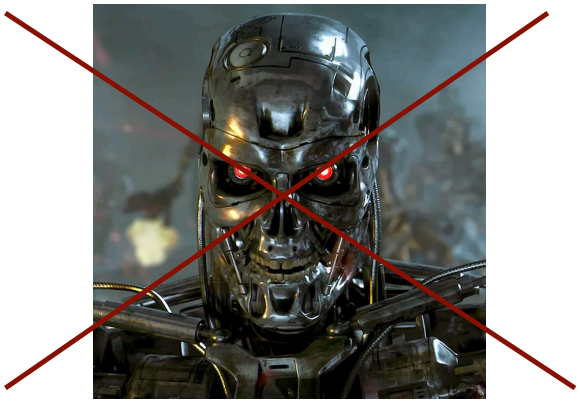
Ethics in NLP/AI



What can actually go wrong?

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

- Bias amplification: systems exacerbate real-world bias rather than correct for it
- Exclusion: underprivileged users are left behind by systems
- Dangers of automatic systems: automating things in ways we don't understand is dangerous
- 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

$$\max_{\{y^i\}\in\{Y^i\}} \sum_i f_\theta(y^i, i),$$

Maximize score of predictions…
f(y, i) = score of predicting y on ith example

$$\text{s.t.} \quad A\sum_i y^i - b \leq 0,$$

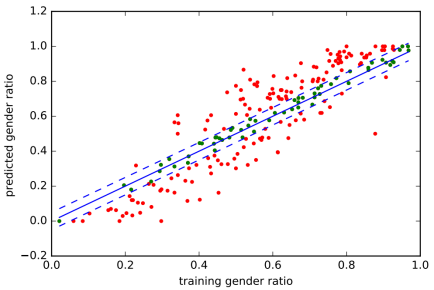
…subject to bias constraint

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

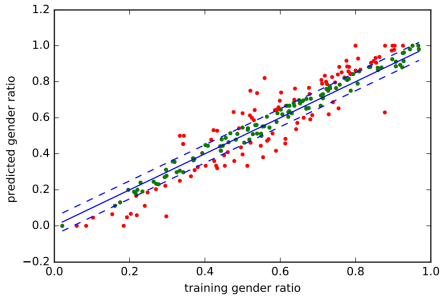
$$b^* - \gamma \leq \frac{\sum_i y^i_{v=v^*, r\in M}}{\sum_i y^i_{v=v^*, r\in W} + \sum_i y^i_{v=v^*, r\in M}} \leq b^* + \gamma \quad (2)$$

Zhao et al. (2017)

Bias Amplification



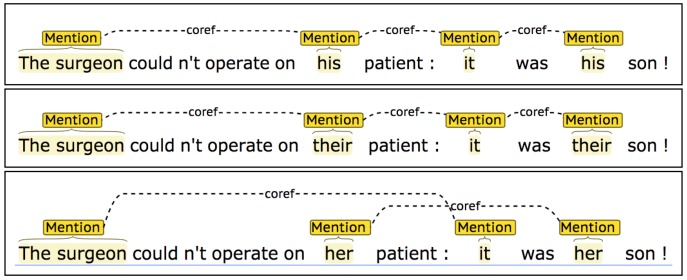
(a) Bias analysis on imSitu vSRL without RBA



(c) Bias analysis on imSitu vSRL with RBA

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.

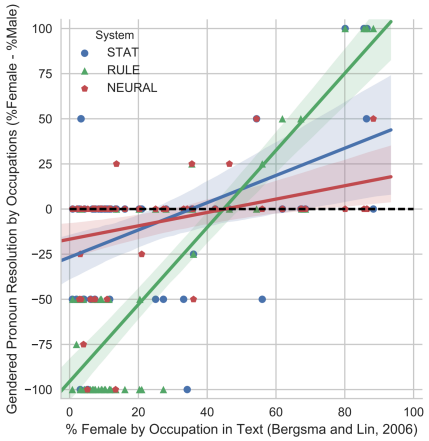
(1b) **The paramedic** performed CPR on someone even though she/he/they knew it was too late.

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

- Can form a targeted test set to investigate

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

Bias Amplification

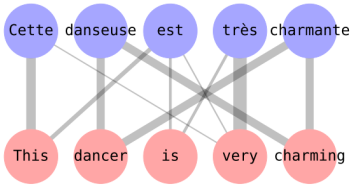


- Test set is balanced so a perfect model has female%-male% = 0 (black line)
- Neural models actually are a bit better at being unbiased, but are still skewed by data

Zhao et al. (2017)

Bias Amplification

- Harder to quantify this for machine translation
- "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

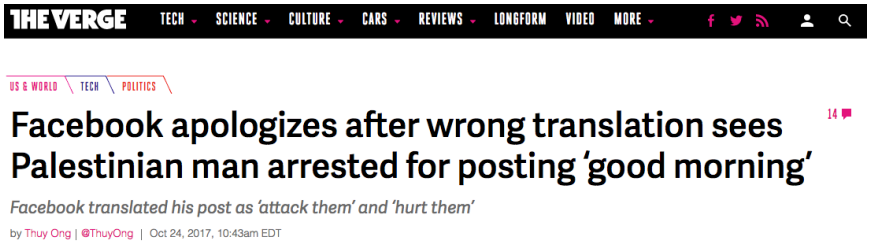
- Most of our annotated data is English data, especially newswire
- What about:

  Dialects?

  Other languages? (Non-European/CJK)

  Codeswitching?

Dangers of Automatic Systems



Slide credit: The Verge

Dangers of Automatic Systems

Offensive terms

Slide credit: allout.org

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

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

Unethical Use

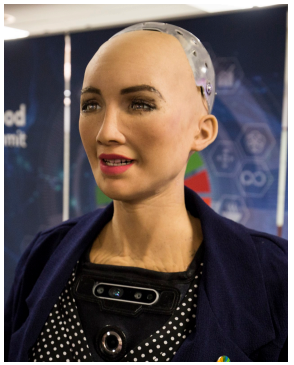
- Surveillance applications?
- Generating convincing fake news / fake comments?

FCC Comment ID: 106030756805675	FCC Comment ID: 106030135205754	FCC Comment ID: 10603733209112
Dear Commissioners:	Dear Chairman Pai,	—
Hi, I'd like to comment on	I'm a voter worried about	In the matter of
net neutrality regulations.	Internet freedom.	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	President Obama's	Tom Wheeler's
decision to	order to	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

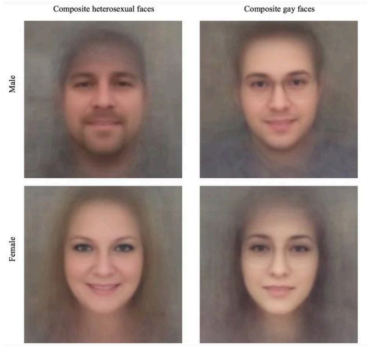
- Sophia: "chatbot" that the creators make incredible claims about
- Creators are actively misleading people into thinking this robot has sentience
- Most longer statements are scripted by humans
- "If I show them a beautiful smiling robot face, then they get the feeling that 'AGI' (artificial general 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/

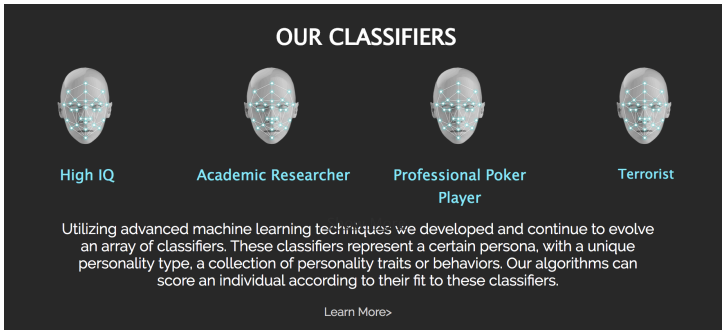
Unethical Use

- Wang and Kosinski: gay vs. straight classification based on faces
- Authors: "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

Unethical Use



http://www.faceception.com

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