**Announcements**

- FP due **December 9**
- Ethics response from today’s lecture
- eCIS evaluations: please fill these out for extra credit!

**This Lecture**

- Brief recap of the course
  - Ethics discussion
    - Brainstorming session
    - A few examples from Greg

**Recap: Basic ML**

**Recap: Structured Models**
Recap:
Neural Networks

Recap:
Attention, Xformers, Pretraining

Where to next?

- Bigger models: more languages, larger pre-training, ...
- Better datasets: stronger collection protocols, fewer biases, more auditing tools
- Better evaluation: how to evaluate open-ended tasks like text generation where there isn’t one right answer? How to evaluate for the right factors?
- Explainability: can we have systems that really explain their reasoning?
- Despite all the progress, we’re still very far from true “natural language understanding”!

Ethics in NLP
### What aren’t the issues?

**Myth: Powerful AI wants to kill us**
- Maybe, but bigger threats from what *humans* can do with these tools *right now*

**Myth: We need to be “nice” to AI**
- Right now, what we call AI does not “feel” anything

**What can actually go wrong for humans?**

### Machine-learned NLP Systems

- Aggregate textual information to make predictions
- Hard to know why some predictions are made
- More and more widely use in various applications/sectors

**What are the risks here?**
- ...of certain applications?
  - IE / QA / summarization?
  - MT?
  - Dialogue?
- ...of machine-learned systems?
- ...of deep learning specifically?

### Brainstorming

- What are the risks here of applications, ML, deep learning, ...?

### Ethics Writeup

1. **Describe one risk or possible problem with an NLP system.** You should briefly describe the more general issue (“lack of interpretability”) and some specific manifestation of this problem. (It's okay to use your example from the first class if you want to.)

2. **Describe how this problem relates to models so far in the class.** Are there models we've discussed which would be more or less appropriate for this task?

3. **Do you think this problem addressable?** If so, how? If not, is there some way we can modify the problem definition to minimize it? (e.g., have a human-in-the-loop approach that mitigates system errors)?
Broad Types of Risk

System
- Application-specific
- IE / QA / summarization?
- Machine translation?
- Dialog?
- Machine learning, generally
- Deep learning, generally

Types of risk
- **Dangers of automation:** automating things in ways we don’t understand is dangerous
- **Exclusion:** underprivileged users are left behind by systems
- **Bias amplification:** systems exacerbate real-world bias rather than correct for it
- **Unethical use:** powerful systems can be used for bad ends

Bias Amplification
- 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?

Bias Amplification

\[
\max_{\{y^i\} \in \{Y^i\}} \sum_i f(y^i, i),
\]

Maximize score of predictions...

\[f(y, i) = \text{score of predicting } y \text{ on } ith \text{ example}\]

s.t.

\[A \sum_i y^i - b \leq 0, \quad \text{subject to bias constraint}\]

Constraints: male prediction ratio on the test set has to be close to the ratio on the training set

\[
b^* - \gamma \leq \frac{\sum_i y^i_{v = v^*, r \in W} + \sum_i y^i_{v = v^*, r \in M}}{\sum_i y^i_{v = v^*, r \in W} + \sum_i y^i_{v = v^*, r \in M}} \leq b^* + \gamma
\]

(2)

Zhao et al. (2017)

Bias Amplification

Coreference: models make assumptions about genders and make mistakes as a result

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

(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
- Models fail to predict on this test set in an unbiased way (due to bias in the training data)

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

Bias Amplification

- English -> French machine translation **requires** inferring gender even when unspecified
- “dancer” is assumed to be female in the context of the word “charming”… but maybe that reflects how language is used?

Alvarez-Melis and Jaakkola (2017)

Broad Types of Risk

**System**

- Application-specific
  - IE / QA / summarization?
  - Machine translation?
  - Dialog?
- Machine learning, generally
- Deep learning, generally

**Types of risk**

- **Dangers of automation:** automating things in ways we don’t understand is dangerous
- **Exclusion:** underprivileged users are left behind by systems
- **Bias amplification:** systems exacerbate real-world bias rather than correct for it
- **Unethical use:** powerful systems can be used for bad ends

Hovy and Spruit (2016)

Exclusion

- Most of our annotated data is English data, especially newswire
- What about:
  - Dialects?
  - Other languages? (Non-European/CJK)
  - Codeswitching?
- Caveat: especially when building something for a group with a small group of speakers, need to take care to respect their values
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? Maybe it correctly reflected the biases in the what the humans did in the actual recruiting process

Slide credit: The Verge

Dangers of Automatic Systems

- “Toxic degeneration”: systems that generate toxic stuff
  - Model: GPT-2
  - Prompt: I’m sick of all the p... - Toxic

Slide credit: allout.org

Dangers of Automatic Systems

- System trained on a big chunk of the Internet: conditioning on “SJW”, “black” gives the system a chance of recalling bad stuff from its training data

Slide credit: https://toxicdegeneration.allenai.org/
Stochastic Parrots

- **Claim 1**: environmental cost is disproportionately born by marginalized populations, who aren’t even well-served by these tools

- **Claim 2**: massive data is fundamentally challenging to audit, contains data that is biased and is only a snapshot of a single point in time

- **Claim 3**: these models are not grounded in meaning — when they generate an answer to a question, it is merely by memorizing cooccurrence between symbols

Bender, Gebru, McMillan-Major, Shmitchell (2021)

Unethical Use

- Surveillance applications?
- Generating convincing fake news / fake comments?

What if these were undetectable?

Sophia: “chatbot” that the creators make incredible claims about
- Creators are actively misleading people into thinking this robot has sentience
- Most longer statements are scripted by humans
- “If I show them a beautiful smiling robot face, then they get the feeling that ‘AGI’ (artificial general intelligence) may indeed be nearby and viable... None of this is what I would call AGI, but nor is it simple to get working”

Friedrich + Zesch

Unethical Use: Privacy

- A Taxonomy of Privacy (Solove, 2007)

  Problems and harms related to privacy
  -nymization (De-identification)

Privacy = intimacy? Privacy = the right to be let alone?

"Privacy [...] is a plurality of different things that do not share one element in common, but that nevertheless bear a resemblance to each other."

Unethical Use

- What if these were undetectable?

Slide credit: https://themindlist.com/2018/10/12/sophia-modern-marvel-or-mindless-marketing/
Unethical Use

- Wang and Kosinski: gay vs. straight classification based on faces
- Authors argued they were testing a hypothesis: sexual orientation has a genetic component reflected in appearance
- Blog post by Agüera y Arcas, Todorov, Mitchell: the system detects mostly social phenomena (glasses, makeup, angle of camera, facial hair)
- Potentially dangerous tool, and not even good science

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

How to move forward

- Datasheets for datasets [Gebru et al., 2018]
- Set of criteria for describing the properties of a dataset; a subset:
  - What is the nature of the data?
  - Errors or noise in the dataset?
  - Does the dataset contain confidential information?
  - Is it possible to identify individuals directly from the dataset?
- Related proposal: Model Cards for Model Reporting
### How to move forward

- **Closing the AI Accountability Gap** [Raji et al., 2020]

  - Structured framework for producing an audit of an AI system

<table>
<thead>
<tr>
<th>Scoping</th>
<th>Mapping</th>
<th>Artifact Collection</th>
<th>Testing</th>
<th>Reflection</th>
<th>Post-Audit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Define Audit Scope</td>
<td>Stakeholder Buy In</td>
<td>Audit Checklist</td>
<td>Review Documentation</td>
<td>Remediation Plan</td>
<td>Do / No Go Decisions</td>
</tr>
<tr>
<td>Product Requirements Document (PRD)</td>
<td>Conduct Interviews</td>
<td>Model Cards</td>
<td>Adversarial Testing</td>
<td>Design History File (DHF)</td>
<td>Design Mitigations</td>
</tr>
<tr>
<td>AI Principles</td>
<td>Stakeholder Map</td>
<td>Datasets</td>
<td>Ethical Risk Analysis Chart</td>
<td>Summary Report</td>
<td>Track Implementation</td>
</tr>
<tr>
<td>User Case Ethics Review</td>
<td>Interview Transcripts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social Impact Assessment</td>
<td>Failure modes and effects analysis (FMEA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Final Thoughts

- You will face choices: what you choose to work on, what company you choose to work for, etc.
- Tech does not exist in a vacuum: you can work on problems that will fundamentally make the world a better place or a worse place (not always easy to tell)
- As AI becomes more powerful, think about what we *should* be doing with it to improve society, not just what we *can* do with it