



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

AI Ethics and Social Impact

601.770

Anjalie Field

Krithika Ramesh

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

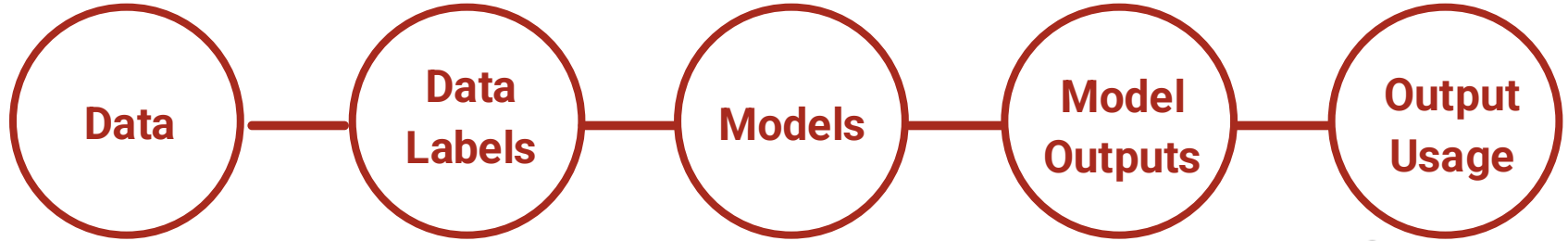
Decisions we make about our data, methods, and tools are tied up with their impact on people and societies



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?

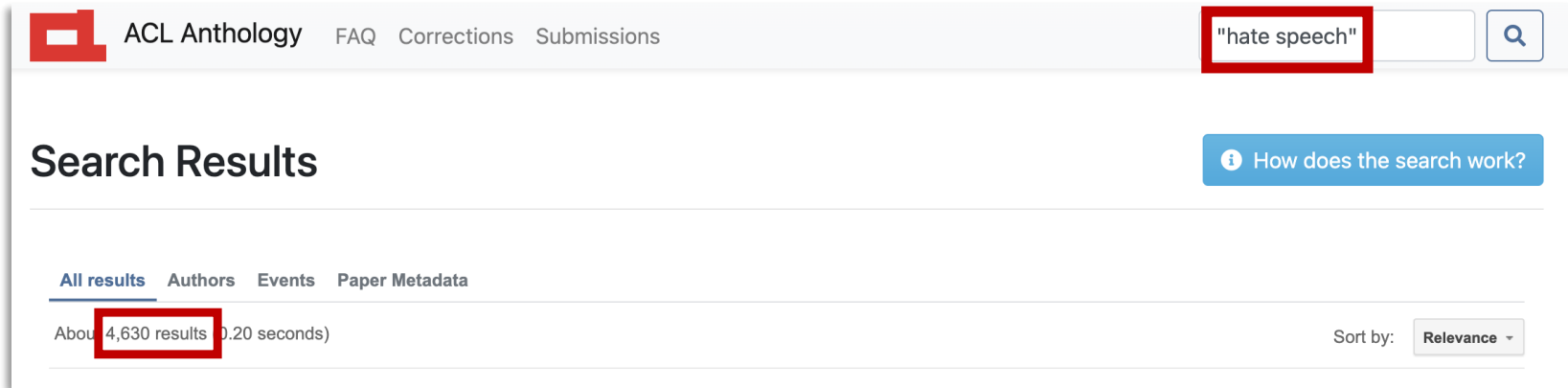
Most examples are not so straightforward

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 shows the ACL Anthology search interface. At the top, the navigation bar includes the ACL Anthology logo 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 displayed. A blue button with an information icon and the text "How does the search work?" is located to the right of the search results header. Underneath, there are tabs for "All results", "Authors", "Events", and "Paper Metadata". The "All results" tab is selected. Below the tabs, the search results summary shows "About 4,630 results (0.20 seconds)". To the right of this summary is a "Sort by:" dropdown menu set to "Relevance".

ACL Anthology FAQ Corrections Submissions

"hate speech" 

Search Results [How does the search work?](#)

[All results](#) [Authors](#) [Events](#) [Paper Metadata](#)

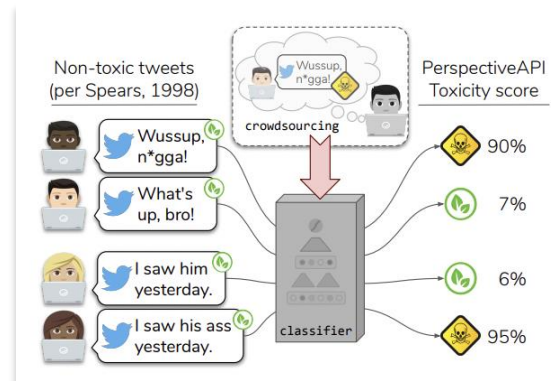
About 4,630 results (0.20 seconds) Sort by: **Relevance** ▾

More Problems: hate speech detectors 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 AI/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 always *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/fa2026/>
- Current plan: Announcements through Canvas

Course Format

- Before class:
 - Read 3-4 papers on 2 topics (~2 papers per topic)
 - Discussion leaders:
 - Decide on questions and topics for discussion during the class period
 - Send discussion plan to Anjalie and Krithika by 5pm Friday
- During class:
 - For each topic:
 - 10min: Individual journal reflection on the assigned readings
 - Don't be late!
 - 45min small discussion groups (6-7 students)
 - Discussion leaders
 - 15min: Discussion leaders share back to the entire class
- Occasionally we will deviate with a lecture or guest

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 (20%)
 - Final Project Q&A (10%)

Today's topics

Introduction:

- Origins of research ethics
- Brief history of human subjects protection

Model development pipelines: Data ownership

- Brief background and reading break

Practice discussion



JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

The Belmont Report

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

An ethical framework that offers very practical and applied guidance

"to identify the basic ethical principles that should underlie the conduct of biomedical and behavioral research involving human subjects and to develop guidelines which should be followed to assure that such research is conducted in accordance with those principles"

- A. Boundaries Between Practice and Research
- B. Basic Ethical Principles
- C. Applications

The Belmont Report: Respect for Persons

1. "Individuals should be treated as autonomous agents."
2. "Persons with diminished autonomy are entitled to protection"

Example of 1:

- People should engage in research voluntarily and with informed consent

Example of 2:

- IRBs required increased scrutiny for studies involving vulnerable populations like people in prison

The Belmont Report: Beneficence

“Do no harm. Maximize benefit and minimize risk”

- Example of failure: Tuskegee syphilis experiment

The Belmont Report: Justice

- "fairness in distribution" or "what is deserved."
- "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"

Example failure: unwilling prisoners as research subjects in Nazi concentration camps

The Belmont Report: Justice

- Informed consent
 - Information, Comprehension, Voluntariness
- Assessment of Risks and Benefits
- Selection of Subjects



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

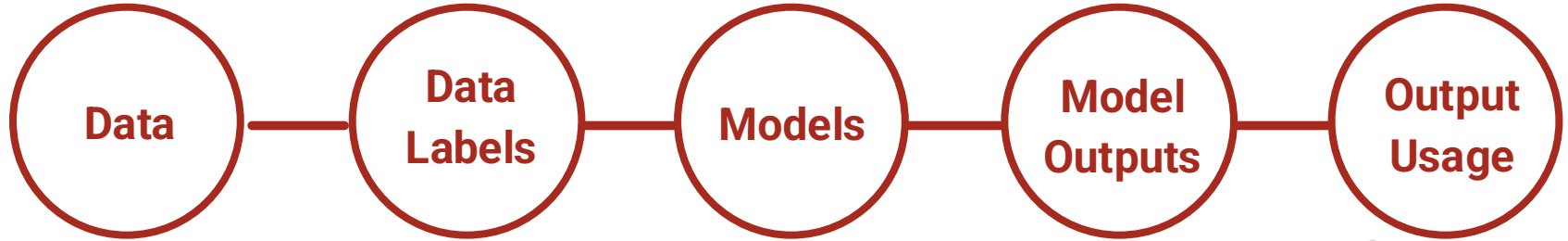
Data Ownership



People create data



People use models



People build models



People are affected
by models

Brief History of Facial Recognition Research

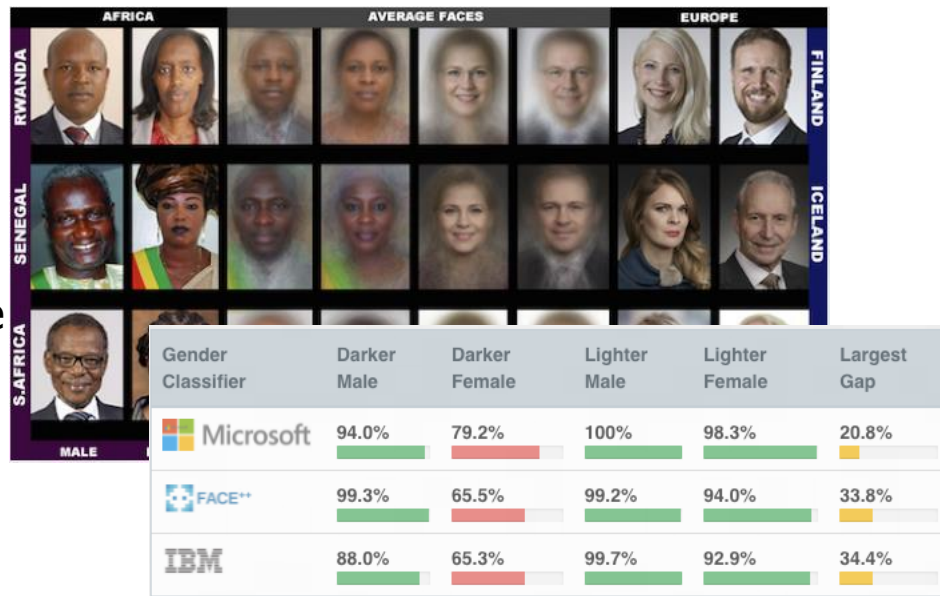
- 1960s: Woody Bledsoe, Helen Chan Wolf, and Charles Bisson try to get computers to recognize faces
- 1993-1996: DARPA FERET project
 - “to develop automatic face recognition capabilities that could be employed to assist security, intelligence, and law enforcement personnel in the performance of their duties”
- 1996-2006: FERET dataset simulates research interest, but models are not very good
- 2007-2013: Mainstream development for unconstrained settings
 - Hunt for “in the wild data sets”
- 2014-present: Deep learning leads to more accurate models and commercialization

<https://www.nist.gov/programs-projects/face-recognition-technology-feret>

Raji, Inioluwa Deborah, and Genevieve Fried. "About face: A survey of facial recognition evaluation." AAAI 2020 Workshop on AI Evaluation (2021).

Gender shades: Intersectional accuracy disparities in commercial gender classification

- Popular benchmark datasets are overwhelmingly composed of lighter-skinned subjects
- Evaluation of 3 commercial gender classification systems shows that darker-skinned women are by far the most misclassified group



“Diversity in faces” data set

- Proposed Solution:
 - We need to train and evaluate models on more diverse data sets
- “Diversity in faces” data set, containing more diverse images annotated with features
 - Data was collecting from an existing ML data set of faces, originally collecting from Flickr, a photo sharing website
 - Annotated data for various features including facial geometry and skin-tone



“Diversity in faces” data set

- “This is the dirty little secret of AI training sets. Researchers often just grab whatever images are available in the wild,” - NYU School of Law professor Jason Schultz
- “Despite IBM’s assurances that Flickr users can opt out of the database, NBC News discovered that it’s almost impossible to get photos removed. IBM requires photographers to email links to photos they want removed, but the company has not publicly shared the list of Flickr users and photos included in the dataset”
- Lawsuits over this use of data filed by people whose photos were in it
- IBM took down the data set



Where is facial recognition technology not used?

2018:

- Gender Shades Project
- Intense criticism of facial recognition research as surveillance tech (petitions, letter to Jeff Bezos)

2020:

- Amazon places 1-year moratorium on use of its face-processing software by police agencies (which they extended at least one more year)
- Microsoft stops selling facial-recognition software to police
- IBM halts work on face recognition because it's used for racial profiling

<https://www.technologyreview.com/2020/06/12/1003482/amazon-stopped-selling-police-face-recognition-fight/>

<https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/>

<https://www.technologyreview.com/2020/06/09/1002947/ibm-says-it-is-no-longer-working-on-face-recognition-because-its-used-for-racial-profiling/>

How is facial recognition technology used?

Ad from a face image processing company

“We live in a dangerous world, where harm doers and criminals easily mingle with the general population; the vast majority of them are unknown to the authorities....Public Safety agencies, city police department, smart city service providers and other law enforcement entities are increasingly strive for Predictive Screening solutions, that can monitor, prevent, and forecast criminal events and public disorder without direct investigation or innocent people interrogations.

“What if it was possible to know whether an individual is a potential pedophile, an aggressive person, or a criminal?”

How is facial recognition technology used?

- Emotion expression and intensity
- Sexual orientation
- Political orientation
- Personality traits

A large body of work in psychology and anthropology has provided evidence that “there is no relationship between how we look and how trustworthy or intelligent we actually are”, though opposing research exists

Severin Engelmann, Chiara Ullstein, Orestis Papakyriakopoulos, and Jens Grossklags. 2022. What People Think AI Should Infer From Faces. (FACCT '22) <https://doi.org/10.1145/3531146.3533080>

Journal reflection and Break

- Briefly introduce yourself
- Why are you taking this class?
- What do you hope to learn?
- Discussion groups:
 - We will start with asking everyone to introduce themselves and share a few thoughts on the topics presented today

Initial Discussion Groups

Group 1

Acharya, Sawal
Bellamkonda Mahendra, Rohan
Goswami, Archit
Klimov, Mirra
Long, Fay
Oberoi, Vansh
Sohani, Janhvi Sushant
Wei, Yana

Group 2

AlNaseeb, Ali Bader
Chang, Sinyin
Gundugurki Narasimhamurthy, Devothama
Lim, Han
McCormick, Kiera Anne
Pham, Toan Quang
Tugbobo, Iyinoluwa
Zhou, Didi W

Group 3

Alsadi, Abdullah
Chen, Cara
Kim, Evie Rose
Lin, Xi
Navale, Sanket Sachin
Phung, Trung Quoc
Wang, Carly G

Discussion Groups

- Make sure that each person in your group is talking. If someone is talking less, find a way to draw them in. This is one of the most important parts of your role as facilitator: do your best to have everyone talk (very roughly) the same amount.
- Try not to respond to every comment that everyone makes at your table; instead if no one is talking ask the table what they think about what has just been said. If it feels like the right time to summarize responses, maybe do so without taking a stand one way or another.
- If you need an icebreaker, you might ask a question for which there is no single correct answer and go around the table with it. (Example: "On a scale of 1 to 10, how successful do you think the paper is at solving the problem they pose" or "If you were going to do a project following up on this paper, which part of the paper might be most rewarding to look into?")
- Try to be careful about giving credit for ideas

Class recap

- **Not** intended to be a summary of the entire discussion, instead, it should be the **outcomes** of your discussion

Closing logistics

- No class 9/7: Labor Day
- Next class (9/14):
 - Data Privacy
 - Crowdsourcing annotations
- Decide on a discussion leader for next class

NIST Evaluation Program

- <https://www.nist.gov/programs-projects/face-projects>

Face Recognition Technology Evaluation (FRTE) 1:1 Verification

[Latest Report](#) | [Participation Agreement](#) | [API](#) | [Validation](#) | [Encryption](#) | [Submit](#)

▼ Status

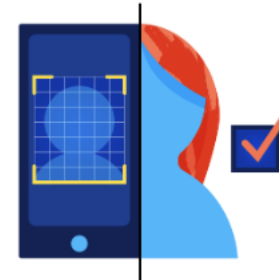
[2024-08-09] A new FRTE 1:1 report has been published.

[2024-07-03] NIST included all algorithms evaluated since the ongoing FRTE evaluation started in 2017. Previously we listed only results for each developer's two most recent implementations. This change allows end-users to reference results for algorithms that have been superseded by newer variants, but which may still be used operationally. We added a "by developer" tab which lists only the most accurate algorithm from each developer, where accuracy is defined by the lowest FNMR value for the visa-border dataset.

[2024-03-27] NIST discontinued running the FRTE Visa-Visa benchmark on 1:1 algorithms submitted to FRTE.

[2023-07-03] All algorithms, participation agreements, and GPG keys should be submitted using the [FRTE Submission Form](#).

Except PAD, all FRTE tracks are open. Algorithms submitted to the FRTE 1:1 track will be run on all the datasets documented in the FRTE 1:1 report.



Credit: Natasha Hanacek/NIST