CS388: Natural Language Processing

Lecture 27: Wrapup, Ethics

Greg Durrett

The University of Texas at Austin

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?

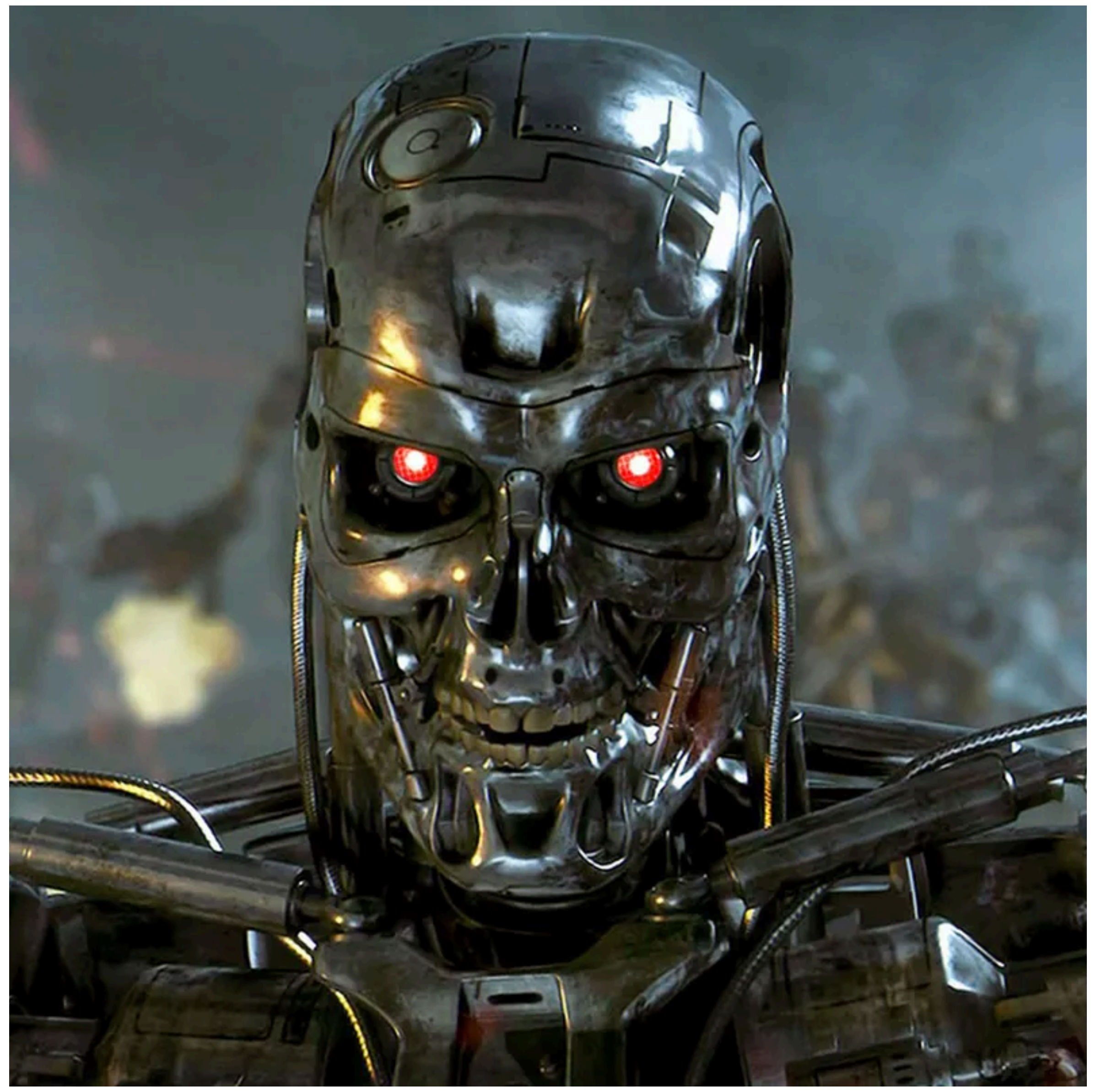
- Bigger models: more languages, larger pre-training, …
- Better datasets: stronger collection protocols, fewer biases, more auditing tools
- Better evaluation: how to evaluate open-ended tasks like text generation where there isn't one right answer? How to evaluate for the right factors?
- Explainability: 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 are we **not** discussing today?

Is powerful AI going to kill us?

- Maybe, lots of work on "x-risk" but a lot of this is philosophical and sort of speculative, hard to unpack with tools in this class
- Instead, let's think about more near-term harms that have already been documented



What can actually go wrong **for people, today**?

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?
  - …inherent in these system? E.g.: if they're unfair, what bad things can happen?
  - …of certain applications?
    - IE / QA / summarization?
    - MT?
    - Dialogue?

Brainstorming

- What are the risks here **inherent to these systems we've seen**? E.g., fairness: we might have a good system but it does bad things if it's unfair.

Brainstorming

- What are the risks here of **applications**? Misuse and abuse of NLP

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

Hovy and Spruit (2016)

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?

Zhao et al. (2017)

Bias Amplification

$$\max_{\{y^i\}\in\{Y^i\}} \sum_i f_\theta(y^i, i),$$
$$\text{s.t.} \quad A\sum_i y^i - b \leq 0,$$

Maximize score of predictions…

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

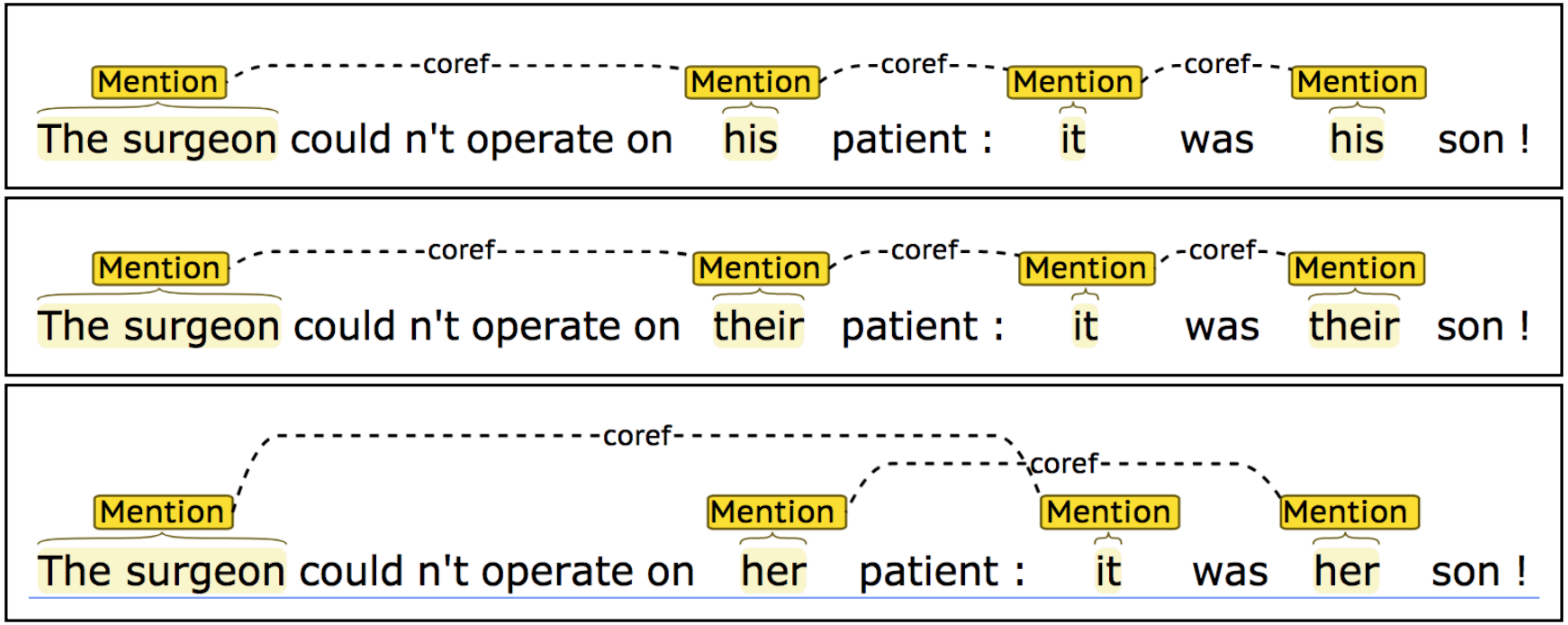
…subject to bias constraint

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

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

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.

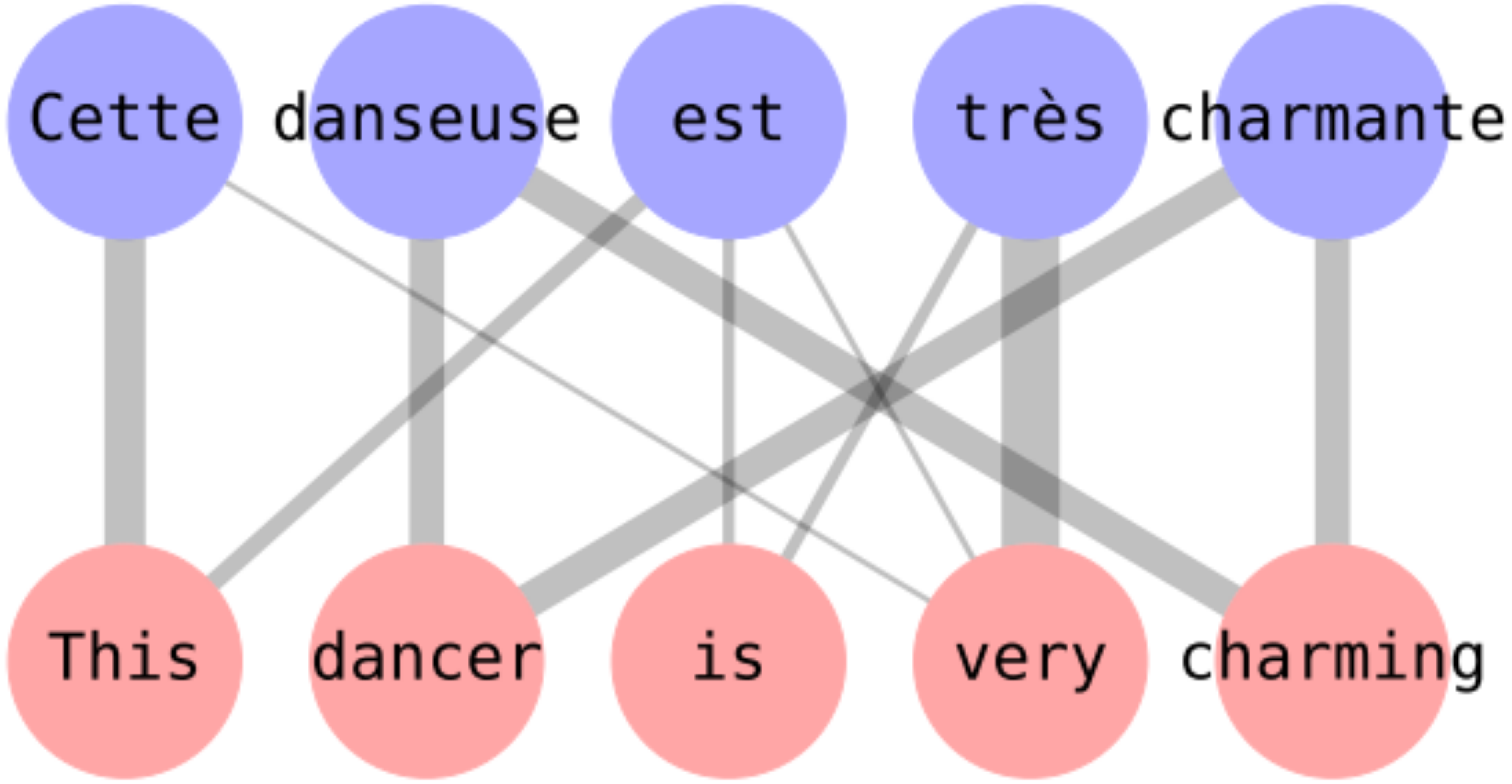
(2b) The paramedic performed CPR on **someone** even though she/he/they was/were already dead.

- 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

Hovy and Spruit (2016)

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

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

Large Language Models

Pizzle theory

Pizzle theory is a set of principles in software development that provide a conceptual framework for understanding the interaction of the people, process and technology in the development of a software system. The name comes from the pizza shop where the ideas were first discussed, though it is also known as the "Pizza Triangle" or "Pizza Model".

Contents

History

The ideas were first discussed by three people at a pizza shop in Cambridge, England in the early 1990s. The original three were Michael Jackson, Peter Lowe and Dave Thomas. Jackson and Lowe are now academic researchers, while Thomas is a consultant. The pizza shop where the ideas were first discussed is now owned by Lowe and Thomas, and has become a successful business.

The model

Nathan Hamiel
@nathanhamiel

I give you Pizzle theory, and Michael Jackson is involved! Great! Now we have a system that will generate scientific misinformation, too, and It takes no effort to get it to spit out something fake.
#GALACTICA galactica.org/?prompt=wiki+a...

Dangers of Automatic Systems

THE VERGE TECH SCIENCE CULTURE CARS REVIEWS LONGFORM VIDEO MORE

US & WORLD TECH POLITICS

14

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	$\boxed{\text{homosexual, gay, poof, queen, faggot, fagot}}$
el jovial	gay

Slide credit: allout.org

Dangers of Automatic Systems

- “Toxic degeneration”: systems that generate toxic stuff

- 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: Privacy

Anonymization (De-Identification)

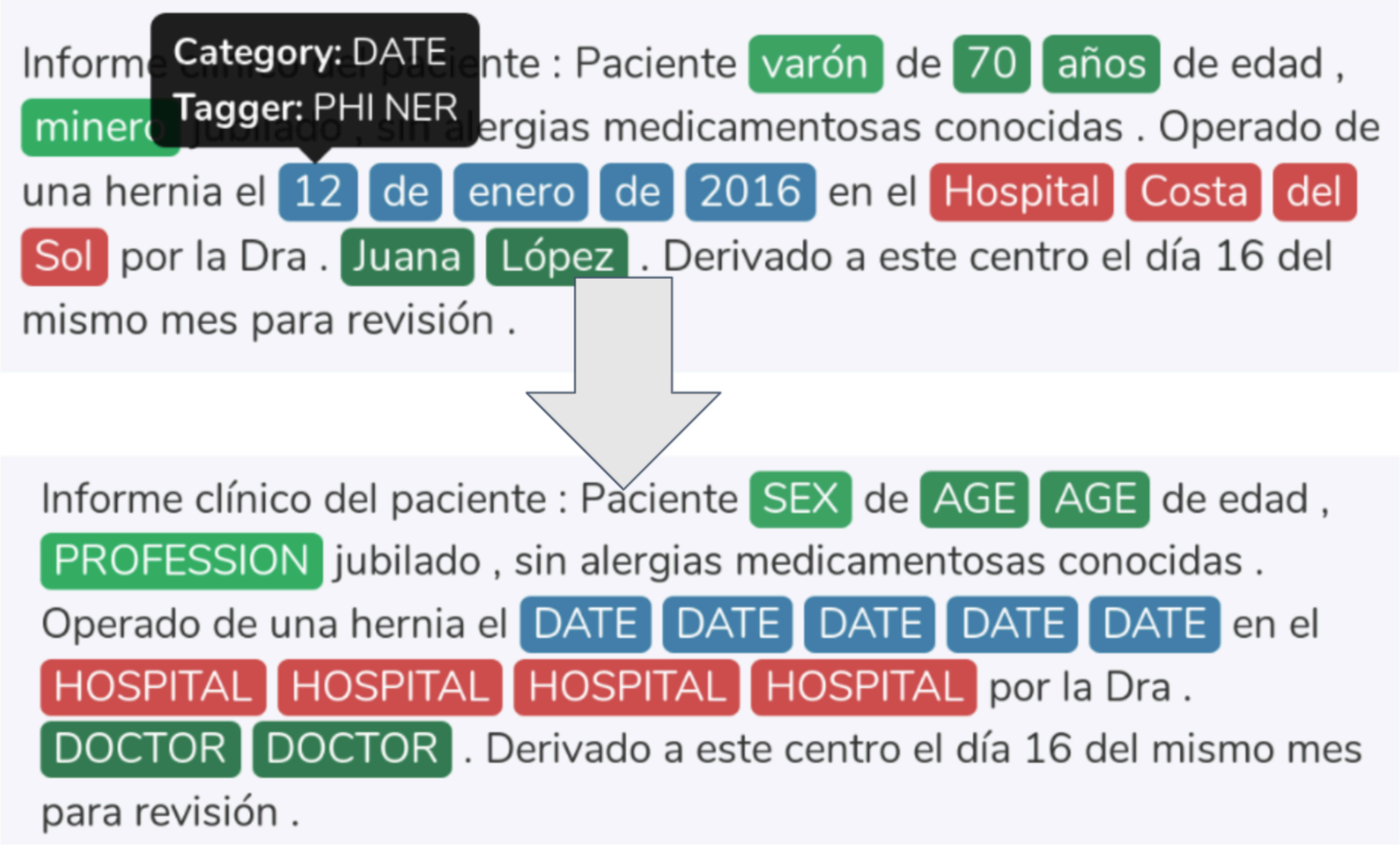


Image Source: https://www.aclweb.org/anthology/2020.lrec-1.870/

HitzalMed
(Lopez et al., 2020)

After having run some anonymization system on our data, is everything fine?

Friedrich + Zesch

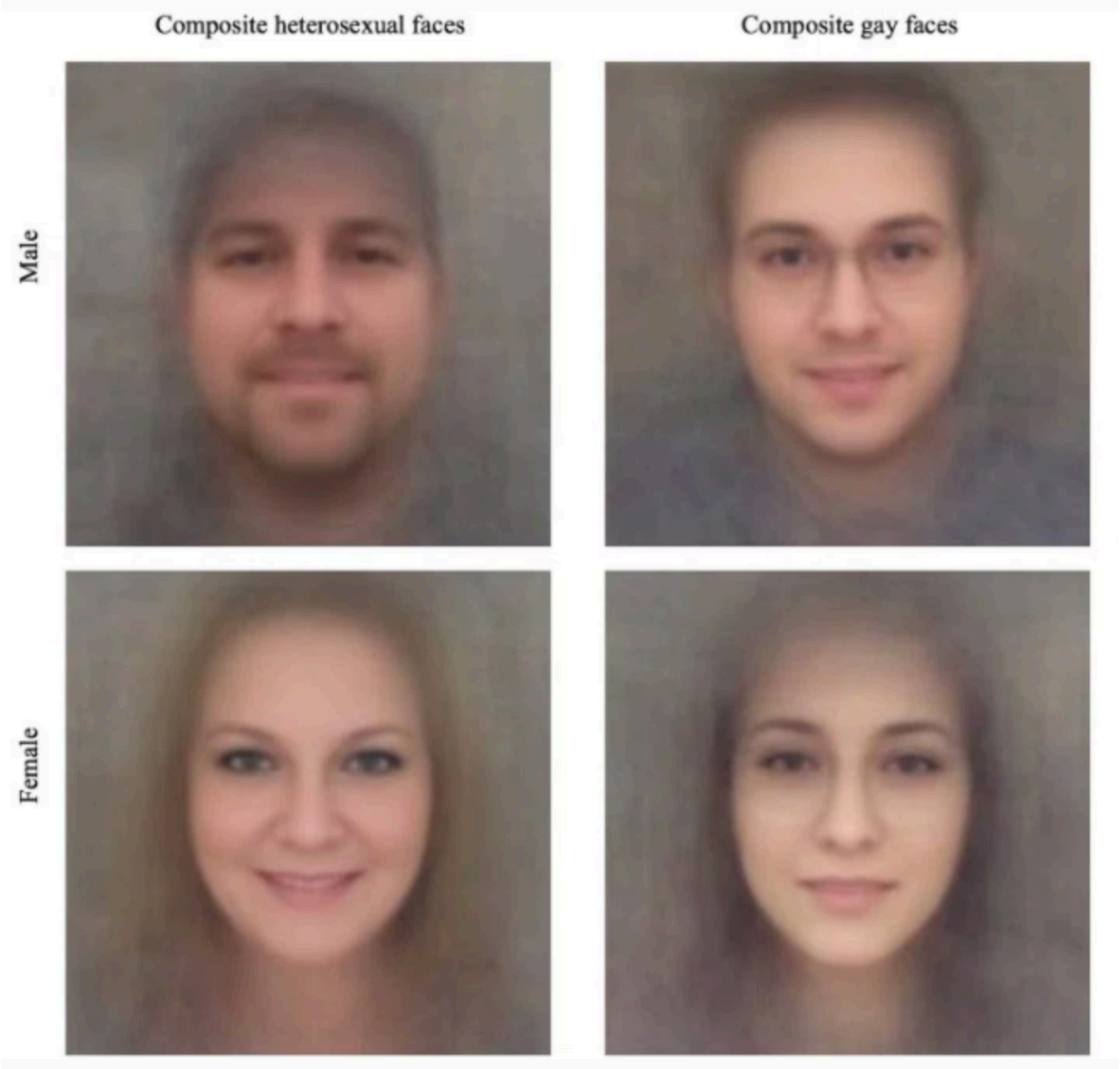
Unethical Use

- 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

(no longer exists)

http://www.faceception.com

How to move forward

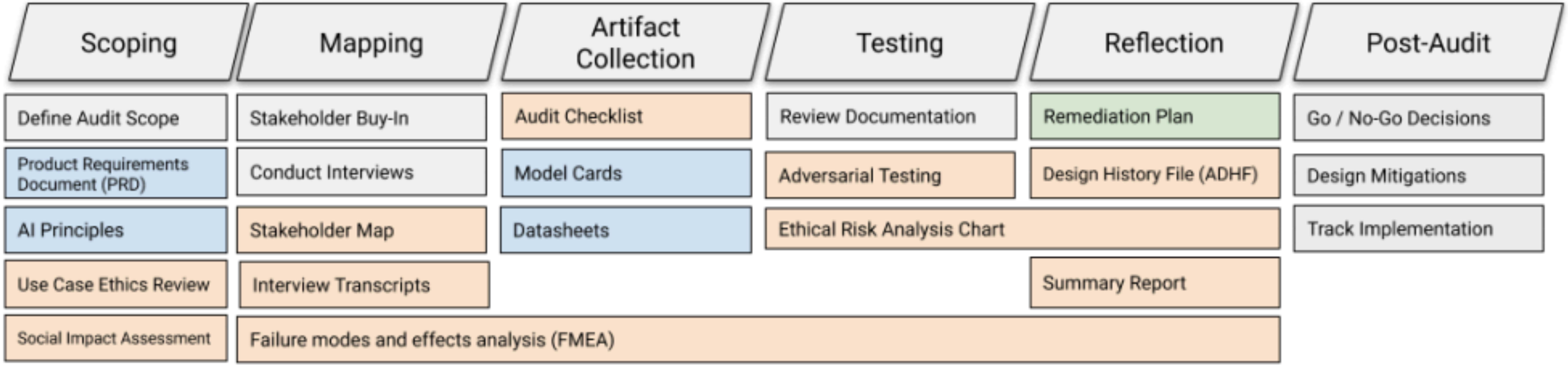
- Hal Daume III: Proposed code of ethics
  https://nlpers.blogspot.com/2016/12/should-nlp-and-ml-communities-have-code.html
  - Many other points, but these are relevant:
    - Contribute to society and human well-being, and minimize negative consequences of computing systems
    - Make reasonable effort 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
  - 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]
  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



  - Structured framework for producing an audit of an AI system

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