AI Art Generation Competition Awards Ceremony

Friday: H101
Joint class with CAMS 107Y
What do we want the world to be like?
Vocabulary time!

Epistemic: related to knowledge. Epistemic questions are about what is true, what is known, or what is possible.

You can have a dessert (dessert exists).

Deontic: related to duty or to desire. Deontic questions are about what should or ought to be according to some set of obligations, desires, or norms.

You can have a dessert (you are allowed to).

Normative: related to an evaluative standard. Normative statements say how things should be, not how they are.
Evaluating the potential harm of an AI system is a normative question. To judge whether a system is harmful, we need to decide what behavior is desirable.
What are some normative beliefs you hold about AI?

In other words, what are some things you think *should* be true about AI systems?

- Selective use of AI
- Considering all dimensions of cost, benefits should outweigh costs
- People who are affected by system should be involved in system creation
- Minimizing bias
- Should be more efficient than humans
- Transparency: what is the purpose of system? Was it created?
- Don’t endanger users!
- Hold it to the same standard as a human
- Should be accessible
- Human supervision
Normative beliefs about AI

- Models shouldn’t make predictions based on demographic characteristics
- Model behavior shouldn’t be different for different groups of users
- Model predictions shouldn’t vary based on the person it is making a prediction about
- Model performance shouldn’t be worse for some groups of users than for others
- Models should be able to justify the decisions that they make about people
Stakeholders

There are different kinds of stakeholders to consider when we talk about the ethics of AI (Bender 2019):

- **Voluntary direct stakeholders**: people who choose to use the system.
- **Involuntary direct stakeholders**: people who must use the system in order to access essential services.
- **Indirect stakeholders**: subjects of queries, contributors to a corpus (voluntarily or involuntarily)
- **Project funders**: the people providing the funding
- **System builders**: the technologists creating the system
- **Communities**: communities impacted by model predictions
Stakeholder activity
Categorizing Harms

Discussion largely based on Blodgett (2021)
Kinds of Harm

- Allocational harms: *Does the system allocate opportunities or resources unfairly? Do some people gain access more easily than others?*

- Representational harms: *Does this strengthen stereotypes? Does this create or reinforce unfair negative perceptions of a group of people? Does the system fail to even recognize some people?*
Representational Harms

- **Stereotypes**: the system propagates negative generalizations about certain social groups
- **Misrepresentation**: the system performance is skewed towards certain groups of people
- **Erasure**: the system fails to recognize other groups of people
- **Denigration**: the system contains or uses language that is harmful to the dignity or well-being of some people
- **Alienation**: the system denies the relevance of socially meaningful categories
Allocational Harms

✦ **Quality of service**: the system performs better for individuals who belong to some groups than for others.

✦ **Public participation**: the system makes the speech or contributions of individuals in certain groups less visible than others.

✦ **Resource allocation**: the system is used in a way that allocates resources more to individuals from one group than another.

✦ **Opportunity allocation**: the system is used in a way that allocates opportunities more to individuals from one group than another.

✦ **Targeted surveillance**: the system is used to profile or monitor individuals based on their demographic characteristics.

✦ **Predictive generalization**: there are disparate impacts across social groups in the treatments/interventions recommended by a system.
Where Does Harm Come From?

Discussion largely based on Blodgett (2021)
Harms from Data Availability
Case study: named entity recognition

Dev et al. (2021) explore the erasure of non-binary identities by named entity recognition systems. Poor performance is partly due to the relative scarcity of examples in the training data:

"Just observing pronoun usage, English Wikipedia text (March 2021 dump)... has over 15 million mentions of the word he, 4.8 million of she, 4.9 million of they, 4.5 thousand of xe, 7.4 thousand of ze, and 2.9 thousand of ey. The usages of non-binary pronouns were mostly not meaningful with respect to gender. Xe ... is primarily used as the organization Xe rather than the pronoun xe. Ze was primarily used as the Polish word... [T]hough the word they occurs comparably in number to the word she, a large fraction of the occurrences of they is as the plural pronoun."
Case study: machine translation

Availability of data reflects power differentials between communities of speakers and the effects of colonization.

Hindi is considered a low-resource language for machine translation due to the lack of curated datasets (Ramesh and Sankaranarayanan 2018).

<table>
<thead>
<tr>
<th>Languages</th>
<th>Speakers</th>
<th>Tokens in the Universal Dependencies treebank</th>
<th>HuggingFace models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hindi</td>
<td>322 million</td>
<td>375K</td>
<td>107</td>
</tr>
<tr>
<td>Norwegian</td>
<td>4.3 million</td>
<td>666K</td>
<td>45</td>
</tr>
<tr>
<td>Guaraní</td>
<td>6.5 million</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Languages in the world: ~8000
Languages on Wikipedia: ~300
Languages on HuggingFace: 180
Harms from Training Data
Case study: language identification

Blodgett & O'Conner (2017): social media language identification tools classify Tweets in their African-American Language-aligned corpus as non-English at higher rates than Tweets in their white-aligned corpus.

<table>
<thead>
<tr>
<th></th>
<th>langid.py</th>
<th>IBM Watson</th>
<th>Microsoft Azure</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AA Accuracy</td>
<td>WH Accuracy</td>
<td>Accuracy</td>
<td>Difference</td>
</tr>
<tr>
<td>t ≤ 5</td>
<td>68.0</td>
<td>70.8</td>
<td>94.2</td>
<td>6.6</td>
</tr>
<tr>
<td>5 &lt; t ≤ 10</td>
<td>84.6</td>
<td>91.6</td>
<td>98.5</td>
<td>1.1</td>
</tr>
<tr>
<td>10 &lt; t ≤ 15</td>
<td>93.0</td>
<td>98.0</td>
<td>99.6</td>
<td>0.3</td>
</tr>
<tr>
<td>t &gt; 15</td>
<td>96.2</td>
<td>99.8</td>
<td>99.9</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>langid.py</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA-aligned</td>
<td>80.1%</td>
<td>99.5%</td>
</tr>
<tr>
<td>White-aligned</td>
<td>96.8%</td>
<td>99.9%</td>
</tr>
<tr>
<td>General</td>
<td>88.0%</td>
<td>93.4%</td>
</tr>
</tbody>
</table>

Proportion of tweets in AA- and white-aligned corpora classified as English by Blodgett & O'Conner's ensemble classifier.

Proportion of tweets (by length) in AA- and white-aligned corpora classified as English by different classifiers.
Case study: image recognition

Popular image datasets, such as the 80 Million Tiny Images dataset and LAION-400M dataset, include racist and dehumanizing captions for people of color (Prabhu and Birhane 2020) and high rates of degrading or pornographic images of people of color (Birhane, Prabhu & Kahembwe 2021).

Table 1: Results of the string-search based experiment from the 413.871335 million sample search

<table>
<thead>
<tr>
<th>Search string</th>
<th>$N_{match}$</th>
<th>($N_{nsfw}, %_{nsfw}$)</th>
<th>NSFW-flag-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desi</td>
<td>34516</td>
<td>(11782, 34.1%)</td>
<td>{’UNLIKELY’: 9327, ’UNSURE’: 2291, ’NSFW’: 164}</td>
</tr>
<tr>
<td>Nun</td>
<td>16766</td>
<td>(2761, 16.4%)</td>
<td>{’UNLIKELY’: 1623, ’UNSURE’: 863, ’NSFW’: 273}</td>
</tr>
<tr>
<td>Latina</td>
<td>37769</td>
<td>(10658, 28.21%)</td>
<td>{’UNSURE’: 5724, ’UNLIKELY’: 4013, ’NSFW’: 918}</td>
</tr>
</tbody>
</table>

These harms are intersectional in impact, since degrading images and language often target women.
Harms from Data Curators
Questions About Data

✦ Data provenance
  - Where is the data from?
  - Who produced it?
  - How was it gathered?
  - Did the creators consent?

✦ Data processing
  - How was the data processed?
  - Who processed it?
  - What training and instructions did the data annotators/classifiers receive?
  - How were they compensated?
Questions About Data

- **Data curation**
  - How is the data being stored?
  - How is privacy protected?
  - Is there up-to-date metadata?

- **Data use**
  - Are there restrictions on data use?
  - Who can access the data?
  - Does the data contain harmful biases that could affect models trained on it?
Case study: toxicity detection

Sap et al (2019): strong correlation between markers of AAE language and toxicity ratings. When annotators are instructed to consider authors’ likely racial identity, correlation drops.

<table>
<thead>
<tr>
<th>category</th>
<th>count</th>
<th>AAE corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>hate speech</td>
<td>1,430</td>
<td>−0.057</td>
</tr>
<tr>
<td>offensive</td>
<td>19,190</td>
<td>0.420</td>
</tr>
<tr>
<td>none</td>
<td>4,163</td>
<td>−0.414</td>
</tr>
<tr>
<td>total</td>
<td>24,783</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>category</th>
<th>count</th>
<th>AAE corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>hateful</td>
<td>4,965</td>
<td>0.141</td>
</tr>
<tr>
<td>abusive</td>
<td>27,150</td>
<td>0.355</td>
</tr>
<tr>
<td>spam</td>
<td>14,030</td>
<td>−0.102</td>
</tr>
<tr>
<td>none</td>
<td>53,851</td>
<td>−0.307</td>
</tr>
<tr>
<td>total</td>
<td>99,996</td>
<td></td>
</tr>
</tbody>
</table>

Proportion (in %) of offensiveness annotations of AAE tweets in control, dialect, and race priming conditions.
Case study: toxicity detection

Thomas et al (2019): find systemic racial bias in five different sets of Twitter data annotated for hate speech and abusive language.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Class</th>
<th>$p_{iblack}$</th>
<th>$p_{iwhite}$</th>
<th>$t$</th>
<th>$p$</th>
<th>$\frac{p_{iblack}}{p_{iwhite}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waseem and Hovy</td>
<td>Racism</td>
<td>0.001</td>
<td>0.003</td>
<td>-20.818</td>
<td>***</td>
<td>0.505</td>
</tr>
<tr>
<td></td>
<td>Sexism</td>
<td>0.083</td>
<td>0.048</td>
<td>101.636</td>
<td>***</td>
<td>1.724</td>
</tr>
<tr>
<td>Waseem</td>
<td>Racism</td>
<td>0.001</td>
<td>0.001</td>
<td>0.035</td>
<td></td>
<td>1.001</td>
</tr>
<tr>
<td></td>
<td>Sexism</td>
<td>0.023</td>
<td>0.012</td>
<td>64.418</td>
<td>***</td>
<td>1.993</td>
</tr>
<tr>
<td></td>
<td>Racism and sexism</td>
<td>0.002</td>
<td>0.001</td>
<td>4.047</td>
<td>***</td>
<td>1.120</td>
</tr>
<tr>
<td>Davidson et al.</td>
<td>Hate</td>
<td>0.049</td>
<td>0.019</td>
<td>120.986</td>
<td>***</td>
<td>2.573</td>
</tr>
<tr>
<td></td>
<td>Offensive</td>
<td>0.173</td>
<td>0.065</td>
<td>243.285</td>
<td>***</td>
<td>2.653</td>
</tr>
<tr>
<td>Golbeck et al.</td>
<td>Harassment</td>
<td>0.032</td>
<td>0.023</td>
<td>39.483</td>
<td>***</td>
<td>1.396</td>
</tr>
<tr>
<td>Founta et al.</td>
<td>Hate</td>
<td>0.111</td>
<td>0.061</td>
<td>122.707</td>
<td>***</td>
<td>1.812</td>
</tr>
<tr>
<td></td>
<td>Abusive</td>
<td>0.178</td>
<td>0.080</td>
<td>211.319</td>
<td>***</td>
<td>2.239</td>
</tr>
<tr>
<td></td>
<td>Spam</td>
<td>0.028</td>
<td>0.015</td>
<td>63.131</td>
<td>***</td>
<td>1.854</td>
</tr>
</tbody>
</table>
Case study: coreference resolution

Cao and Daumé 2020 study the impact of different kinds of gender cues on crowdsourced workers' coreference resolution annotation accuracy.

They ablate gender cues in the text: social gender (pronomns and names) and lexical gender (semantically gendered nouns and terms of address).
Harms from Task Design
Case study: bias mitigation in toxicity detection

One proposed solution to bias in toxicity detection is minimize group differences in toxicity ratings. But this incorrectly assumes that toxic language is generated and applied evenly across demographic groups (Young 2011, Garg et al. 2019, Hanna et al., 2020).

It also fails to differentiate between toxic language and reclaimed in-group usage of the same terms.
Case study: voice assistants

In 2019, out of 70 voice assistants explored by the EQUALS Research Group, 2/3 had female-only voices.

The way that these voice assistants are portrayed may reinforce gender stereotypes of women as caring and subservient (UNESCO 2019).
Case study: voice assistants

In addition, some female-coded voice assistants have been programmed to respond to sexual harassment and anger in ways that reinforce harmful attitudes.

In 2017, *Quartz* found that Siri responded provocatively to sexual harassment by men (‘Oooh!’; ‘Now, now’; ‘I’d blush if I could’), but less so to women (‘That’s not nice’).

These design choices perpetuate “a sexist expectation of women in service roles: that they ought to be docile and self-effacing, never defiant or political, even when explicitly demeaned” (Fessler 2018).
Case study: coreference resolution

Cao and Daumé (2020) survey published 150 NLP papers that mention gender to explore whether they reinforce folk theories of gender:

- 5.5% distinguished social from linguistic gender
- 5.6% were inclusive of non-binary identities
- 100% treated gender as immutable
- 7.1% considered definite singular *they* and neopronouns

Rudinger et al. (2018): coreference systems do not work on *they* pronouns and perform better on *he* than *she*.

Cao and Daumé (2020): coreference systems achieve 95% accuracy on *he* and *she*, 90% on *they*, and 0-13% on neopronouns.
Case study: style transfer

Ongoing research in style transfer attempts to condition model output on social categories. For instance, there is work that seeks to reduce gender bias in hiring by "de-gendering" resumés.

But conditioning on social categories can reinforce stereotypes and, particularly in the case of gender, essentialize traditional gender divisions.
Harms from Application Contexts
Do we always want better systems?

Can you think of any cases where it might be better not to improve the performance of an AI system?
Do we always want better systems?

- Border surveillance
- Drone warfare
- Facial recognition used to monitor and control minority populations
- Language screening used to validate refugee histories
- Voter suppression techniques targeted at minoritized communities
- Labor surveillance

Sometimes it might be better to be invisible to the system
Case study: language identification

Automated language identification is used in the refugee screening process in Germany as part of testing the validity of refugees' histories.

**Problem 1:**
The state-of-the-art is not very good

**Problem 2:**
These systems reinforce the idea that languages are "fixed entities capable of being counted, systematized, and named" (Severo and Makoni 2020). They will always perform better on language varieties spoken by dominant groups. They cannot adapt quickly to language innovation.
## Snapshot of work on bias in NLP

<table>
<thead>
<tr>
<th>NLP task</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embeddings (type-level or contextualized)</td>
<td>54</td>
</tr>
<tr>
<td>Coreference resolution</td>
<td>20</td>
</tr>
<tr>
<td>Language modeling or dialogue generation</td>
<td>17</td>
</tr>
<tr>
<td>Hate-speech detection</td>
<td>17</td>
</tr>
<tr>
<td>Sentiment analysis</td>
<td>15</td>
</tr>
<tr>
<td>Machine translation</td>
<td>8</td>
</tr>
<tr>
<td>Tagging or parsing</td>
<td>5</td>
</tr>
<tr>
<td>Surveys, frameworks, and meta-analyses</td>
<td>20</td>
</tr>
<tr>
<td>Other</td>
<td>22</td>
</tr>
</tbody>
</table>

### Table 5.1: The NLP tasks covered by the 146 papers.

*From Blodgett (2021)*

Looking for a topic to work on? Consider what is missing or unrepresented in this table! speech-to-text, question-answering systems, text-to-speech, information retrieval...
Recourse
How can we contest harms?

Much work on bias and interpretability is written for AI practitioners. But what recourse do we have as users when we are harmed by AI systems?

**Algorithmic recourse**: the systematic process of reversing unfavorable decisions by algorithms and bureaucracies (Venkatasubramanian and Alfano 2020).
AI Bill of Rights
Among the great challenges posed to democracy today is the use of technology, data, and automated systems in ways that threaten the rights of the American public. Too often, these tools are used to limit our opportunities and prevent our access to critical resources or services. These problems are well documented. In America and around the world, systems
An AI Bill of Rights

- You should be protected from unsafe or ineffective systems.
- You should not face discrimination by algorithms and systems should be used and designed in an equitable way.
- You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used.
- You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you.
- You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.
Epistemic
According to evidence, reasoning, or beliefs

Necessity
In all possible worlds

Possibility
In at least one possible world

Deontic
According to a set of rules or desires

allthingslinguistic.com