RQ
    - Letting individuals opt out
    - Sharing data at aggregate levels
3. Ethical deliberation with colleagues
4. Caution in sharing results
5. Respect the norms of the contexts in which online data was generated.

Emerging Consensus

- No differences in agreement on these practices between CS, IS, and social science scholars
- Qualitative responses indicate both thoughtful, deliberative processes
  - Lots of variation on Common Rule, “do no harm”
  - “*I try to consider whether the research context is a significant departure from the original context the data were published in, before embarking on collection. For this reason, I generally choose not to scrape/crawl public sources.*”
- And significant room for growth
  - *“Flexible”*
  - *“Under construction :S”*
  - *“It is ok to break the rules”*

Comparative Data Culture Cases

- Citizen science
  - Data shared includes info on volunteer location & other sensitive personal information
  - But volunteers do not typically express privacy concerns
  - Overall, citizen science volunteers and practitioners share and promote openness and data sharing over protecting privacy.
    - *Sharing and contribution* rather than *taking*.
  - Citizen science is an example of *contextually-appropriate* data sharing

- Cybersecurity research
  - With Megan Finn at UW, research just beginning

Bowser, A., Shilton, K., Warrick, E., & Preece, J. (2017). Accounting for privacy in citizen science: ethical research in a context of openness. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017)*. Portland, OR: ACM.