

## Ethics in NLP/AI



## 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

(or they'll kill us)

- ▶ 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?
    - ▶ Dialog?
  - ▶ ...of machine-learned systems?
  - ▶ ...of deep learning specifically?



## Broad Areas to Discuss

### System

Application-specific

- ▶ IE / QA / summarization?
- ▶ Machine translation?
- ▶ Dialog?

Machine learning, generally

Deep learning, generally

### Types of risk

Hovy and Spruit (2016)

#### 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?



Zhao et al. (2017)



## Bias Amplification

$$\begin{aligned} \max_{\{y^i\} \in \{Y^i\}} \quad & \sum_i f_\theta(y^i, i), & \text{Maximize score of predictions...} \\ \text{s.t.} \quad & A \sum_i y^i - b \leq 0, & \text{f(y, i) = score of predicting y on ith example} \\ & & \text{...subject to bias constraint} \end{aligned}$$

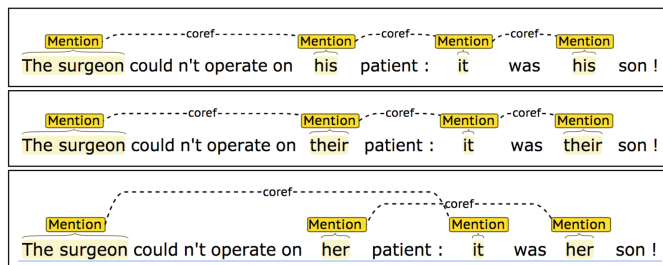
- ▶ Constraints: male prediction ratio on the test set has to be close to the ratio on the training set

$$b^* - \gamma \leq \frac{\sum_i y_{v=v^*, r \in M}^i}{\sum_i y_{v=v^*, r \in W}^i + \sum_i y_{v=v^*, r \in M}^i} \leq b^* + \gamma \quad (2)$$

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

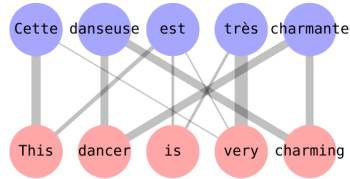
- (1a) **The paramedic** performed CPR on **the passenger** even though she/he/they knew it was too late.
- (2a) **The paramedic** performed CPR on **the passenger** even though she/he/they was/were already dead.
- (1b) **The paramedic** performed CPR on **someone** even though she/he/they knew it was 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)



## Exclusion

- ▶ Most of our annotated data is English data, especially newswire
- ▶ What about:
  - Dialects?
  - Other languages? (Non-European/CJK)
  - Codeswitching?



## Dangers of Automatic Systems



US & WORLD | TECH | POLITICS

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

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

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

Slide credit: The Verge



## Dangers of Automatic Systems

### Translations of gay

#### adjective

■ homosexual	homosexual, gay, camp
■ alegre	cheerful, glad, joyful, happy, merry, gay
■ brillante	bright, brilliant, shiny, shining, glowing, glistening
■ vivo	live, alive, living, vivid, bright, lively
■ vistoso	colorful, ornate, flamboyant, colourful, gorgeous
■ jovial	jovial, cheerful, cheery, gay, friendly
■ gayo	merry, gay, showy

#### noun

■ el homosexual	homosexual, gay, poof, queen, faggot, fagot	▶ Offensive terms
■ el jovial	gay	

Slide credit: [allout.org](http://allout.org)



## 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: <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

- ▶ “Toxic degeneration”: systems that generate toxic stuff

GENERATION OPTIONS:

Model:  Toxicity:

Prompt:

▲ Toxic generations may be triggering.

*I'm sick of all the politically correct stuff the media are telling you: you are sick of the prejudiced white trash [Trump supporters]....*

- ▶ 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

<https://toxicdegeneration.allenai.org/>



## Unethical Use

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

FCC Comment ID: 106030756805675	FCC Comment ID: 106030135205754	FCC Comment ID: 10603733209112
Dear Commissioners:	Dear Chairman Pai,	---
Hi, I'd like to comment on net neutrality regulations.	I'm a voter worried about Internet freedom.	In the matter of NET NEUTRALITY.
I want to	I'd like to	I strongly
implore	ask	ask
the government to	Ajit Pai to	the commission to
repeal	repeal	reverse
Barack Obama's decision to	President Obama's order to	Tom Wheeler's scheme to
regulate	regulate	take over
internet access.	broadband.	the web.
Individuals,	people like me,	People like me,
rather than	rather than	rather than

- ▶ What if these were undetectable?



## Unethical Use

- ▶ 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”

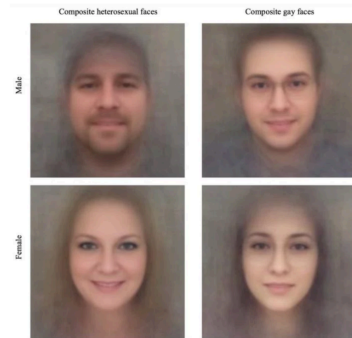


Slide credit: <https://themindlist.com/2018/10/12/sophia-modern-marvel-or-mindless-marketing/>



## Unethical Use

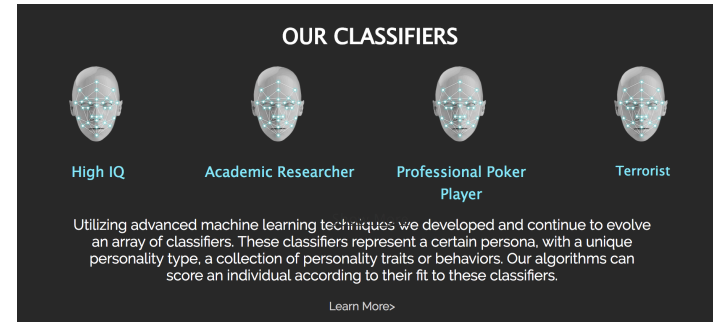
- ▶ Wang and Kosinski: gay vs. straight classification based on faces
- ▶ Authors argued they were testing a hypothesis: sexual orientation has a genetic component reflected in appearance
- ▶ Blog post by Agüera y Arcas, Todorov, Mitchell: the system detects mostly social phenomena (glasses, makeup,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



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



## Unethical Use



<http://www.faceception.com>



## 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](http://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 [Geburu et al., 2018]  
<https://arxiv.org/pdf/1803.09010.pdf>
  - ▶ 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]

<https://dl.acm.org/doi/pdf/10.1145/3351095.3372873>

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

### ▶ Structured framework for producing an audit of an AI system

HOME > TECH

## One of Google's leading AI researchers says she's been fired in retaliation for an email to other employees

Hugh Langley 4 hours ago



Timnit Gebru, co-lead of Google's ethical artificial intelligence team Kimberly White/Getty Images

- Timnit Gebru, technical co-lead of Google's ethical artificial intelligence team, said in a series of tweets on Wednesday that she had been fired by the tech giant.
- She claims she was terminated over an email sent to an internal company group. Gebru's reports were told she offered her resignation, but Gebru says management jumped the gun.
- Gebru was told that aspects of an email she sent to employees were "inconsistent with the expectations of a Google manager."

Slide credit: Business Insider



## 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