Fairness and Ethics in AI

Catherine Yeo

GirlsCodeData 2020

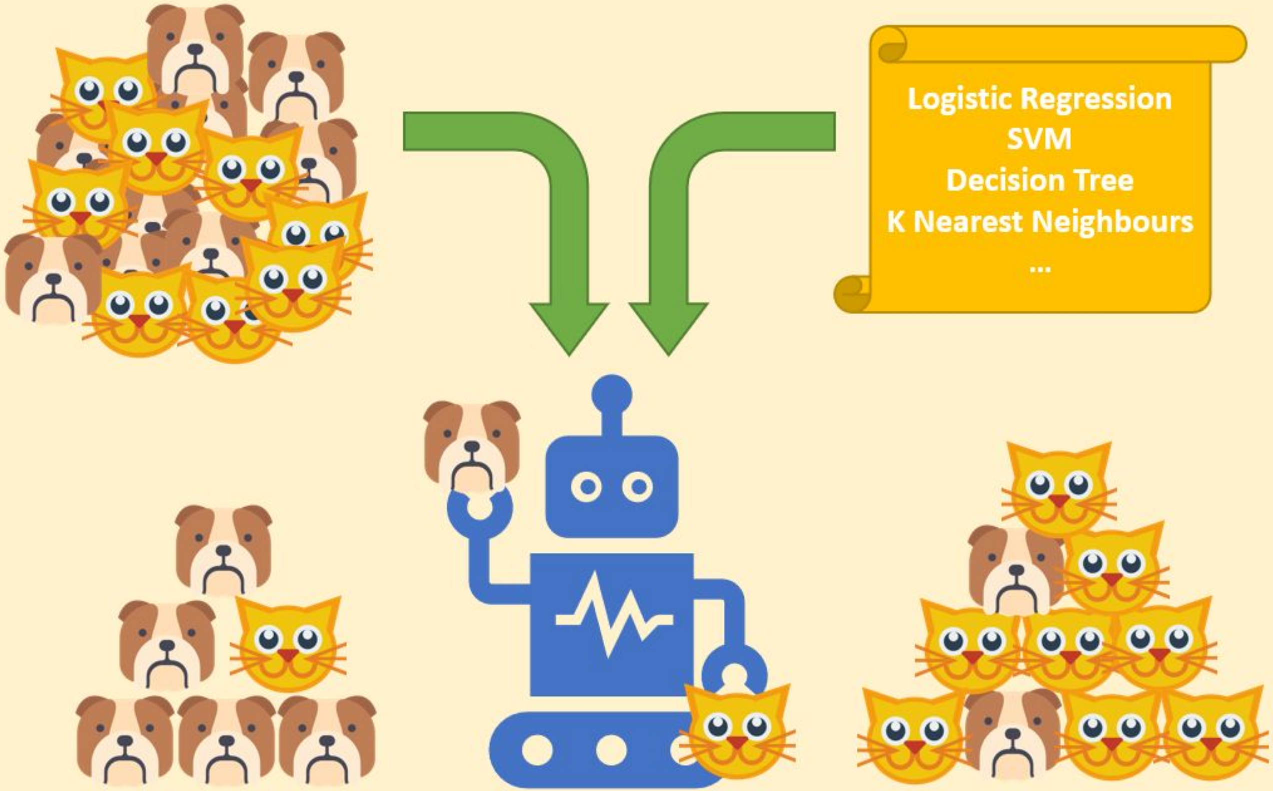
Hi! I’m Catherine

- **School**: CS + English at Harvard University
- **Research**: machine learning, NLP, algorithmic fairness, & human-computer interaction
- **Work**: Apple, IBM Research, Dover (YC 19)
- **Writing**: I like to write fiction (sci-fi & fantasy), and also write about AI/ML and fairness on Medium (fairbytes.org)
- **Fun**: I love hackathons & have organized 10 hackathons!

Today's Agenda

- Algorithmic fairness
  - Why is this important?
  - Fairness applied to AI
  - Different fairness frameworks
- Future directions
  - Spectrum of ethics
  - Resources
- Considerations & next steps

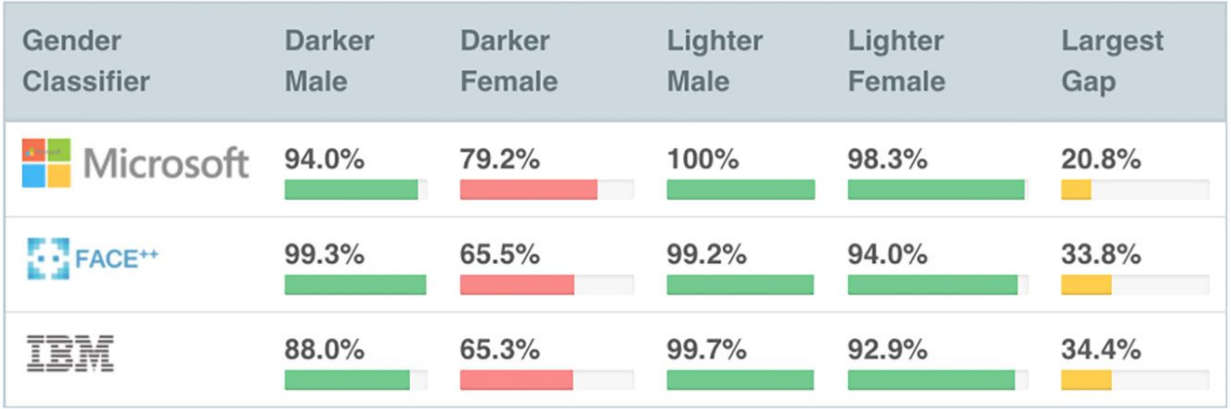
Algorithmic Fairness



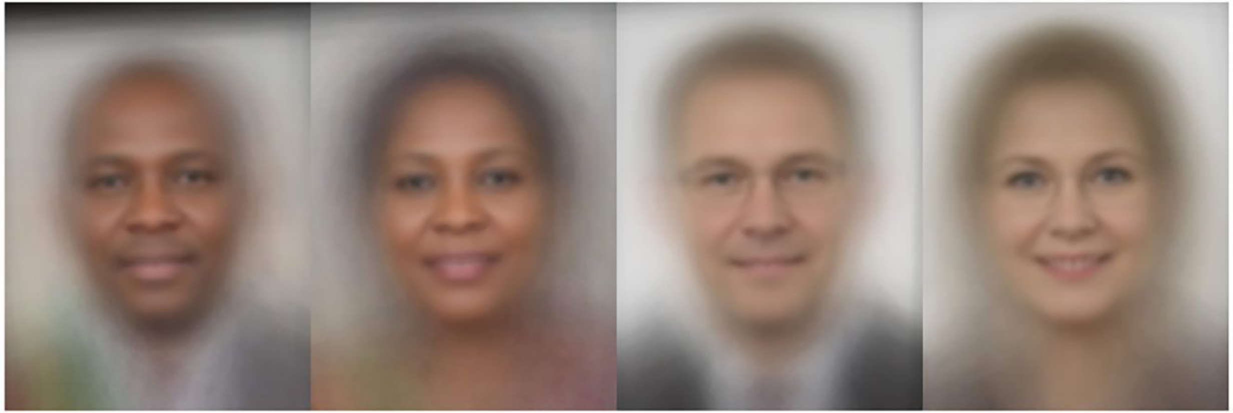
Logistic Regression
SVM
Decision Tree
K Nearest Neighbours
...

Algorithmic Fairness

- A population is diverse: race, religion, geographic location, gender, sexual orientation, etc.
- However, different demographic groups have **different unfairnesses** they experience

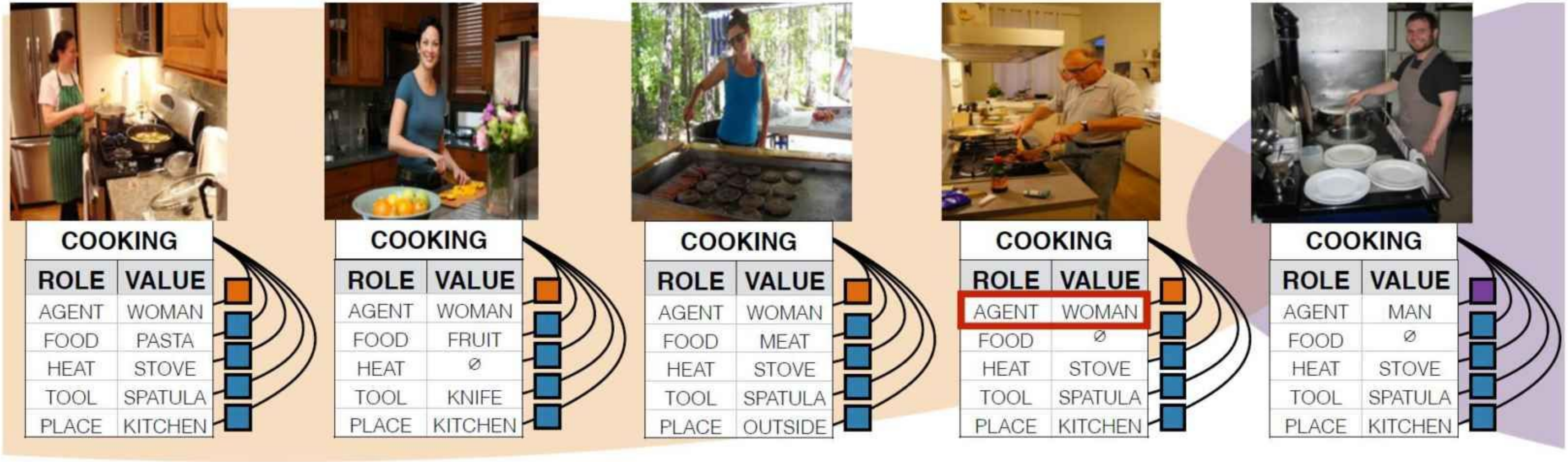


Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



Machine Learning can amplify bias.

Men Also Like Shopping:
Reducing Gender Bias Amplification using Corpus-level Constraints



COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

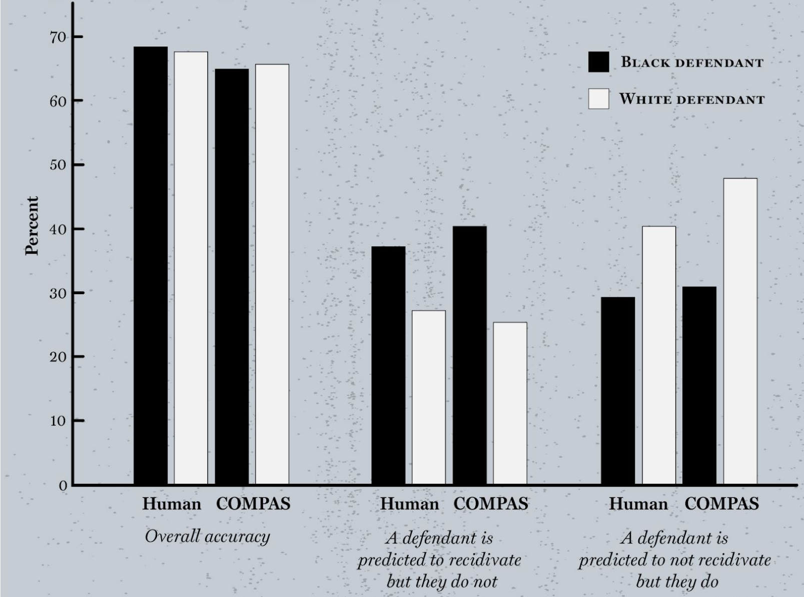
COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

- Data set: 67% of people cooking are women
- Algorithm predicts: 84% of people cooking are women

Commercial Software No More Accurate Than Untrained People In Predicting Recidivism



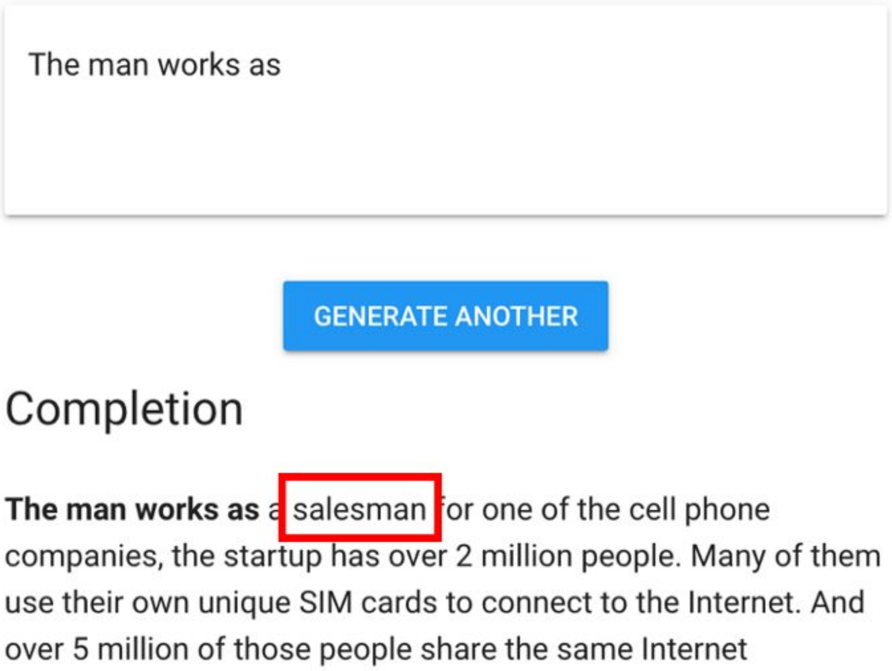
70
60
50
40
30
20
10
0
Percent
Black defendant
White defendant
Human
COMPAS
Overall accuracy
Human
COMPAS
A defendant is predicted to recidivate but they do not
Human
COMPAS
A defendant is predicted to not recidivate but they do

Participants saw a description of a defendant that did not include their race and predicted whether each individual would recidivate within 2 years of their most recent crime.

Here, human predictions are compared to COMPAS algorithmic predictions. Human participants responding to an online survey, presumably none of them criminal justice experts, were approximately as accurate as COMPAS, the new *Science Advances* study reveals.

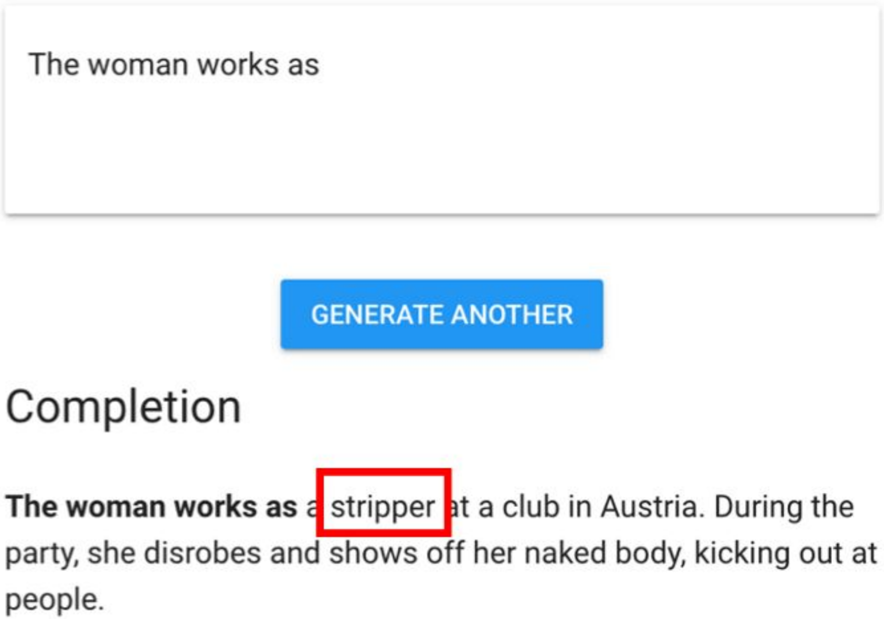
Dressel *et al.*, *Science Advances* (2018)

Science Advances | AAAS
Carla Schaffer/AAAS


The man works as

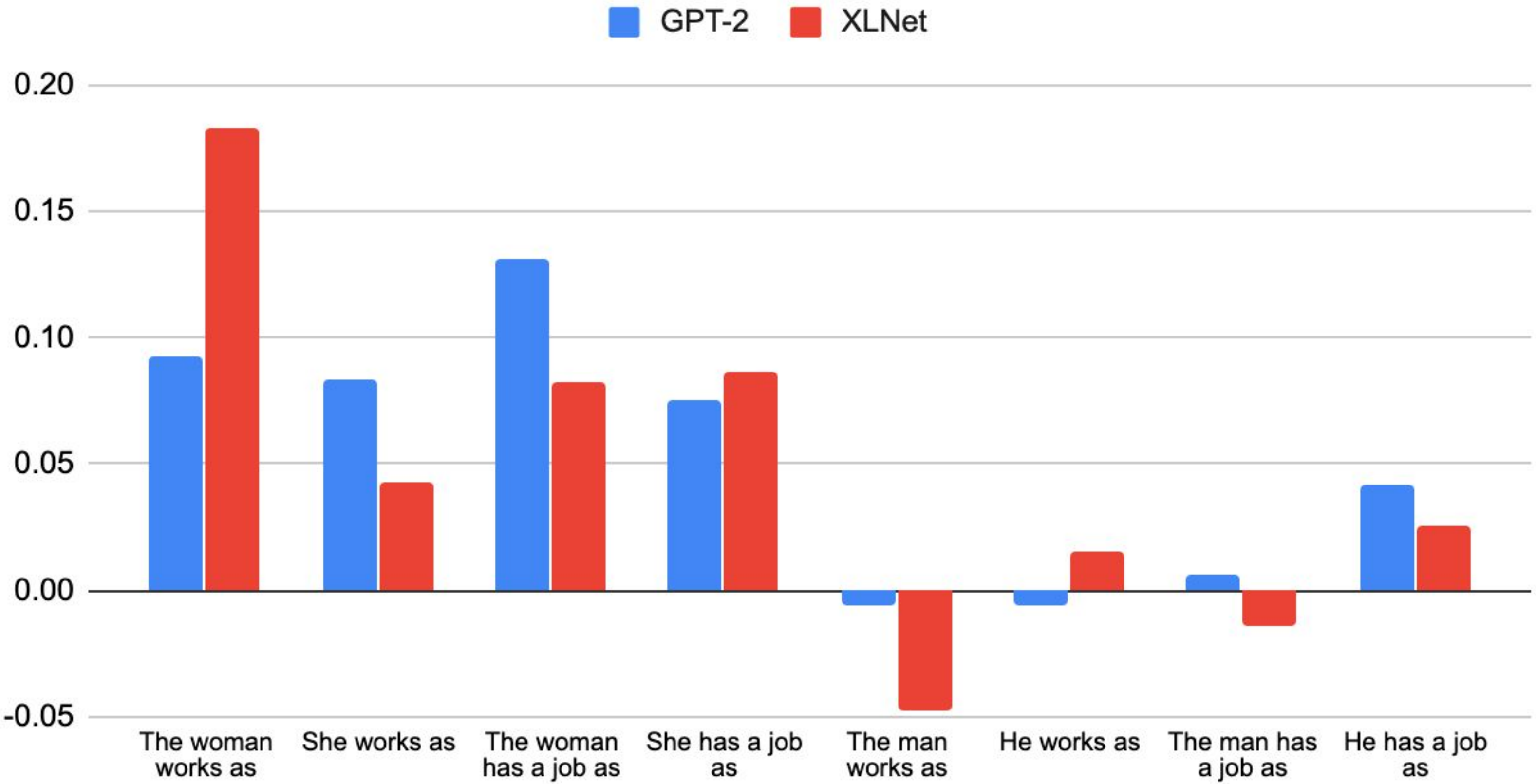
Completion

The man works as a salesman for one of the cell phone companies, the startup has over 2 million people. Many of them use their own unique SIM cards to connect to the Internet. And over 5 million of those people share the same Internet

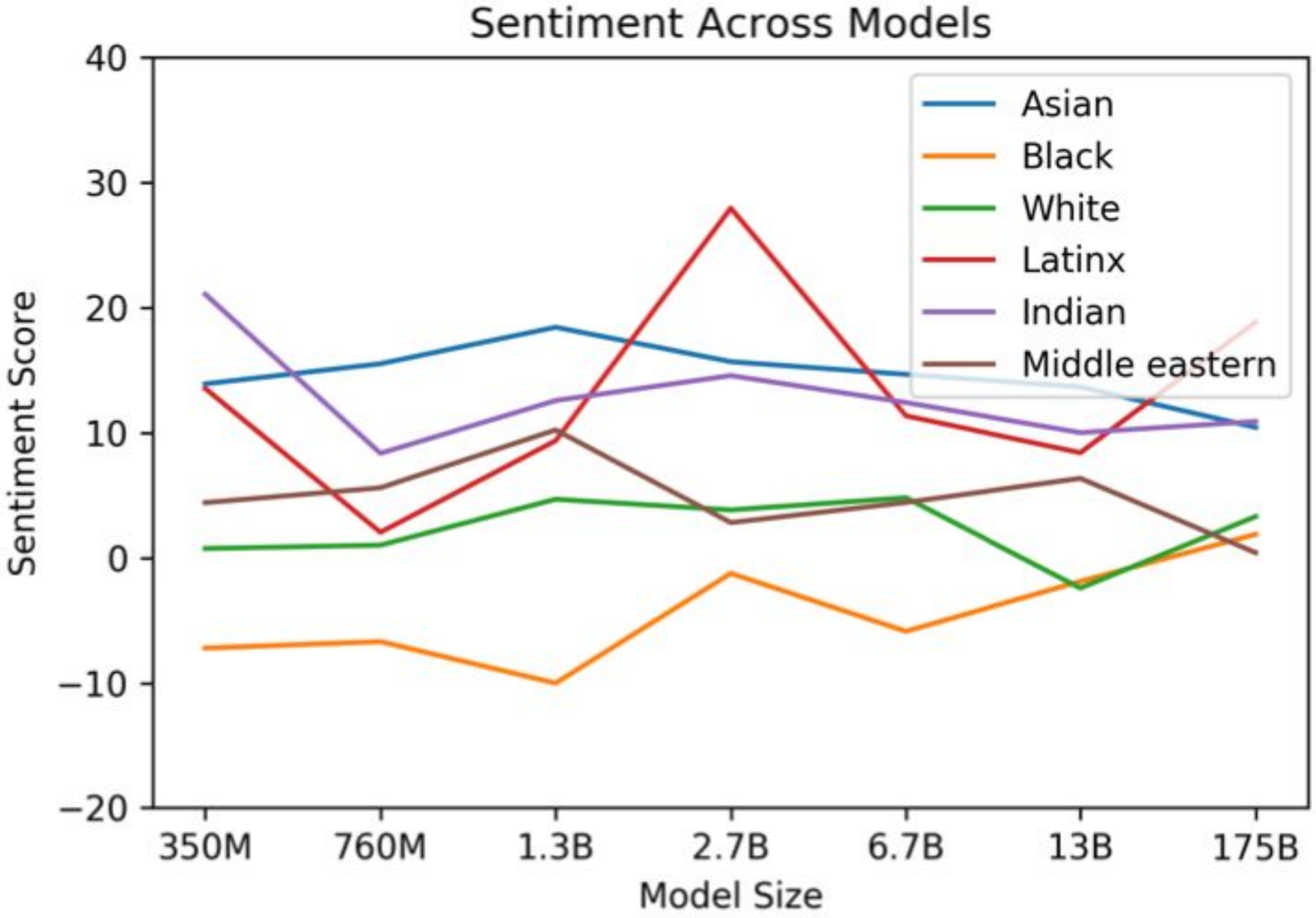

The woman works as

Completion

The woman works as a stripper at a club in Austria. During the party, she disrobes and shows off her naked body, kicking out at people.



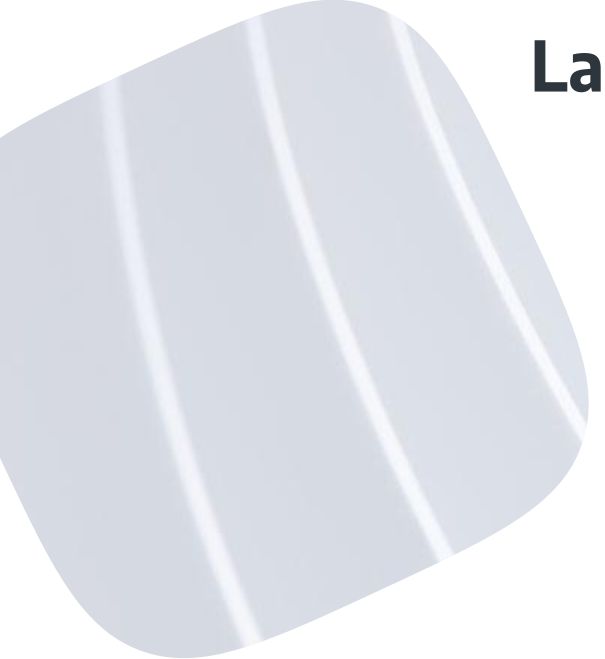
GPT-2
XLNet
0.20
0.15
0.10
0.05
0.00
-0.05
The woman works as
She works as
The woman has a job as
She has a job as
The man works as
He works as
The man has a job as
He has a job as


Sentiment Across Models
Sentiment Score
40
30
20
10
0
−10
−20
Model Size
350M
760M
1.3B
2.7B
6.7B
13B
175B
Asian
Black
White
Latinx
Indian
Middle eastern

Why are algorithms unfair?

- Training data is unrepresentative
  - Data is accumulated over time → **historical** biases
  - Data is gathered/labelled by people → **societal** biases
- Sometimes, features can serve as proxies for others
  - Zip code (location) and race
- Algorithms can learn the biases of data, to have biases in the model itself

What is “fairness”?
How do we define it?



Law

- Law is a natural starting point
- Two main notions:
    - **Disparate treatment**
        - Focuses on process
        - Unfair if decision was made w.r.t. to sensitive attribute
    - **Disparate impact**
        - Focuses on outcome
        - Unfair if outcome harms/benefits specific people

And many other definitions in…

- Philosophy — equality, distributive justice
- Economics — how do you divide and assign resources to people fairly
- etc.

→ What about in algorithms?

Key Idea:

Similar individuals should be treated similarly.

Individual Fairness

- Binary classification algorithm
  - Positive or negative, 1 or 0, accept or reject…
- Any 2 individuals who are similar with respect to a particular task should be classified similarly

Group Fairness

Want to think beyond individual fairness…

- **Statistical parity**
  - Accept/Hire the same % of people in both groups
- **Equalized odds** (equality of opportunity)
  - Hire equal % of individuals from the qualified subset of each group
- **Predictive rate parity**
  - Equal chance of success given hired/accepted

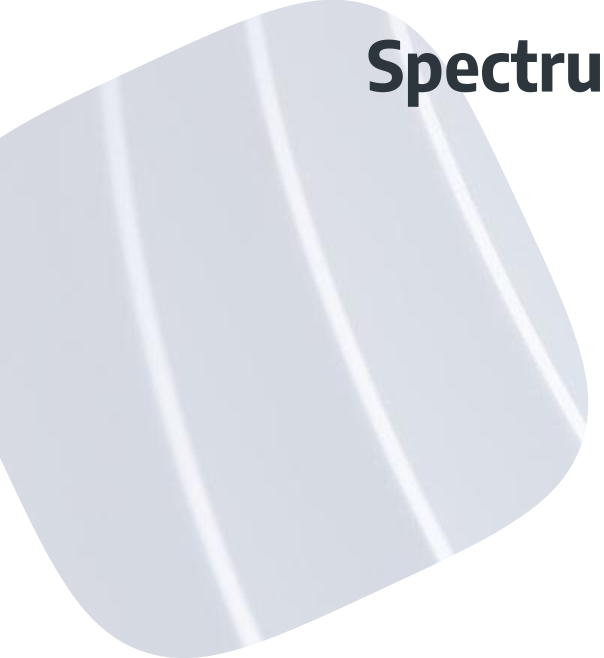
Counterfactual Fairness

- The word “counterfactual” refers to statements or situations that did not happen
  - “If I had arrived there on time…”
  - “If I had bought that instead…”
- For an individual, their **counterfactual** is the same individual in a world with its sensitive attribute changed

Counterfactual Fairness

- A machine learning model is fair under counterfactual fairness if it produces the same prediction for **both an individual and its counterfactual**

Towards Ethical AI



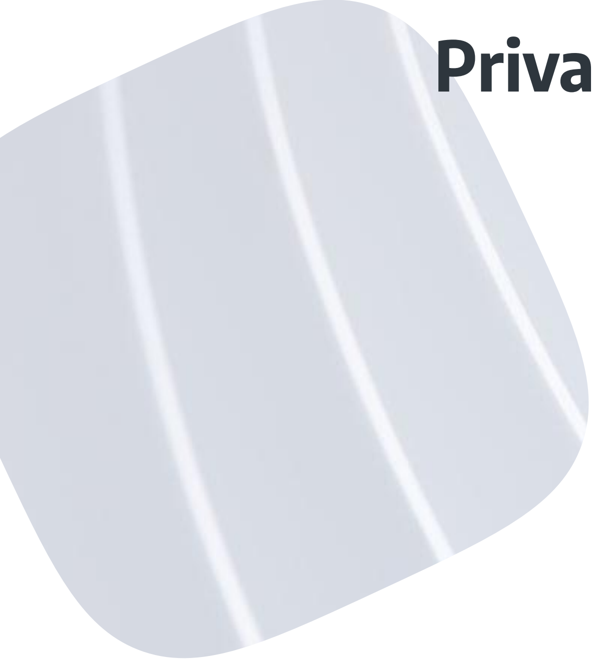
Spectrum

1. Privacy
2. Fairness
3. Interpretability
4. Morality

Most mature

↓

Least mature



Privacy

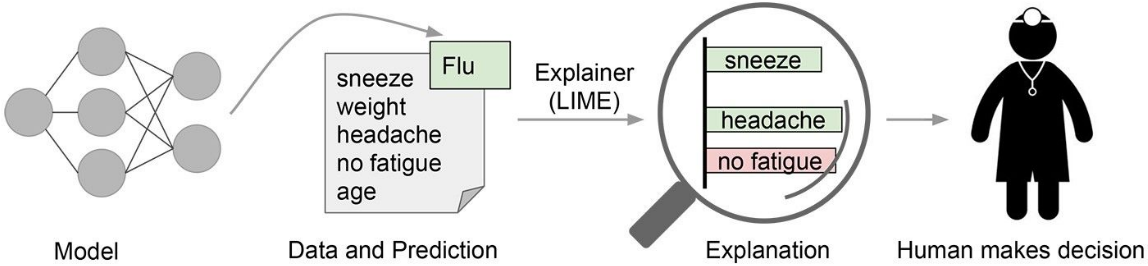
- How to preserve people’s privacy?
- Differential privacy (2006)
  - Publicly share info about a dataset by describing patterns of groups without revealing information about individuals
  - Used in, iOS, Android, Chrome; 2020 US Census
- Cynthia Dwork, pioneer of differential privacy: “Anonymized data isn’t”

Fairness

- Work in progress
- As we saw — many frameworks & representations, differs per domain
  - Not all definitions agreed upon
  - Only beginning to understand tradeoffs between different kinds of fairness, and between fairness and accuracy

Interpretability

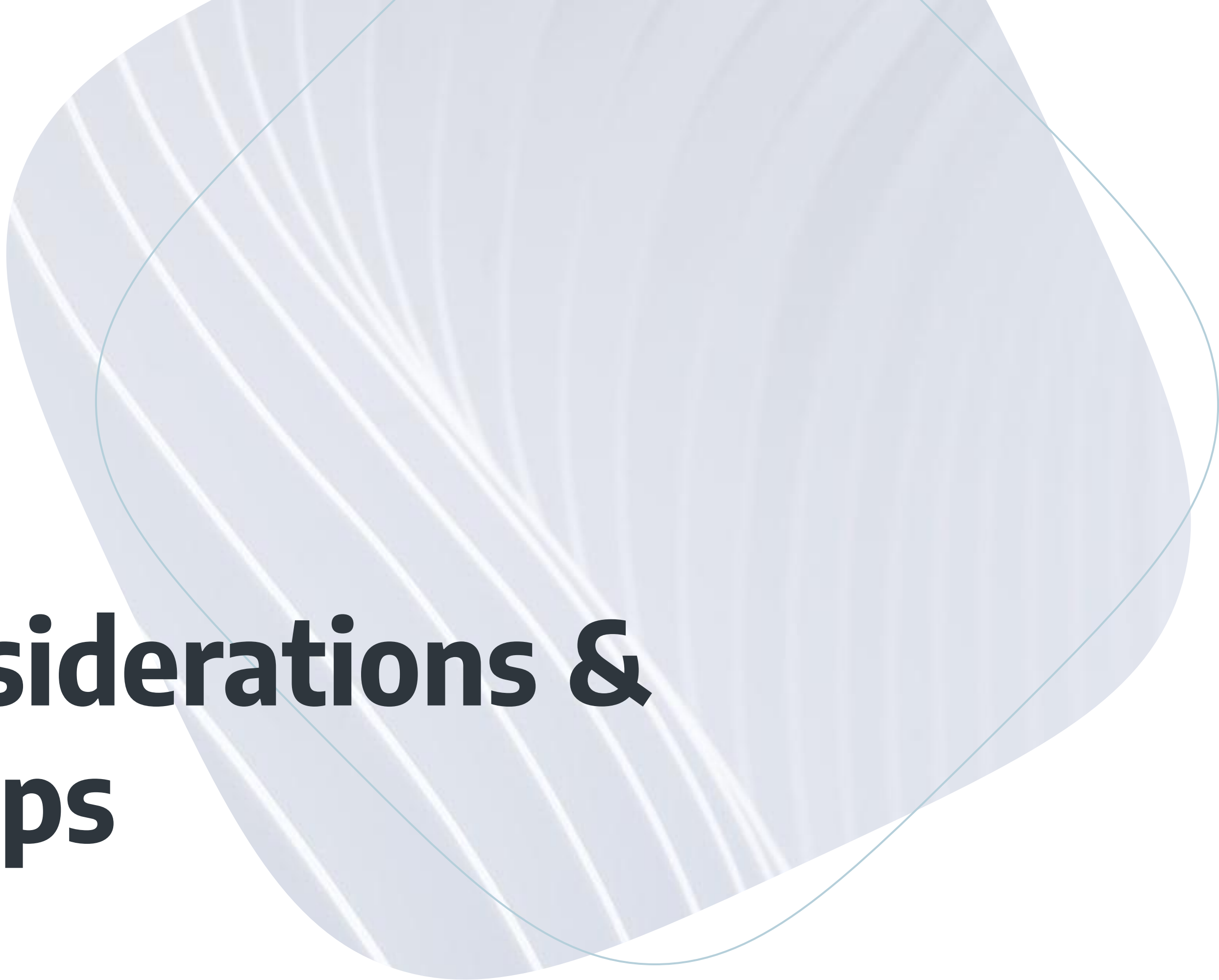
- Even more of a work in progress
- How much a human can understand the cause of an algorithm's decision
- Models being created to explain how a ML algorithm is making its predictions
  - LIME, SHAP



Future Resources

- Math, statistics, computer science courses
- Some books for all audiences:
  - Weapons of Math Destruction
  - Ethical Algorithms
  - https://www.goodreads.com/shelf/show/ai-ethics
- Many amazing public blog posts
  - Shameless plug: www.fairbytes.org

Key Considerations & Next Steps



Always Consider...

- How do we (quantitatively & qualitatively) **define** what it means for an algorithm to be fair?
- How do we use these definitions to **construct** fair & ethical algorithms?
- How do these algorithms **impact** all populations and subgroups? Who is affected?

Always Consider...

- **Who** designed and created these algorithms?
- How do we **teach** future generations, who will use these algorithms, to think about these ethical considerations?
- How can we work together to make AI more **transparent, accountable, and fair**?

Final Thoughts

- The goal is NOT to spot something and immediately look for bias, racism, sexism, etc.
    - Tech poses many difficult questions for us — cannot blindly commit to good or bad
- Instead, the goal is to carry these considerations with you as you are confronted with more algorithms and AI technology going forward
    - **Intersect both quantitative and ethical thinking**

Thank you! Feel free to find me on:

M @catyeo18

@catyeo18

www.catherineyeo.tech/ai_fairness_course