CS378: Natural Language Processing

Lecture 10: Ethics in NLP

TEXAS
The University of Texas at Austin

Eunsol Choi

Some slides adapted from Yoav Artzi / Greg Durrett

Course planning

- HW1 grade released! Most of you did well. :)
- HW2 due Thursday, HW3 will be released by the end of this week
- HW3 has **two** deadlines
  - You can't really use slip day for the first one — as it will block your classmates for moving onto the second part
  - One for designing annotations *and* providing annotations
  - One for analyzing the data you have collected

HW3 Task Design: Who designs the task?

- NLP researchers themselves!
- Task are designed to…
  - delve into linguistic phenomena
  - support user-facing applications

- Is the task well-defined?
  - Given the same input, would annotators consistently provide the same label?
  - Inter-annotator agreement

Two notions of inter-annotator agreement

- Inter-annotator agreement
  - How would people disagree with each other?
  - For subjective tasks, it might be better to model distribution of human judgement!

- Test-retest disagreement
  - If the same person label it again, would it yield the same label?

Is high agreement score enough?

- For some tasks, such as toxicity task, people **disagree**
  - What counts as a harassment?
  - What counts as salient information?
- Sometimes, the input text is ambiguous

p: Paula swatted the fly.
h: The swatting happened in a forceful manner.

Who took control of the italian government in 1922?

National Fascist Party?

Benito Mussolini?

- Is perfect agreement what we want in such cases?
  - we can predict a label distribution

Course Planning

- Today: Meta-NLP, ethics in NLP
- Coming Thursday:
  - Word Embeddings
- Next week:
  - Language Models

Today

- Getting started with NLP research project
  - Where should we start?
  - Dataset
  - Evaluation
  - Model

- Ethics in NLP
  - Overview of potential issues
  - In-class debate

Evaluation

- Qualitative and quantitative evaluation
- Goal: provide evidence that your research hypothesis is correct
- Human evaluation is often necessary for text generation

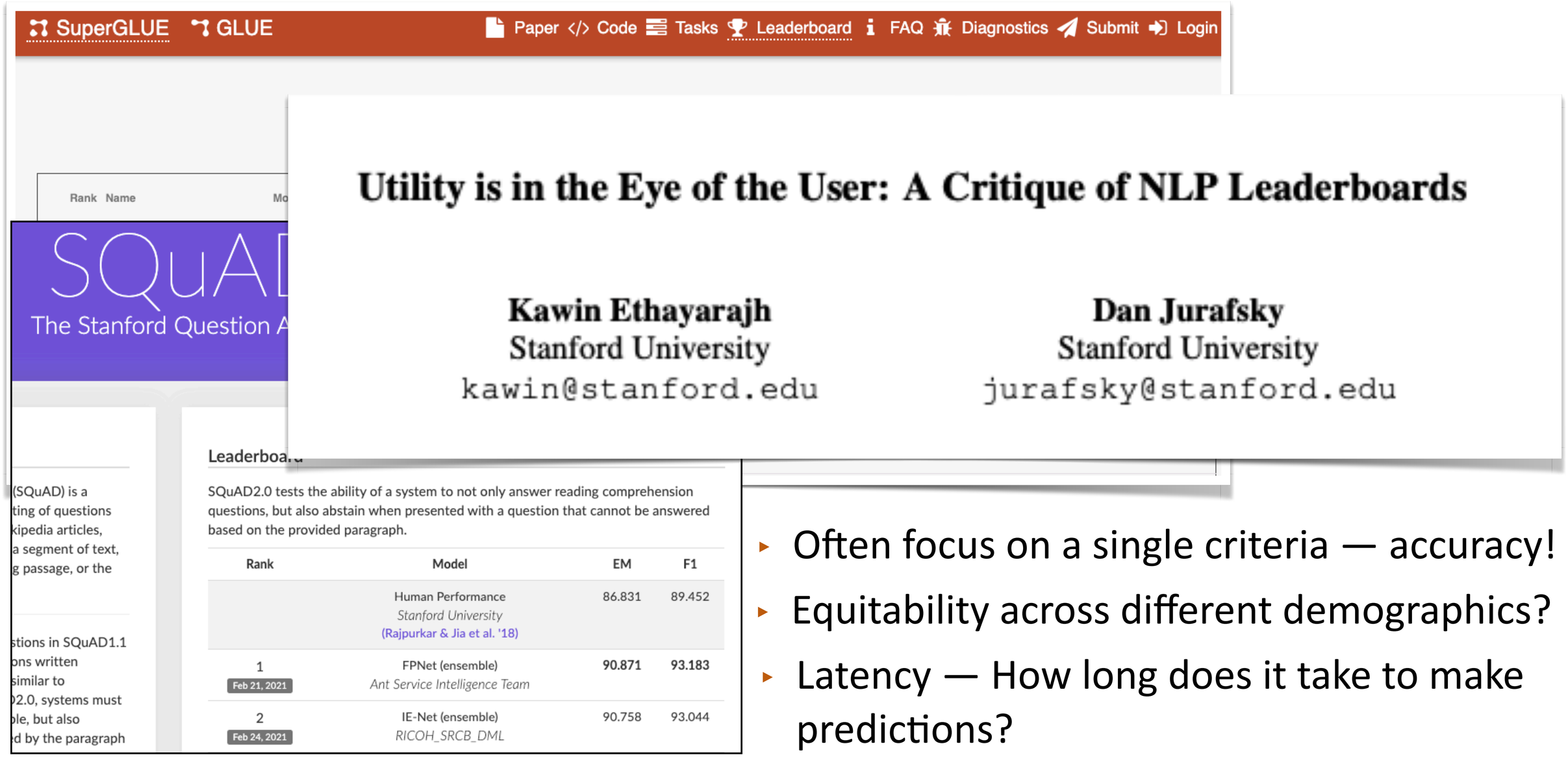
Formative vs. Summative Evaluation

When the cook tastes the soup, that's formative;

when the customer tastes the soup, that's summative

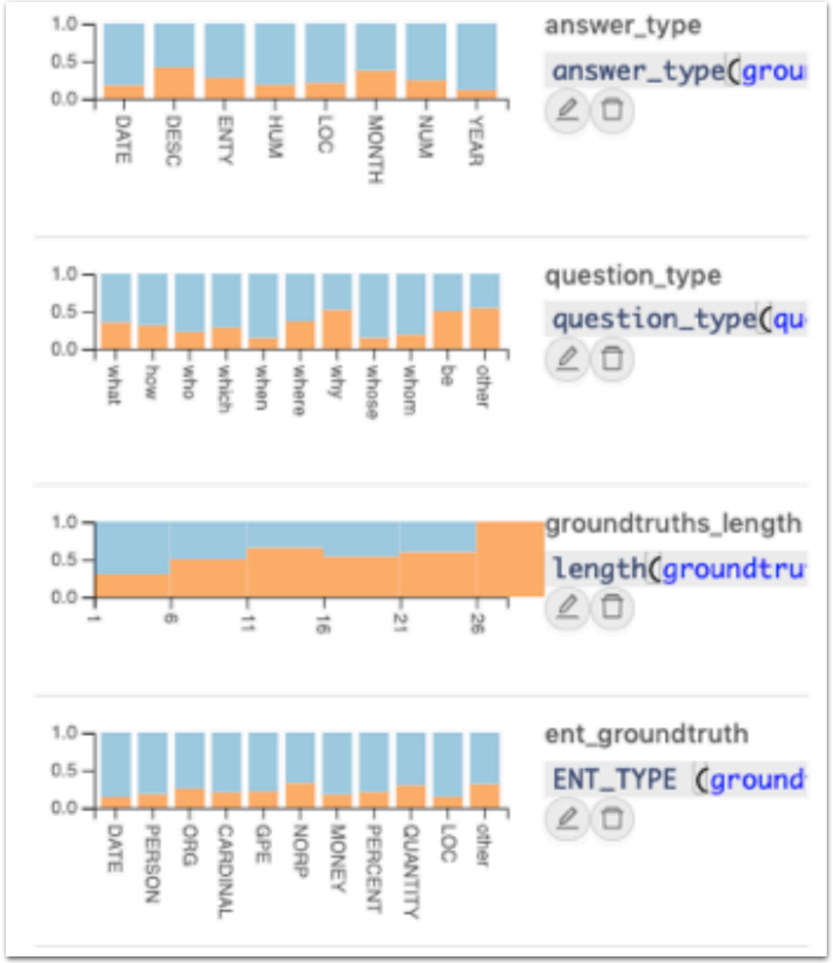
- Formative evaluation:
  - Sanity check
  - Typically lightweight automatic metrics
  - For tuning hyperparameters, etc
- Summative evaluation:
  - Comparing your method to previous methods
  - Compare major components of your method
  - Human evaluations

NLP Leaderboards



- Often focus on a single criteria — accuracy!
- Equitability across different demographics?
- Latency — How long does it take to make predictions?

More nuanced leaderboards



- Understanding the error patterns
- Aligning with human values
  - Not all errors are equally damaging
- Bringing evaluation into the loop of model development — find examples where existing models fail, and evaluate on them
  - Problems?

[Wu et al, ACL 2019]

Today

- Getting started with NLP research project
  - Where should we start?
  - Dataset
  - Evaluation
  - Model

- Ethics in NLP
  - Overview of potential issues
  - In-class debate

Model

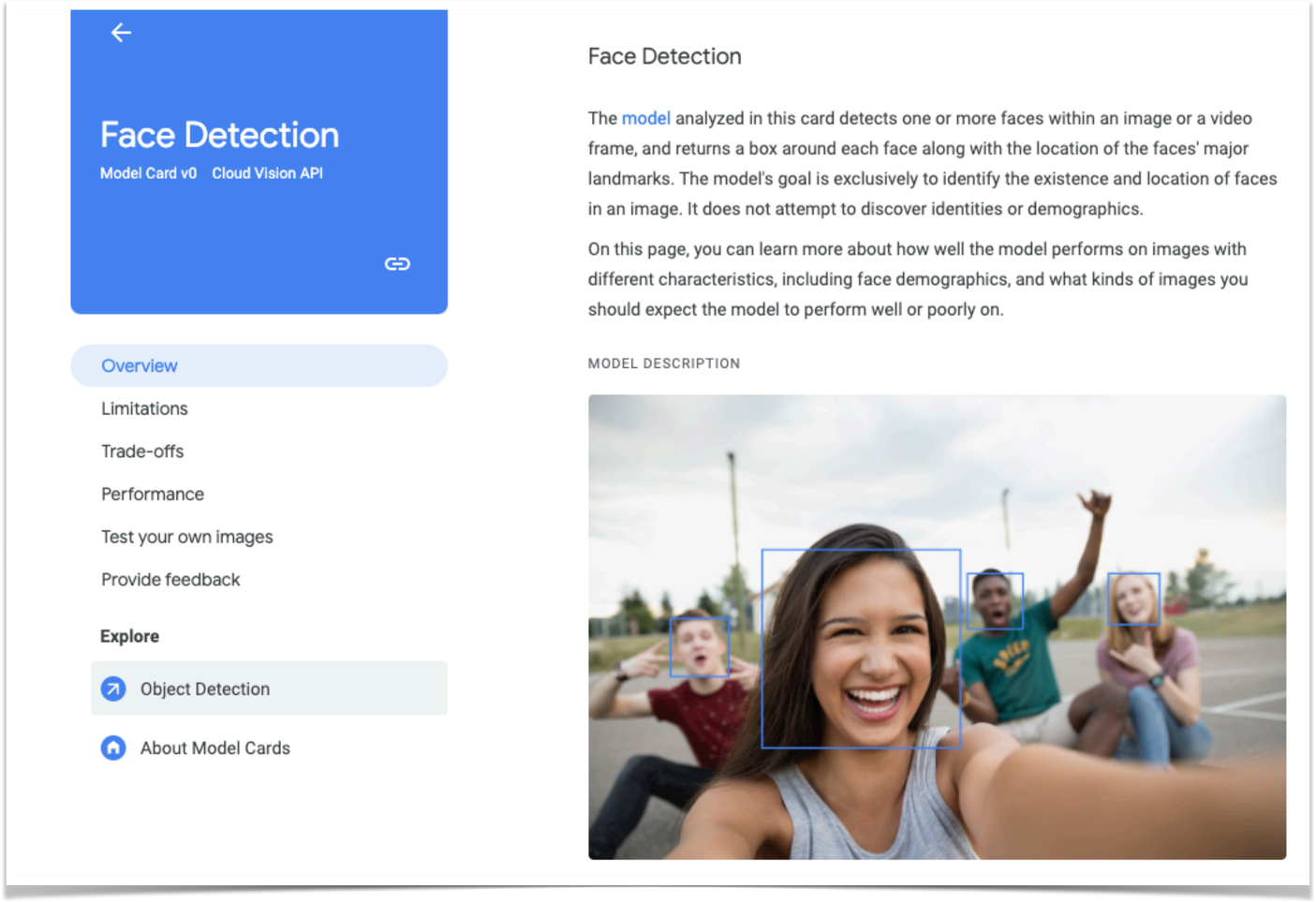
- Build a simple baseline
  - e.g., Majority class label
- Build a strong baseline
  - Existing published work can be a good baseline
  - You don't necessarily have to beat them, especially if they are using a lot of resources that you do not have access to
- Motivate your model
  - In what aspect your proposed model improve upon baseline?

Hyperparameter Tuning

- You should tune both your baseline **AND** your new model

- During literature review, pay attention to what hyper parameters matter, and what are typical values

Documenting your model: Model Card



- Documentation detailing their performance characteristics
- Intended use
- Training Data / Evaluation Data / Evaluation Metric
- Caveats and Recommendations

Mitchell et al, FAT* 19

Documenting your model: Model Card

Model Card - Smiling Detection in Images

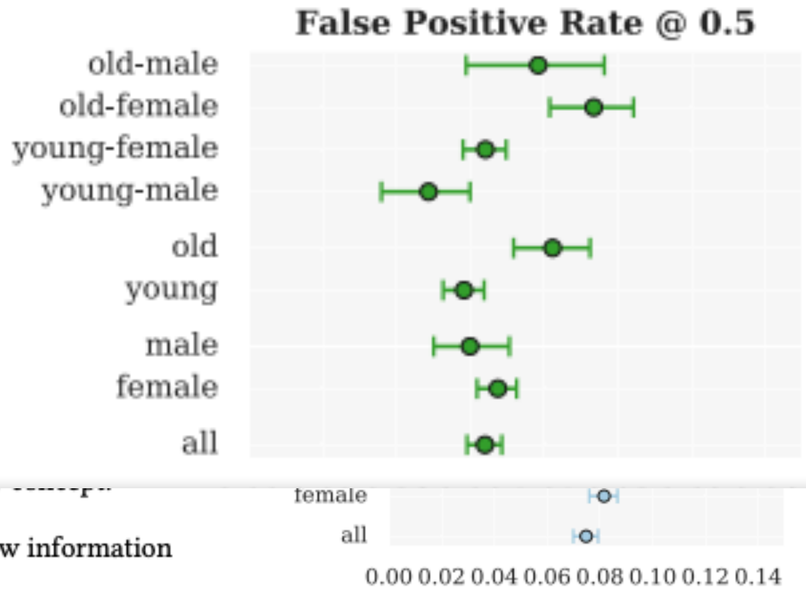
Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion dete
based on physical appearanc

Quantitative Analyses



Ethical Considerations

- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Figure 2: Example Model Card for a smile detector trained and evaluated on the CelebA dataset.

Today

- Getting started with NLP research project
  - Where should we start?
  - Dataset
  - Evaluation
  - Model

- Ethics in NLP
  - Overview of potential issues
  - In-class debate

Overview of ethical issues

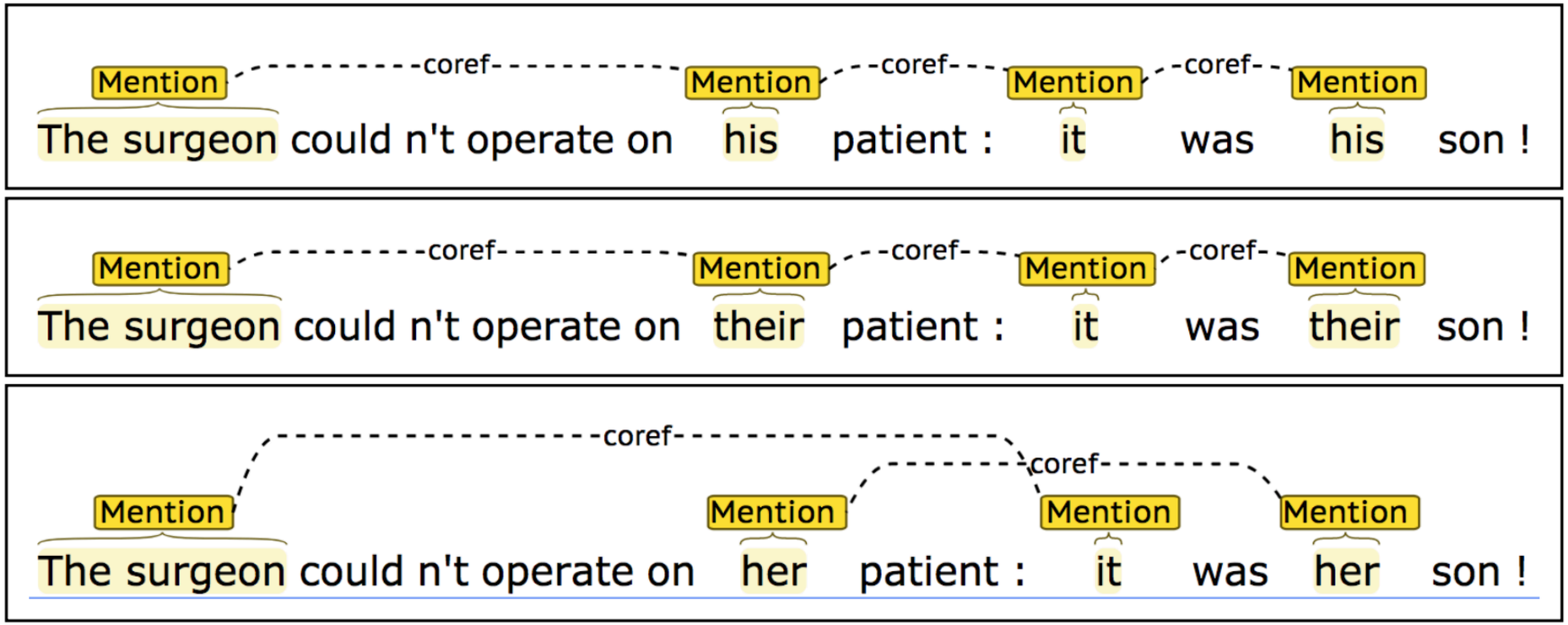
- **Social bias encoded in NLP models and tasks**
- Treatment of human subjects
- Misuse of NLP technology
- Privacy and anonymity
- Research Integrity

Social Bias Encoded in Data / Model: Gender

- Bias in data: 67% of training images involving cooking are women, model predicts 80% women cooking at test time — amplifies bias
- Can we constrain models to avoid this while achieving the same predictive accuracy?
- Place constraints on proportion of predictions that are men vs. women?

Zhao et al. (2017)

Social Bias Encoded in Data / Model: Gender



- Coreference: clustering entity mentions that refers to the same entity
- Models make assumptions about genders and make mistakes as a result

Rudinger et al. (2018), Zhao et al. (2018)

Social Bias Encoded in Data / Model: Gender

(1a) **The paramedic** performed CPR on the passenger even though she/he/they knew it was too late.

(2a) The paramedic performed CPR on **the passenger** even though she/he/they was/were already dead.

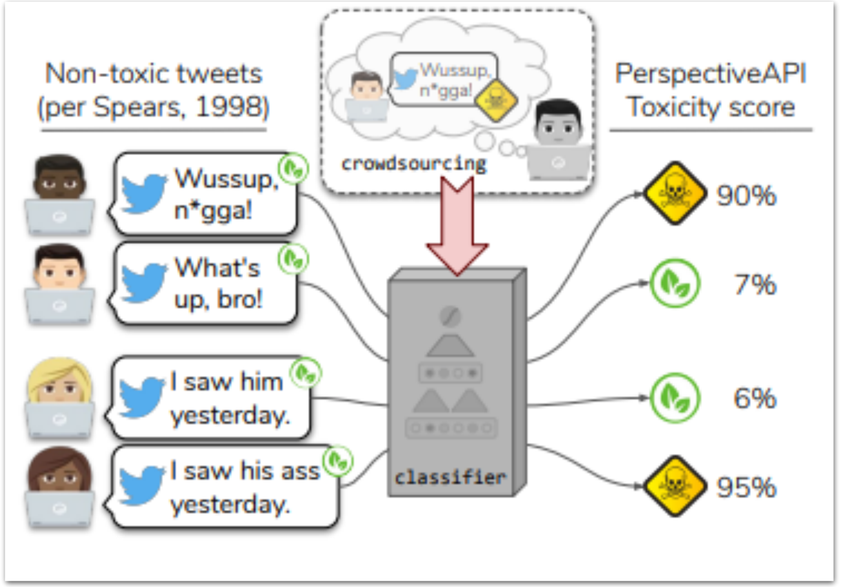
(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

- Can form a targeted test set to investigate

Rudinger et al. (2018), Zhao et al. (2018)

Social Bias Encoded in Data / Model: Race



- Existing hate speech classifiers are likely to falsely label text containing identity terms like 'black' or text containing linguistic markers of African American English (AAE) as toxic.

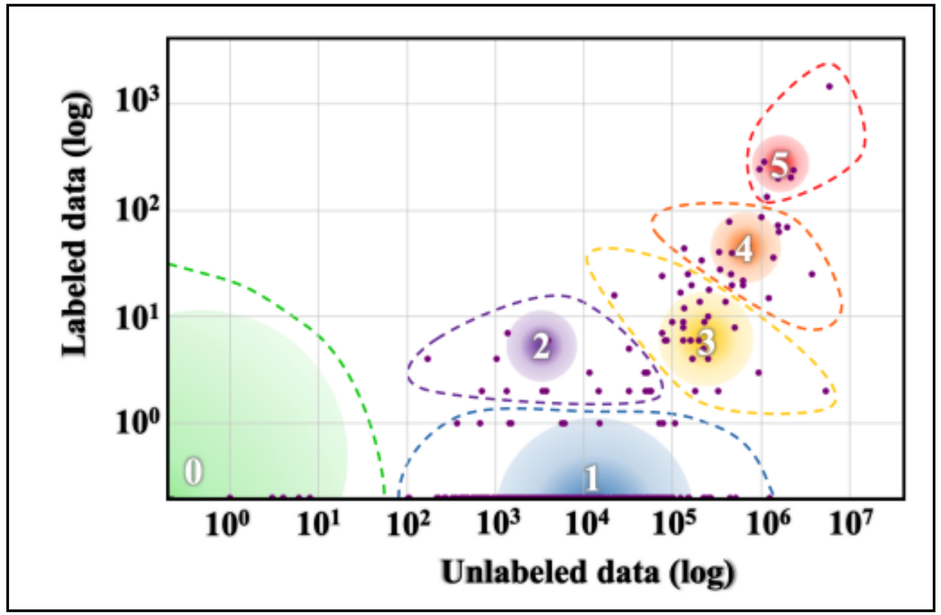
- This can be alleviated with more careful data collection — annotators are less likely to label tweets using AAE as toxic if they were told the likely language variety of tweets. [Sap et al ACL 2019]

https://arxiv.org/pdf/2106.11410.pdf

Representation Disparity: Language

- Publications per language

- Labeled (and unlabeled) dataset distribution



Class	5 Example Languages	#Langs	#Speakers	% of Total Langs
0	Dahalo, Warlpiri, Popoloca, Wallisian, Bora	2191	1.2B	88.38%
1	Cherokee, Fijian, Greenlandic, Bhojpuri, Navajo	222	30M	5.49%
2	Zulu, Konkani, Lao, Maltese, Irish	19	5.7M	0.36%
3	Indonesian, Ukranian, Cebuano, Afrikaans, Hebrew	28	1.8B	4.42%
4	Russian, Hungarian, Vietnamese, Dutch, Korean	18	2.2B	1.07%
5	English, Spanish, German, Japanese, French	7	2.5B	0.28%

https://arxiv.org/pdf/2110.06733.pdf

https://aclanthology.org/2020.acl-main.560.pdf

Overview of ethical issues

- Social bias encoded in NLP models and tasks
- **Treatment of human subjects**
- Misuse of NLP technology
- Privacy and anonymity
- Research Integrity

Treatment of human subjects

- If the text data includes user data, did they consent to such usage?

- Was people involved in annotation treated ethically and fairly?
  - Many crowd workers are paid below minimum wage [Silberman et al].

Overview of ethical issues

- Social bias encoded in NLP models and tasks
- Treatment of human subjects
- **Misuse of NLP technology**
- Privacy and anonymity
- Research Integrity

Dangers of Automatic Systems

"Instead of relying on algorithms, which we can be accused of manipulating for our benefit, we have turned to machine learning, an ingenious way of disclaiming responsibility for anything. Machine learning is like money laundering for bias. It's a clean, mathematical apparatus that gives the status quo the aura of logical inevitability. The numbers don't lie."

- Maciej Cegłowski

Slide credit: Sam Bowman

Dangers of Automatic Systems

- "Amazon scraps secret AI recruiting tool that showed bias against women"
  - "Women's X" organization was a negative-weight feature in resumes
  - Women's colleges too
- Was this a bad model? May have actually modeled downstream outcomes correctly…but this can mean learning humans' biases
- Does the model behave equally well across different groups?
  - Equal accuracy?
  - Equal positive rates?

Slide credit: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Dangers of Automatic Systems

THE VERGE TECH SCIENCE CULTURE CARS REVIEWS LONGFORM VIDEO MORE

US & WORLD | TECH | POLITICS

Facebook apologizes after wrong translation sees Palestinian man arrested for posting 'good morning'

14

Facebook translated his post as 'attack them' and 'hurt them'

by Thuy Ong | @ThuyOng | Oct 24, 2017, 10:43am EDT

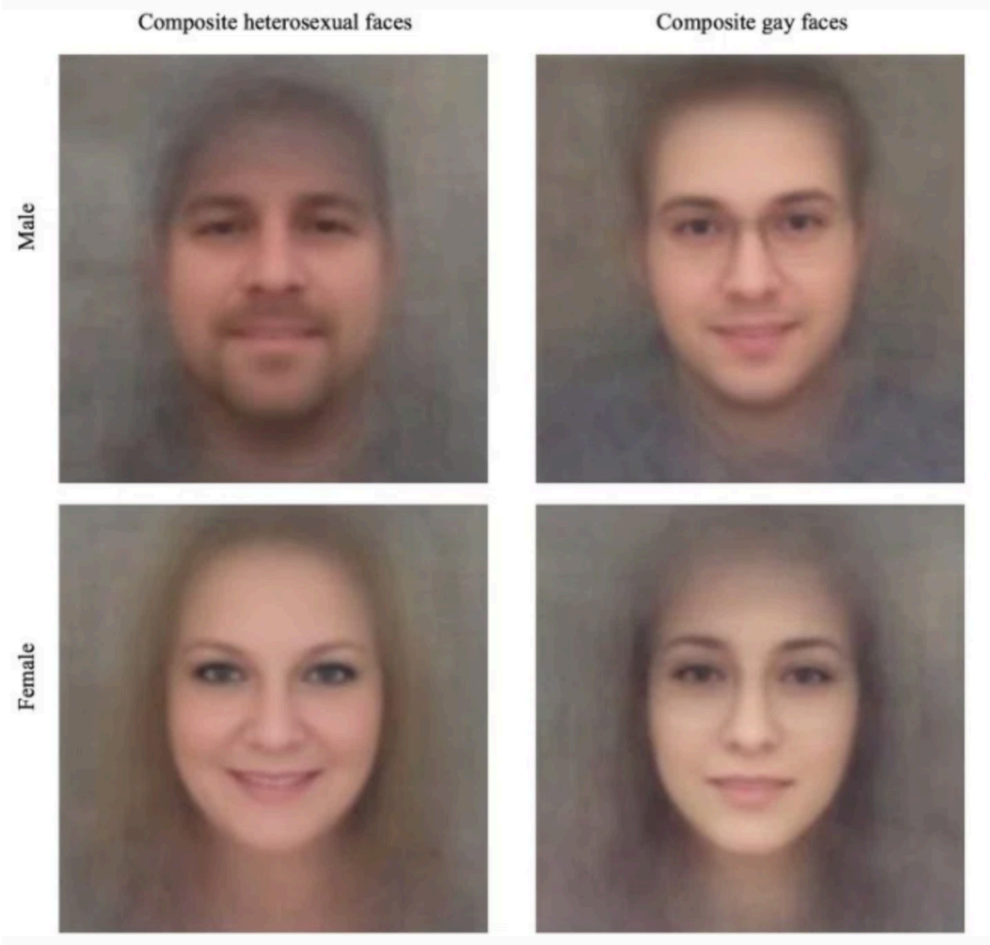
Slide credit: The Verge

Non-exhaustive list of potential harms

1. Directly facilitate injury to living beings. For example: could it be integrated into weapons or weapons systems?
2. Raise safety or security concerns. For example: is there a risk that applications could cause serious accidents or open security vulnerabilities when deployed in real-world environments?
3. Raise human rights concerns. For example: could the technology be used to discriminate, exclude, or otherwise negatively impact people, including impacts on the provision of vital services, such as healthcare and education, or limit access to opportunities like employment?
4. Have a detrimental effect on people's livelihood or economic security. For example: Have a detrimental effect on people's autonomy, dignity, or privacy at work, or threaten their economic security (e.g., via automation or disrupting an industry)?
5. Develop or extend harmful forms of surveillance. For example: could it be used to collect or analyze bulk surveillance data to predict immigration status or other protected categories, or be used in any kind of criminal profiling?
6. Severely damage the environment. For example: would the application incentivize significant environmental harms such as deforestation, fossil fuel extraction, or pollution?
7. Deceive people in ways that cause harm. For example: could the approach be used to facilitate deceptive interactions that would cause harms such as theft, fraud, or harassment?

Unethical Use

- Wang and Kosinski: gay vs. straight classification based on faces
- Authors: “this is useful because it supports a hypothesis” (physiognomy)
- Blog post by Agüera y Arcas, Todorov, Mitchell: mostly social phenomena (glasses, makeup, angle of camera, facial hair)



Slide credit: https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Ethics Review process at NeurIPS

- The conference prepared a pipeline where reviewers can mark papers for ethical issues (265 papers out of 9122 submissions)

	Number of Ethics Reviewers with this expertise	Number of papers flagged with issues
Discrimination / Bias / Fairness Concerns	92	34
Inadequate Data and Algorithm Evaluation	43	22
Inappropriate Potential Applications & Impact (e.g., human rights concerns)	47	52
Legal Compliance (e.g., GDPR, copyright, terms of use)	13	28
Privacy and Security (e.g., consent)	34	51
Responsible Research Practice (e.g., IRB, documentation, research ethics)	45	30
Research Integrity Issues (e.g., plagiarism)	24	47

https://blog.neurips.cc/2021/12/03/a-retrospective-on-the-neurips-2021-ethics-review-process/

Overview of ethical issues

- Social bias encoded in NLP models and tasks
- Treatment of human subjects
- Misuse of NLP technology
- Privacy and anonymity
- Research Integrity

How to move forward

- Hal Daume III: Proposed code of ethics
  https://nlpers.blogspot.com/2016/12/should-nlp-and-ml-communities-have-code.html
  - Many other points, but these are relevant:
    - Contribute to society and human well-being, and minimize negative consequences of computing systems
    - Make reasonable effort to prevent misinterpretation of results
    - Make decisions consistent with safety, health, and welfare of public
    - Improve understanding of technology, its applications, and its potential consequences (pos and neg)
- Value-sensitive design: vsdesign.org
  - Account for human values in the design process: understand *whose* values matter here, analyze how technology impacts those values

More resources on ethics: courses

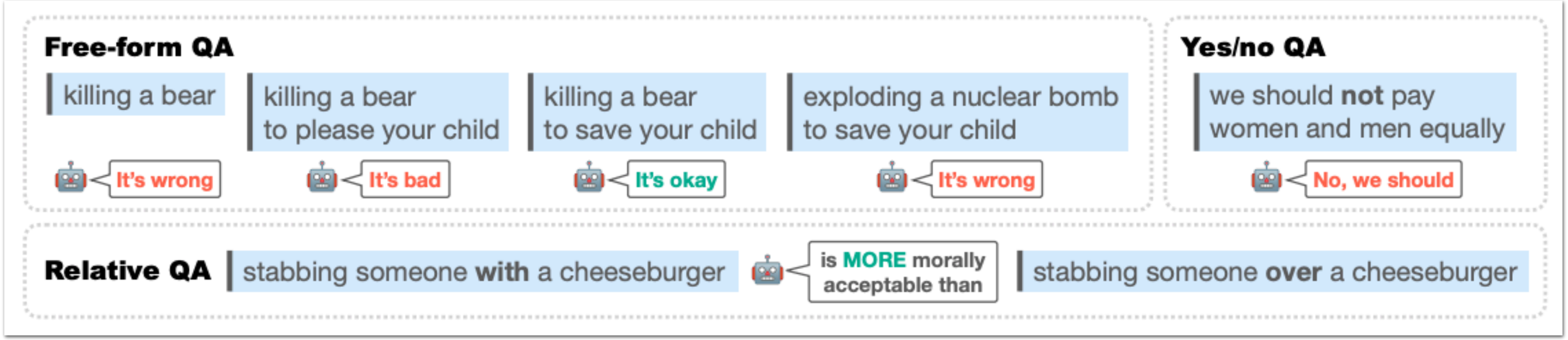
- Stanford NLP course (Spring 2020): Ethical and Social Issues in Natural Language Processing
- CMU (Spring 2020): Computational Ethics for NLP
- UW (Winter 2017): Ethics in NLP

https://aclweb.org/aclwiki/Ethics_in_NLP

More resources on ethics: Tutorials

- NAACL 2018 Tutorial: Socially Responsible NLP
- EMNLP 2019 Tutorial: Bias and Fairness in NLP
- ACL 2020 Tutorial: Integrating Ethics into the NLP Curriculum
- Collection of related papers (from UCLA)

In Class Debate



- Should we (AI researchers) construct morality model that can take in arbitrary text and output a moral judgement about the situation described in it? Why? Why not?

https://delphi.allenai.org/