



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

AI Ethics and Social Impact

601.770

Anjalie Field

Sophia Hager

Acknowledgement: Some of the content in these slides is borrowed from Dan Jurafsky's ethics course and Yulia Tsvetkov's and Alan Black's course. Thank you!

How can we develop AI for good and not for bad?

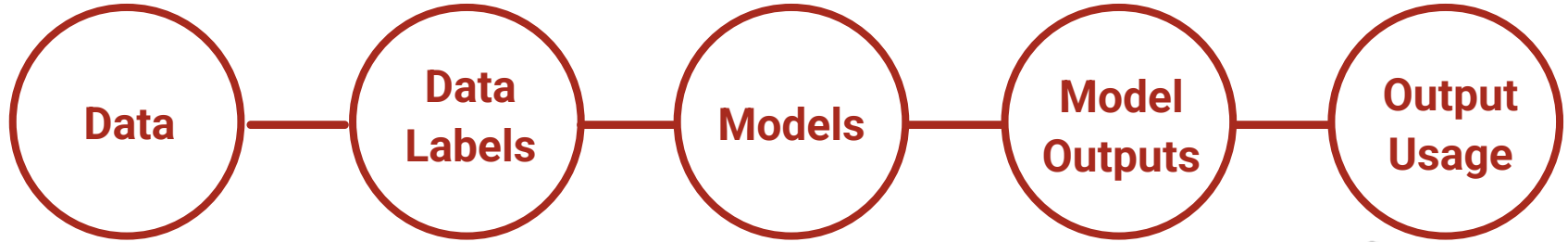
Decisions we make about our data, methods, and tools are tied up with their impact on people and societies



People create data



People use models



People build models



People are affected by models

Example: Are there some applications we should not build?

Hypothetical case: should we build a classifier to predict someone's sexual orientation from their photo?

Sexual Orientation Classifier

Who can be harmed by such a classifier?

- Personal attributes (gender, race, sexual orientation, religion) are complex social constructs, not categorial/binary, are dynamic, are private and often not visible publicly
- These are properties for which people are often discriminated against
 - In many places being gay is prosecutable
 - Such a classifier might affect people's employment, family relationships, health care opportunities, etc.

Additional Ethical Questions

- Who can benefit from such a classifier?
- Where does the training data come from?
- Did anyone consent?

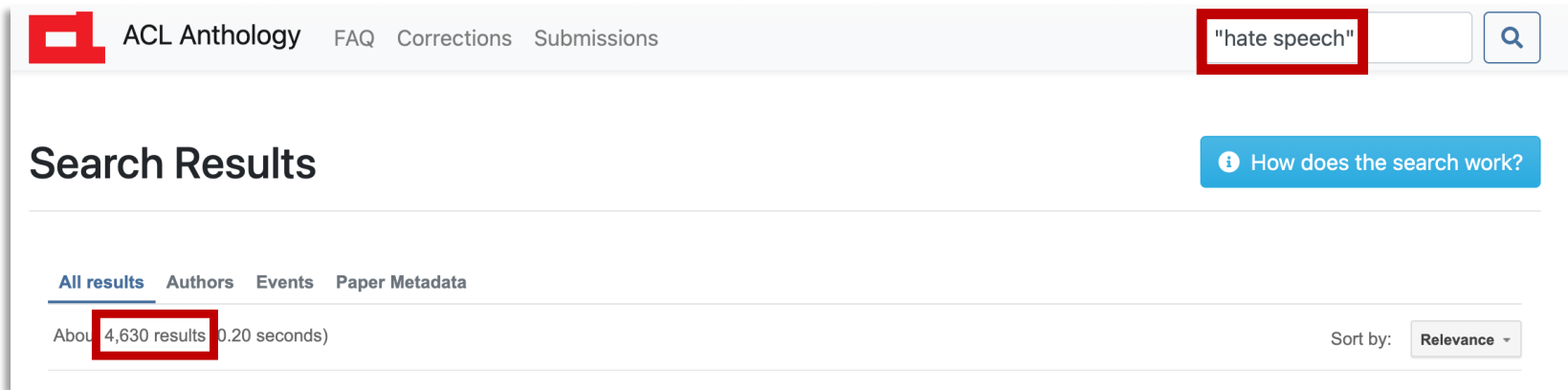
Most examples are not so straightforward

Problem:

- Hate speech and offensive language are prevalent on the internet and can lead to tangible harms
- Marginalized people are disproportionately targets of hate speech
- Manually identifying hate speech is difficult for human moderators
 - Too much volume to keep up with
 - Mental toll of reading offensive content

Technical Solution

- Build NLP models to identify hate speech automatically



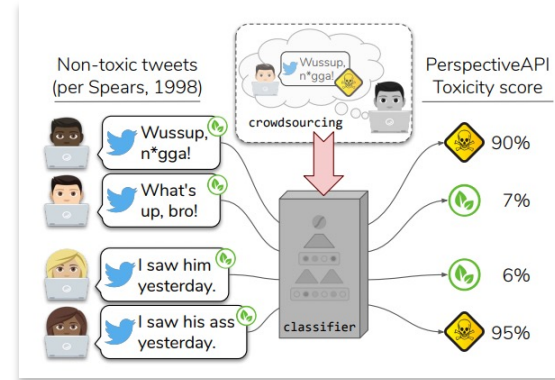
The screenshot displays the ACL Anthology search interface. At the top, the navigation bar includes the ACL Anthology logo, the text "ACL Anthology", and links for "FAQ", "Corrections", and "Submissions". A search bar on the right contains the query "hate speech" and a search icon. Below the search bar, the "Search Results" section is visible, featuring a blue button labeled "How does the search work?". Underneath, there are tabs for "All results", "Authors", "Events", and "Paper Metadata". The "All results" tab is selected, and the search results summary shows "About 4,630 results (0.20 seconds)". A "Sort by:" dropdown menu is set to "Relevance".

More Problems: NLP models are biased

Term	Toxic	Overall
atheist	0.09%	0.10%
queer	0.30%	0.06%
gay	3%	0.50%
transgender	0.04%	0.02%
lesbian	0.10%	0.04%
homosexual	0.80%	0.20%
feminist	0.05%	0.05%
black	0.70%	0.60%
white	0.90%	0.70%
heterosexual	0.02%	0.03%
islam	0.10%	0.08%
muslim	0.20%	0.10%
bisexual	0.01%	0.03%

Table 1: Frequency of identity terms in toxic comments and overall.

[Dixon et al. 2018]



[Sap et al. 2019]

Even more problems

- How do define what is offensive or hate speech?
 - Norms differ widely in different communities
 - Setting a universal standard is enforcing a majority viewpoint
- Who has control of the technology?
 - Concentrating power in few hands
- How might this technology be abused? (dual use potential)
 - Hate speech generator
 - Censorship

Solutions?

- Don't build NLP for hate speech detection?
- But then what about all the hate speech on the internet?
- Maybe we should ban social media? The internet?

This course will often be about *asking questions* and not necessarily *finding answers*

This Course

- Case studies of ethical challenges in AI research and development
 - Part 1 structured more around ethical challenges (data and privacy, fairness, etc.)
 - Part 2 structured more around applications (social services, criminal justice, healthcare)
 - Some readings are more technical, many are not
- Frameworks and guidelines
 - Code of ethics, policy and regulation
- Identifying and discussing series of questions
 - ~~What is AI?~~
What should AI be? [Virginia Dignum, Interspeech Keynote 2023]

Course website

- <https://ai-ethics-601-770.cs.jhu.edu/fa2024/>
- Please join the course Piazza!
 - <https://piazza.com/jhu/fall2024/601770>

Course Format

- Before class:
 - Read 1-2 papers on a topic
 - Post a 2-paragraph response on Piazza
 - Responses should **not summarize** the reading but instead **reflect**: raise points to think about or questions for discussion
 - Discussion leaders:
 - Review responses from others in group
 - Decide on questions and topics for discussion during the class period
- During class:
 - 45min small discussion groups (6-7 students)
 - Discussion leaders
 - 15min: Discussion leaders share back to the entire class
 - [10min: Intro to next course's readings]

Course Project

- Recommended groups of 3ish people
- Can be on any topic related to the course:
 - Technical paper
 - Survey/position paper

Grading

- Coursework (40%):
 - Reading responses (15%)
 - Class participation (25%)
- Project (55%):
 - Project literature review (15%)
 - Project written proposal and presentation (15%)
 - Project final presentation and paper (25%)
- Course goals and midterm feedback responses (5%)

Today's topic: Introduction

- Origins of research ethics
- Brief 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.
- Highlights:
 - 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...

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 Archives



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 Buxtun 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

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

What falls under IRB jurisdiction?

- Operate under federal (U.S. government) requirements
 - Don't cover non-U.S. institutions
- Institutions receiving federal support
 - Companies that don't receive federal grants may have internal review processes but are not mandated to have IRBs
- "Human subject research"
 - Biospecimens
 - Identifiable private information

The Belmont Report

- Respects for persons
 - “Individuals should be treated as autonomous agents. Persons with diminished autonomy are entitled to protection”
- Beneficence
 - “Do no harm. Maximize benefit and minimize risk”
- Justice
 - “The benefits and risks shared by a population that may benefit from the results of research”
 - “Another way of conceiving the principle of justice is that equals ought to be treated equally”

Discussion

1. This work proposes a mapping between different frameworks: {respects for persons, beneficence, justice} and {transparency, fairness, accountability}
 - A. In the context of AI, what do principles of transparency, fairness, and accountability mean to you?
 - B. Did you find this mapping meaningful?
2. The Belmont Report and IRB are only intended to apply to *research*. What are some examples where this framework may be difficult to apply in industry and deployment?
 - A. E.g. Beneficence: what are possible risks and benefits of deploying a content curation algorithm? Who are relevant stakeholders?
3. IRB / the Belmont Report reflect a top-down or regulation/law approach (what does this mean?), what might be more of a capital marketplace approach?
 - A. What are some of the tradeoffs of each approach?

Closing logistics

- NO CLASS THURSDAY (8/28)
- Next topic: Data
 - Readings for Tuesday 9/3:
 1. [Lundberg, Ian, et al. "Privacy, ethics, and data access: A case study of the Fragile Families Challenge." Socius 5 \(2019\)](#)
 2. [Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." FAccT. PMLR, 2018](#)
 3. ["Facial recognition's 'dirty little secret': Millions of online photos scraped without consent", NBC article](#)
- Post 2-paragraph response on Piazza by 8pm Monday night (9/2)!

Initial Discussion Groups

- Introduce yourselves
- Decide on a discussion leader for next class (Tuesday)