CS 384: Ethical and Social Issues in NLP

Dan Jurafsky, Ria Kalluri, Peter Henderson

Stanford University

Spring 2023

Where does the Data come from?

Participants, Data, & labelers in NLP

1. History of Human Subjects Protection

Nuremberg Code of 1947

Ten principles of research developed for the "Doctors' Trial": American judges trying Nazi doctors accused of murder and torture in their human experiments in the concentration camps.

1. The voluntary consent of the human subject is … essential.

2. The experiment should be .... for the good of society

6. …risk ... should never exceed ... the humanitarian importance of the problem

9. ...subject should be at liberty to bring the experiment to an end...

Shuster, Evelyne. 1997. "Fifty years later: the significance of the Nuremberg Code." *New England Journal of Medicine* 337, 20: 1436-1440.

Videos and photos from Holocaust Memorial Museum (trigger warning)

United States Public Health Services Study in Tuskegee

40-year study by the US Public Health Service begun in 1932

Goal: observe natural history of untreated syphilis

Enrolled 600 poor African American sharecropper men

- 400 with syphilis, 200 controls

Told they would be treated for "bad blood"

Were not treated, merely studied

- Were not told they had syphilis
- Sexual partners not informed
- By 1940s penicillin becomes standard treatment for syphilis
  - Subjects were not told or given penicillin

Wikimedia Commons,
from National Archive

United States Public Health Services Study in Tuskegee

1964 Protest letter from a doctor who reads one of the papers

- "I am utterly astounded by the fact that physicians allow patients with a potentially fatal disease to remain untreated when effective therapy is available"

1965 Memo from authors:

- "This is the first letter of this type we have received. I do not plan to answer this letter"

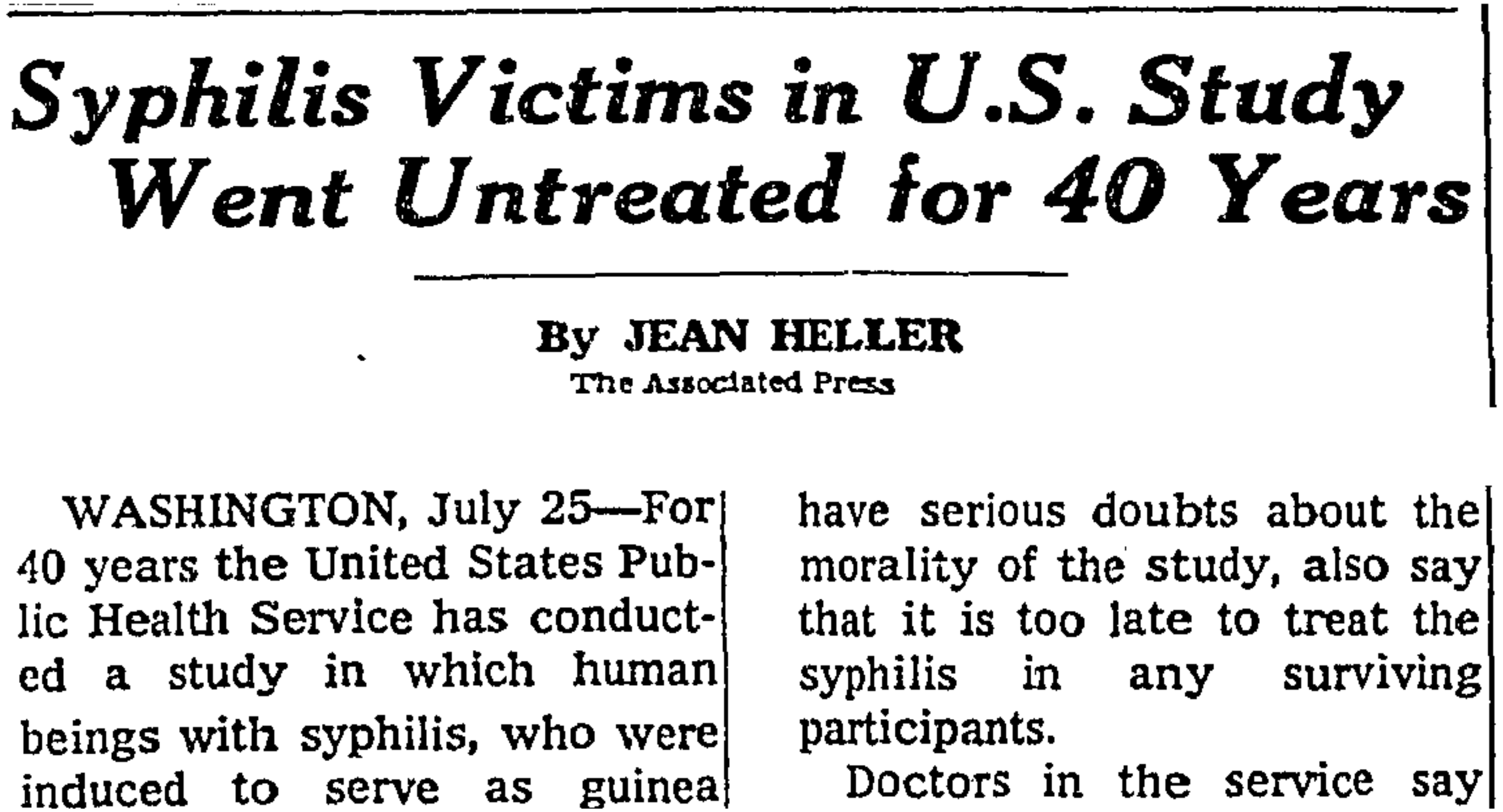
United States Public Health Services Study in Tuskegee

1966 Peter Buxtun, a PHS researcher in San Francisco, sent a letter to the CDC but study was not stopped.

1972 Buxton goes to the press.

Senator Edward Kennedy calls congressional hearings

1974 Congress passes National Research Act



Syphilis Victims in U.S. Study Went Untreated for 40 Years

By JEAN HELLER
The Associated Press

WASHINGTON, July 25—For 40 years the United States Public Health Service has conducted a study in which human beings with syphilis, who were induced to serve as guinea

have serious doubts about the morality of the study, also say that it is too late to treat the syphilis in any surviving participants.

Doctors in the service say

NY Times July 26, 1972

Non-medical experiments

Stanford prison experiment

Conducted by Philip Zimbardo in 1971 (in the basement of Jordan Hall).

Modeled after the "Toyon Hall experiment", a final project of one of the students in his Psychology seminar

Le Texier, Thibault. "Debunking the Stanford Prison Experiment." *American Psychologist* (2019).

Stanford prison experiment

College students were chosen to be either "prisoners" or "guards"

Results as published by Zimbardo:

- Guards humiliated and abused prisoners
- Prisoners became depersonalized
- Evidence for "ugly side of human nature"

Experiment stopped after 6 days

Le Texier, Thibault. "Debunking the Stanford Prison Experiment." *American Psychologist* (2019).

Stanford prison experiment Scientific and ethical flaws

Participants were not random: respondents to an ad for “a psychological study of prison life.”

- Carnahan and MacFarland 2007: word "prison" selects personalities

Guards were told the expected results ("conditions which lead to mob behavior, violence")

Researchers intervened in experiment to instruct guards how to behave ("We can create a sense of frustration. We can create fear")

Guards not told they were participants

Researcher refused to allow prisoner participants to leave experiment.

Le Texier, Thibault. "Debunking the Stanford Prison Experiment." *American Psychologist* (2019).

National Research Act 1974

Required institutional review of all federally funded experiments

- Institutional Review Boards (IRBs)

Created National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research

- Issued Belmont Report in 1976/1979

The Common Rule: Title 45, Part 46 of the Code of Federal Regulations: Protection of Human Subjects.

- Informed consent

2. Current Human Participants Rules

The Belmont Report
Three Basic Ethical Principles

1. Respect for Persons

- Individuals should be treated as autonomous agents
  - "Informed Consent"
- Persons with diminished autonomy are entitled to protection

The Belmont Report
Three Basic Ethical Principles

2. Benificence

- Do no harm
- Maximize possible benefits and minimize possible harms.

The Belmont Report
Three Basic Ethical Principles

3. Justice

Who ought to receive the benefits of research and bear its burdens?

- Fair procedures and outcomes in the selection of research subjects
- Advances should benefit all

The Common Rule

The Federal Policy for the Protection of Human Subjects

45 CFR part 46

https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html

IRB: Institutional Review Board

Internal to each academic institution

- Most universities (including Stanford) have 2 distinct boards
  - Medical and Non-medical
  - https://researchcompliance.stanford.edu/panels/h

Reviews all human subjects experiments

- Consent forms
- Risks/benefits
- Contributions of research
- Protection of privacy

The Common Rule

Human subject: a living individual about whom an investigator (professional or student) conducting research:

(i) Obtains information … through intervention or interaction with the individual, and uses, studies, or analyzes the information …; or

(ii) Obtains, uses, studies, analyzes, or generates identifiable private information ….

Exempt Research

…

Research that only includes … **survey procedures, interview procedures, or observation of public behavior** (including visual or auditory recording) **if** one of::

(i) …the **identity** of the human subjects **cannot** readily **be ascertained**, …

(ii) Any **disclosure** of the human subjects' **responses** outside the research **would not …**place the subjects at risk of criminal or civil liability or **be damaging** to the subjects' financial standing, employability, educational advancement, or reputation; or

(iii) The information obtained is recorded by the investigator in such a manner that the identity of the human subjects can readily be ascertained, and an IRB conducts a limited IRB review …

…

[When] consent is not required: Secondary research uses of identifiable private information … if… the identifiable private information … [is] **publicly available**;

Deceiving participants

Belmont Report:

"incomplete disclosure" is allowed when:

(1) incomplete disclosure is truly necessary to accomplish the goals of the research

(2) there are no undisclosed risks to subjects that are more than minimal, and

(3) there is an adequate plan for debriefing subjects, when appropriate, and for dissemination of research results to them

CITI training

If you intend to be on any research project that runs human subjects

You must do CITI certification

- Required by Stanford IRB
- Required for all federally funded research
- Short course
- https://researchcompliance.stanford.edu/panels/hs/forms/training/citi

3. What about data from corpora?

Authors

Using social media data: author

From IRB perspective this kind of corpus data is exempt if it is public

- E.g., public twitter data

But are there still questions?

Issues with social media data: author

Williams, M. L., Burnap, P., Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.

"Are consent, confidentiality and anonymity required where the research is conducted in a public place where people would reasonably expect to be observed by strangers?"

What counts as a public vs. private space on/off the web?

- If people are whispering in a public square is that private?
- What about religious ceremonies?

Issues with social media data: author

Williams, M. L., Burnap, P., Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.

What are the potential harms?

- Demographic info (age, ethnicity, religion, sexual orientation)
- Associations (membership in groups or associations with particular people)
- Communications that are person or potentially harmful (extreme options? Illegal activities?)
- Others?

What do Twitter authors think?

Casey Fiesler and Nicholas Proferes. 2018. “Participant” Perceptions of Twitter Research Ethics. Social Media + Society, 4(1). 22

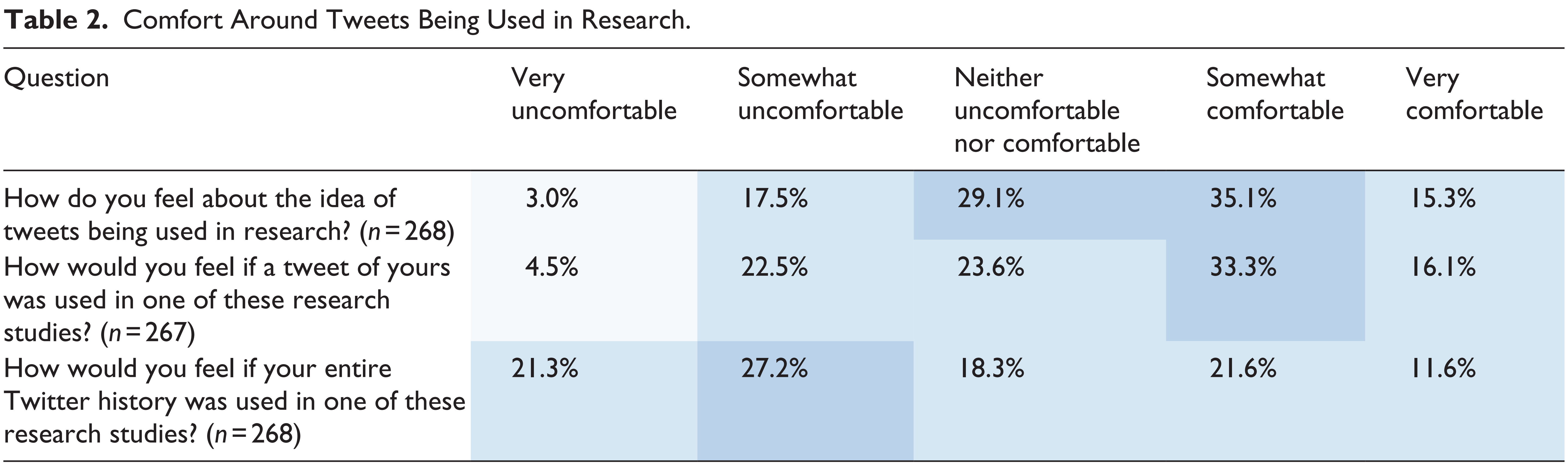


Table 2. Comfort Around Tweets Being Used in Research.

Question	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
How do you feel about the idea of tweets being used in research? (*n* = 268)	3.0%	17.5%	29.1%	35.1%	15.3%
How would you feel if a tweet of yours was used in one of these research studies? (*n* = 267)	4.5%	22.5%	23.6%	33.3%	16.1%
How would you feel if your entire Twitter history was used in one of these research studies? (*n* = 268)	21.3%	27.2%	18.3%	21.6%	11.6%

Fiesler et al

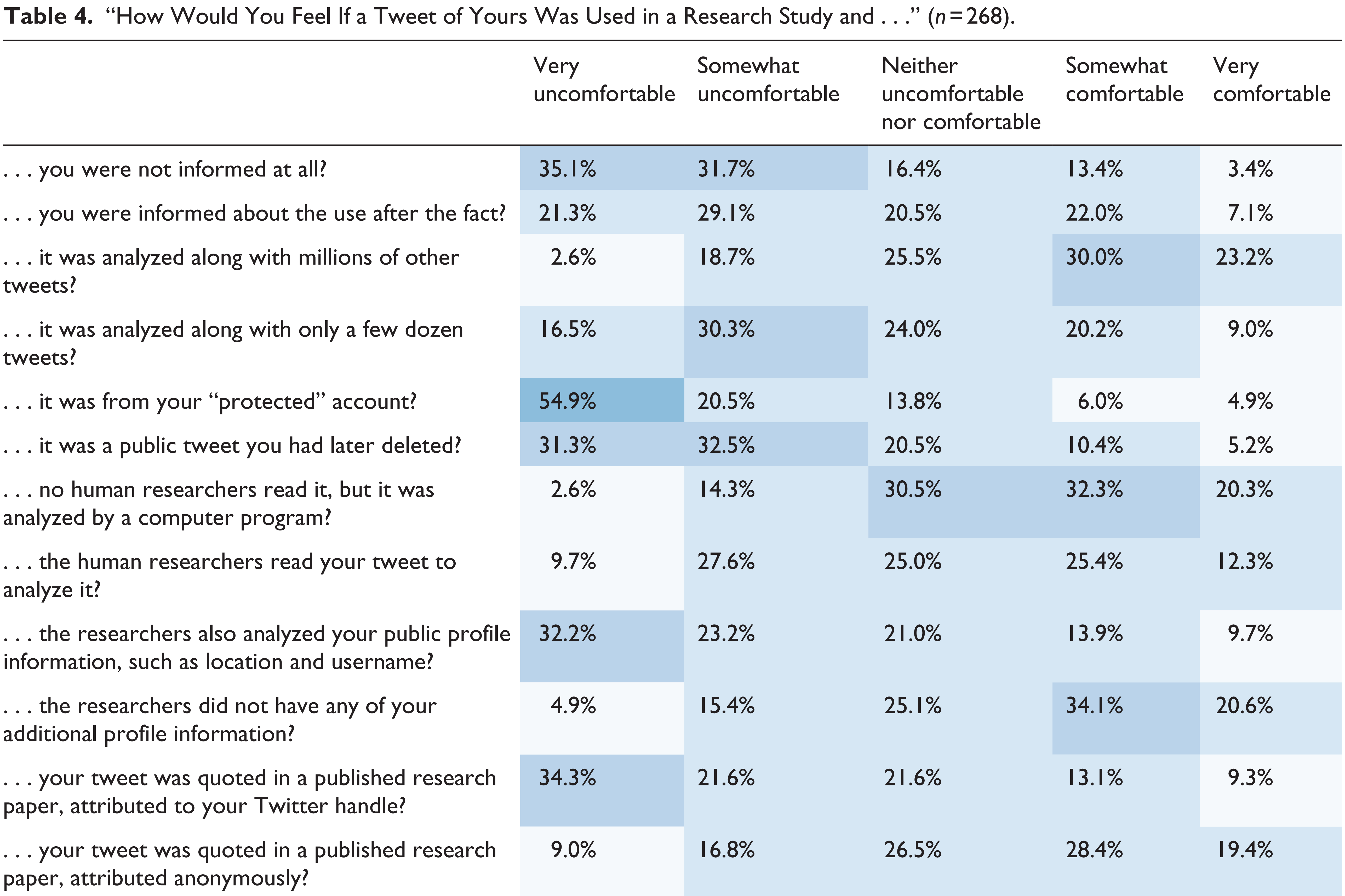
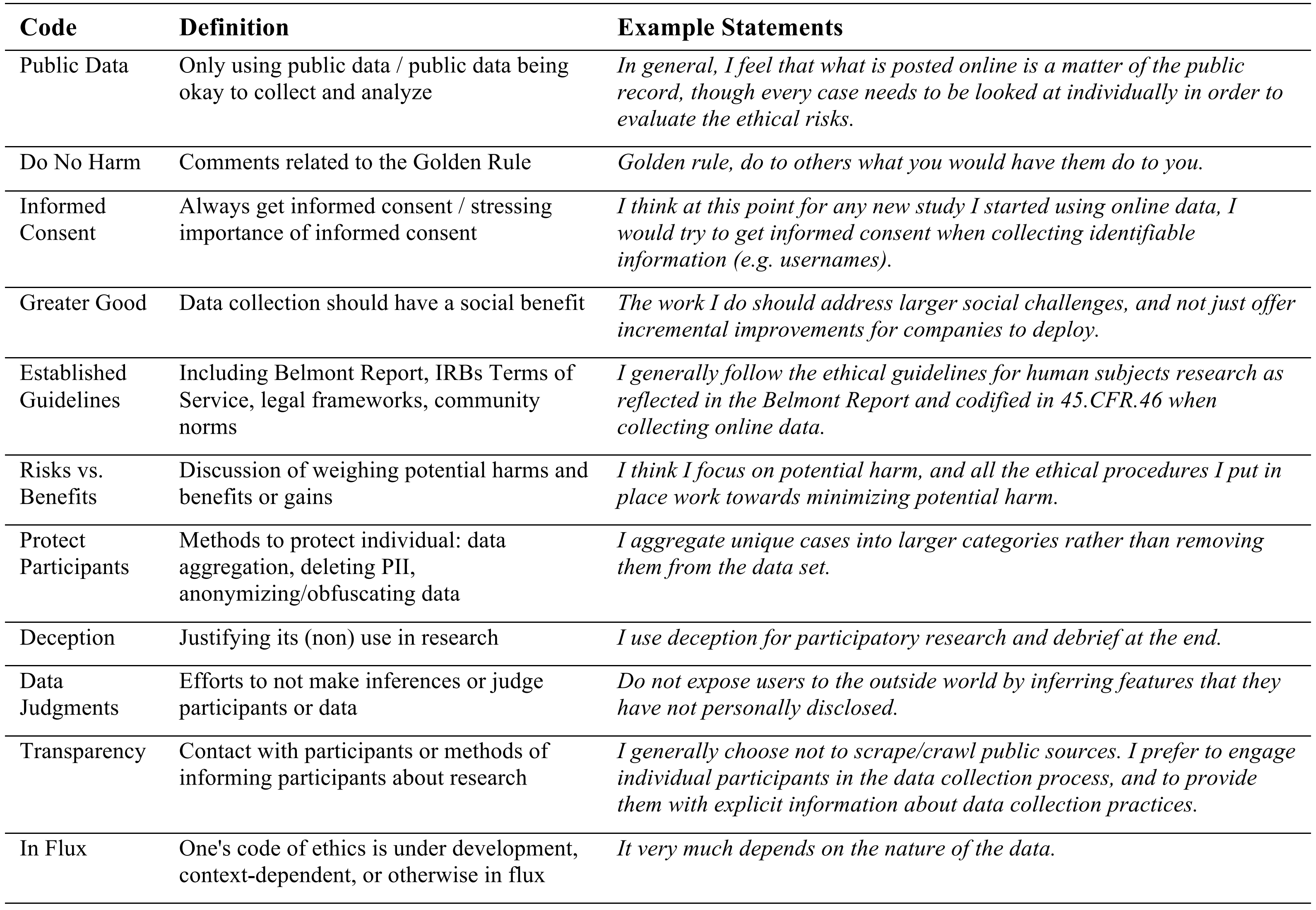


Table 4. "How Would You Feel If a Tweet of Yours Was Used in a Research Study and . . ." (*n* = 268).

	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
. . . you were not informed at all?	35.1%	31.7%	16.4%	13.4%	3.4%
. . . you were informed about the use after the fact?	21.3%	29.1%	20.5%	22.0%	7.1%
. . . it was analyzed along with millions of other tweets?	2.6%	18.7%	25.5%	30.0%	23.2%
. . . it was analyzed along with only a few dozen tweets?	16.5%	30.3%	24.0%	20.2%	9.0%
. . . it was from your "protected" account?	54.9%	20.5%	13.8%	6.0%	4.9%
. . . it was a public tweet you had later deleted?	31.3%	32.5%	20.5%	10.4%	5.2%
. . . no human researchers read it, but it was analyzed by a computer program?	2.6%	14.3%	30.5%	32.3%	20.3%
. . . the human researchers read your tweet to analyze it?	9.7%	27.6%	25.0%	25.4%	12.3%
. . . the researchers also analyzed your public profile information, such as location and username?	32.2%	23.2%	21.0%	13.9%	9.7%
. . . the researchers did not have any of your additional profile information?	4.9%	15.4%	25.1%	34.1%	20.6%
. . . your tweet was quoted in a published research paper, attributed to your Twitter handle?	34.3%	21.6%	21.6%	13.1%	9.3%
. . . your tweet was quoted in a published research paper, attributed anonymously?	9.0%	16.8%	26.5%	28.4%	19.4%

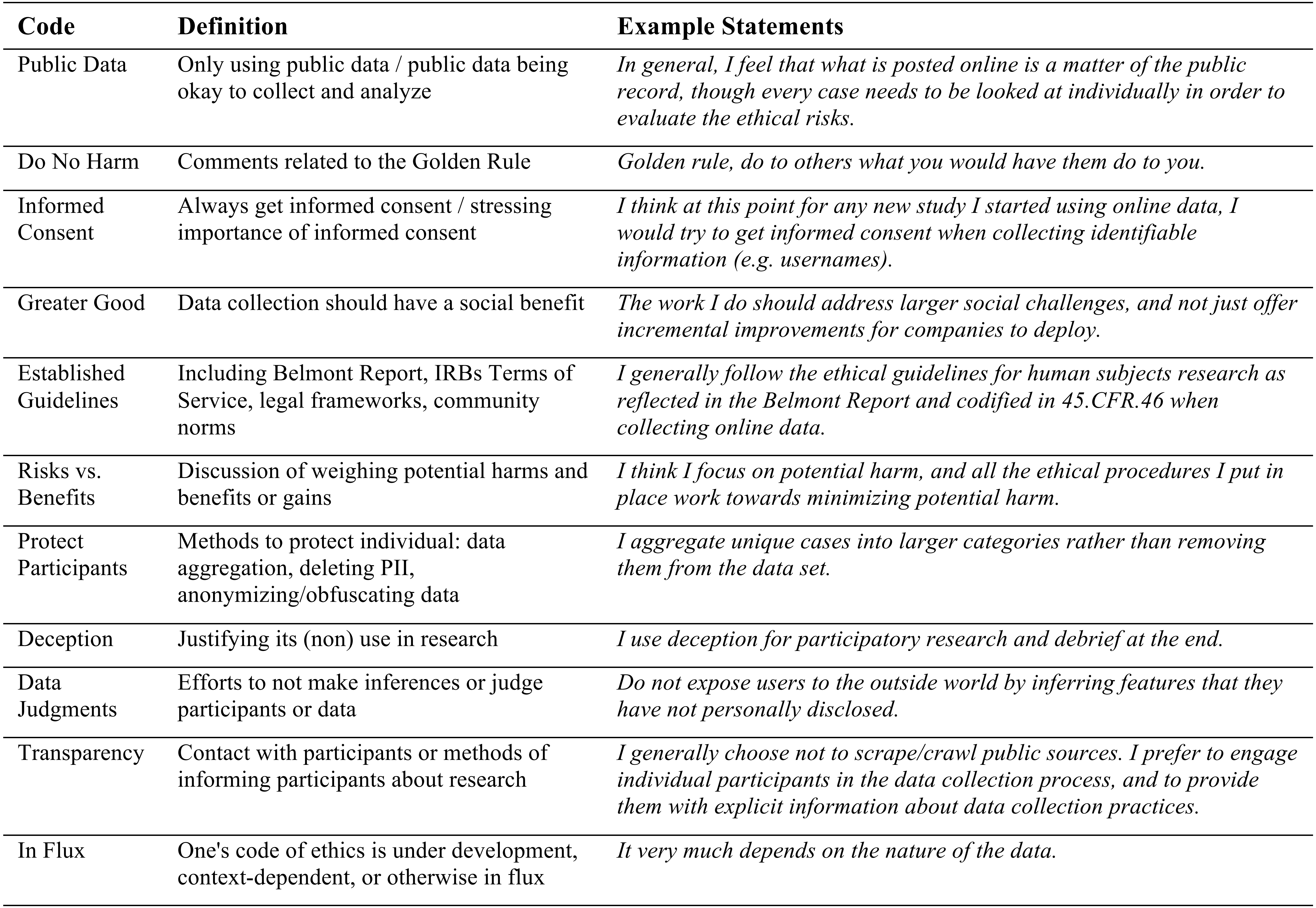
What do Twitter researchers do/think?

Vitak, Jessica, Katie Shilton, and Zahra Ashktorab. 2016. "Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community." ACM CSCW, pp. 941-953. 2016.



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 the public record, though every case needs to be looked at individually in order to evaluate the ethical risks.*
Do No Harm	Comments related to the Golden Rule	*Golden rule, do to others what you would 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	Methods to protect individual: data aggregation, deleting PII, anonymizing/obfuscating data	*I aggregate unique cases into larger categories rather than removing them from the data set.*
Deception	Justifying its (non) use in research	*I use deception for participatory research and debrief at the end.*
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 generally choose not to scrape/crawl public sources. I prefer to engage individual participants in the data collection process, and to provide them with explicit information about data collection practices.*
In Flux	One's code of ethics is under development, context-dependent, or otherwise in flux	*It very much depends on the nature of the data.*

Vitak, Jessica, Katie Shilton, and Zahra Ashktorab. 2016. "Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community." ACM CSCW, pp. 941-953. 2016.



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 the public record, though every case needs to be looked at individually in order to evaluate the ethical risks.*
Do No Harm	Comments related to the Golden Rule	*Golden rule, do to others what you would 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	Methods to protect individual: data aggregation, deleting PII, anonymizing/obfuscating data	*I aggregate unique cases into larger categories rather than removing them from the data set.*
Deception	Justifying its (non) use in research	*I use deception for participatory research and debrief at the end.*
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 generally choose not to scrape/crawl public sources. I prefer to engage individual participants in the data collection process, and to provide them with explicit information about data collection practices.*
In Flux	One's code of ethics is under development, context-dependent, or otherwise in flux	*It very much depends on the nature of the data.*

Some proposals

- OK to programmatically collect data without explicit consent
- But seek informed consent for all directly quoted content in publications
  - Twitter's view is that users retain rights to the content they post.

Williams, M. L., Burnap, P., Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users’ Views, Online Context and Algorithmic Estimation. Sociology, 51(6), 1149–1168.

More suggestions

Transparency with research communities

- Ask/inform
- Ethical deliberation with colleagues (in addition to IRBs)
- Be cautious about sharing results that include potentially identifiable outliers

Vitak, Jessica, Katie Shilton, and Zahra Ashktorab. 2016. "Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community." ACM CSCW, pp. 941-953. 2016.

4. What about data from corpora?

Data and Labelers

More ethical issues re: data

NLP systems (and machine learning models) can "reproduce or amplify unwanted societal biases reflected in training data" (Gebru et al 2020).

Data issues can cause NLP systems to fail for some populations (children, the elderly, speakers of dialects, minority languages)

Data has scientific implications

- What is the training/test split?
- Is the data appropriate for the task?
- How was the data labeled?

Datasheets, data statements, etc

Dataset creators:

- Encourage careful reflection on assumptions, risks, implications

Dataset consumers:

- Support informed decisions about using a dataset

Data sheets

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford. 2020. Datasheets for Datasets. Arxiv.

Motivation

- Why collected, who, how funded

Composition

- How many instances, how sampled, data split

Collection Process

- How collected, how metadata assigned, IRB, timeline, consent

Pre-processing

Uses

Distribution

Maintenance

Data Statements

Emily M. Bender and Batya Friedman. 2018. Data statements for NLP: Toward mitigating system bias and enabling better science. TACL 6, 587–604.

A "design solution and professional practice" for NLP

Should be included in NLP writings:

- papers presenting new datasets
- papers reporting experimental work with datasets
- system documentation

Data Statements Sample

Hate Speech Twitter (Waseem and Hovy 2016) https:// github.com/zeerakw/hatespeech

CURATION RATIONALE:

- In order to study the automatic detection of hate speech

LANGUAGE VARIETY:

- Twitter search API in late 2015. Information about which varieties of English are represented is not available, but at least Australian (en-AU) and US (en-US) mainstream Englishes are both included.

BCP-47 TAG:

- https://tools.ietf.org/rfc/bcp/bcp47.txt

SPEAKER DEMOGRAPHIC:

- Speakers were not directly approached for inclusion in this dataset and thus could not be asked for demographic information. More than 1,500 different Twitter accounts are included.

Emily M. Bender and Batya Friedman. 2018. Data statements for NLP: Toward mitigating system bias and enabling better science. TACL 6, 587–604.

Data Statement Sample (con't)

Annotator Demographic:

- This dataset includes annotations from both crowdworkers and experts. A total of 1,065 crowdworkers were recruited through Crowd Flower, primarily from Europe, South America, and North America. The expert annotators were recruited specifically for their understanding of intersectional feminism. They ranged in age from 20–40 years, included 3 men and 13 women, and gave their ethnicity as…

Speech Situation:

- All tweets were initially published between April 2013 and December 2015. Tweets represent informal,….

Text Characteristics:

- For racist tweets the topic was dominated by Islam and Islamophobia.

Emily M. Bender and Batya Friedman. 2018. Data statements for NLP: Toward mitigating system bias and enabling better science. TACL 6, 587–604.

Bender's Question:
What language is the paper studying?

"Surveys of EACL 2009 (Bender, 2011) and ACL 2015 (Munro, 2015) found 33–81% of papers failed to name the language studied. (It always appeared to be English.) "

- Bender and Friedman 2018.

What about labeling?

Did the paper use labels from an external dataset or were some data relabeled?

Who were they? Experts? Crowdworkers?

How were they trained?

- Are training example given in the paper?

How screened?

How were they compensated?

How aggregated to form final labels?

R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, Jenny Huang. 2020. Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? ACM FAT* 2020

What about harms to labelers?

E.g., the Kenyan labelers for OpenAI in the reading

Labeler Pay

Were the labelers paid minimum wage?

Whiting, Mark E., Grant Hugh, and Michael S. Bernstein. "Fair Work: Crowd Work Minimum Wage with One Line of Code." In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, pp. 197-206. 2019.

The community source of the data

Labov (1982:173), Wolfram

- Investigators who have obtained linguistic data from members of a speech community have obligations:
  - To make knowledge of that data available to the community
  - To actively pursue ways to return linguistic favors to the community

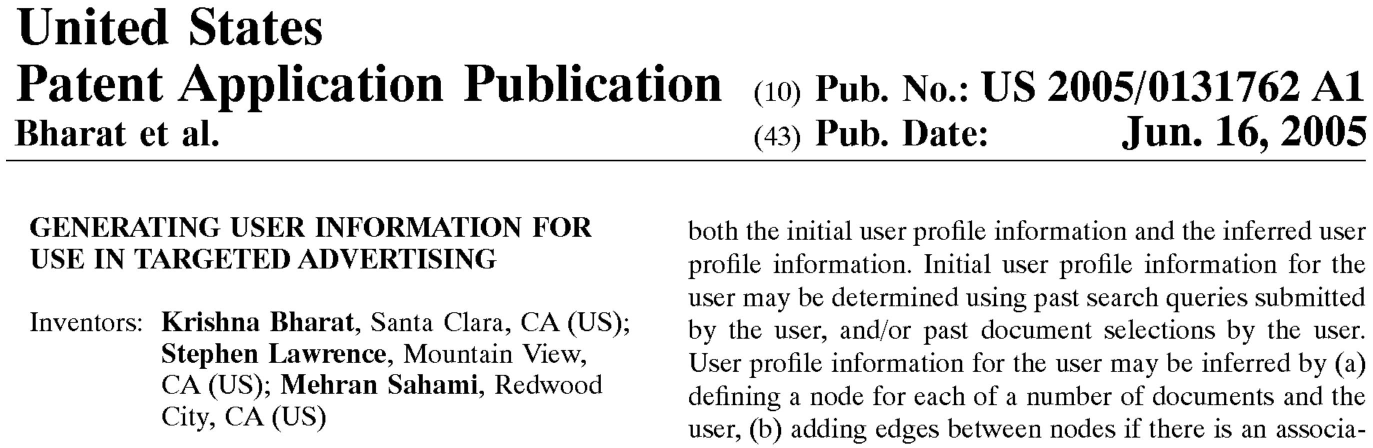
Rickford 1997

- We have "drawn substantially on data from the African American speech community… but… given relatively little in return."
- The contributions could… include the induction of African Americans into linguistics, the representation of African Americans in our writings, and involvement in courts, workplaces, and schools, especially with respect to the teaching of reading…

Rickford, John Russell. "Unequal partnership: Sociolinguistics and the African American speech community." *Language in Society* 26, no. 2 (1997): 161-197.

Case study

Famous Google 2003 patent:



United States
Patent Application Publication (10) **Pub. No.: US 2005/0131762 A1**
Bharat et al. (43) **Pub. Date: Jun. 16, 2005**

GENERATING USER INFORMATION FOR USE IN TARGETED ADVERTISING

Inventors: **Krishna Bharat**, Santa Clara, CA (US); **Stephen Lawrence**, Mountain View, CA (US); **Mehran Sahami**, Redwood City, CA (US)

both the initial user profile information and the inferred user profile information. Initial user profile information for the user may be determined using past search queries submitted by the user, and/or past document selections by the user. User profile information for the user may be inferred by (a) defining a node for each of a number of documents and the user, (b) adding edges between nodes if there is an associa-

Proposed to automatically induce information about the user even if, for privacy reasons, the user is purposely trying to conceal it from Google and does not give permission.

Question for discussion:

- Is this OK if Google does not give out the data to anyone else?
- When is it OK for us to infer demographics of users?

Zuboff, Shoshana. 2019. The Age of Surveillance Capitalism

Sample questions for our discussion

When and how is it OK to use data from the web?

What demographics of users is it OK to infer without permission?

When do we need to ask consent for NLP research?

For what kinds of NLP papers is it crucial to investigate multiple languages?

What human subjects issues should apply to crowdworkers or other data labelers?

Who should be an author on a paper?

Should we be putting data statements in our class papers?

Are there examples of your prior research practices that you now think you might want to change?