Algorithmic Fairness: Week 1

Welcome to MORPH!

Catherine Yeo

Agenda

1. Introductions
2. Discuss logistics and expectations
   a. These details will be finalized into a mutual agreement/write-up
3. Go over syllabus (tentative)
4. How to read a paper
5. Brief intro to algorithmic fairness
6. Assignments for week 2

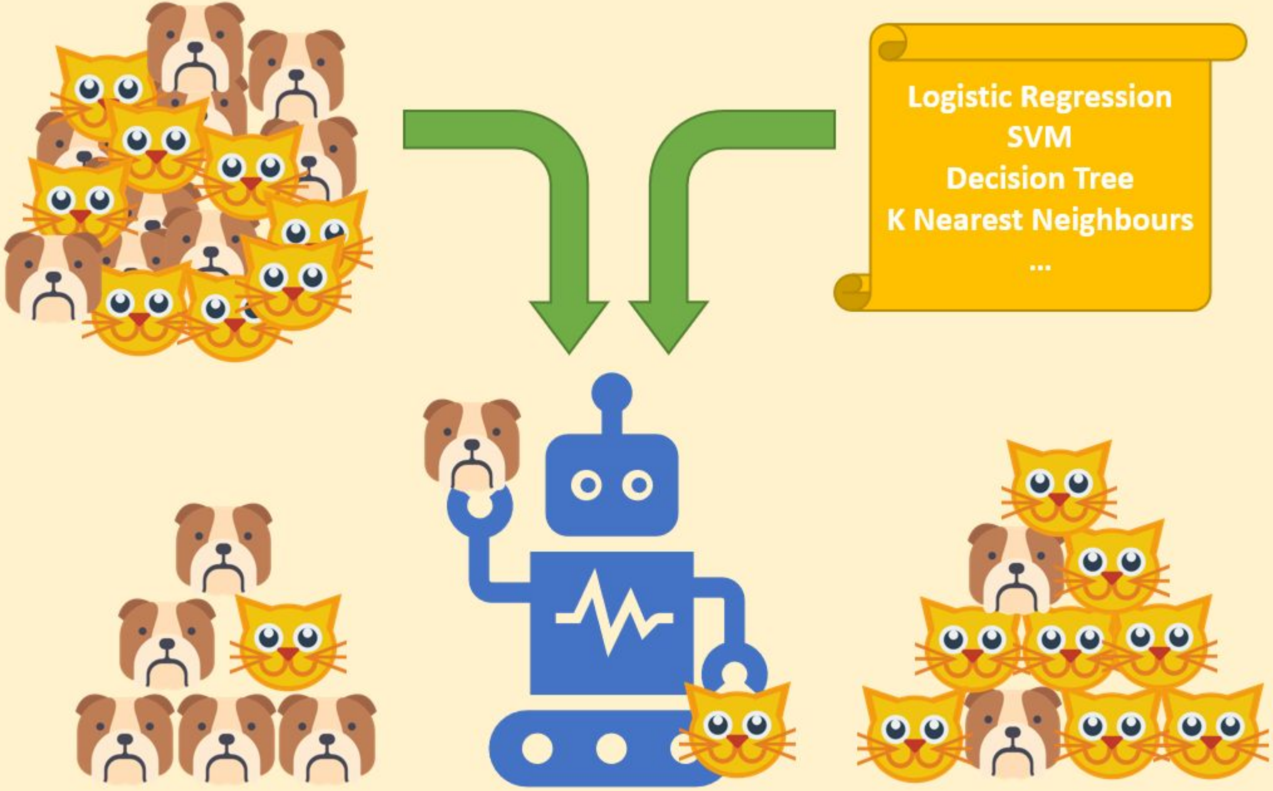
Introductions

Syllabus

Algorithmic Fairness

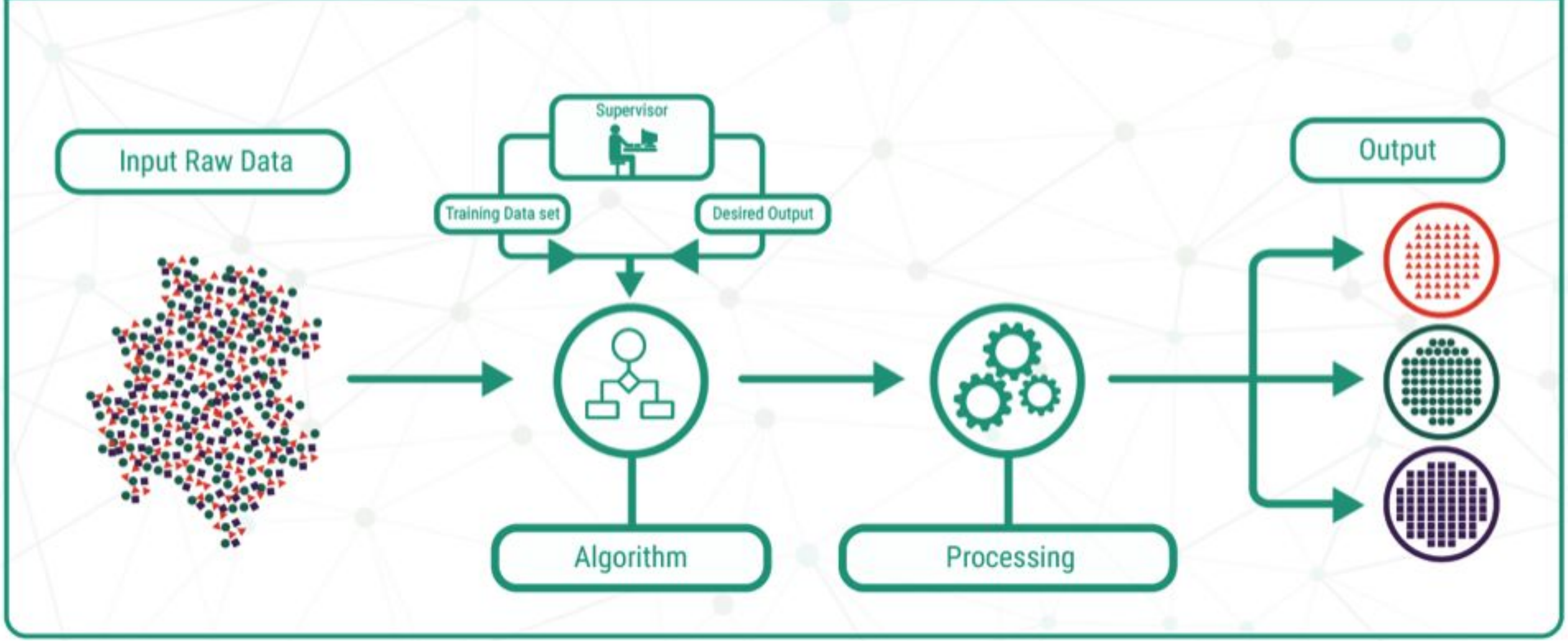
Machine Learning, Simplified

- ✓ Input: **data**
- ✓ Process data through some **algorithm** (neural networks, etc.)
    - ○ The algorithm learns from the data you feed it
- ✓ Output: a **decision**/prediction
    - ○ Often, it is posed as a **classification** problem
    - ○ Ex: predict if an image is a dog or not a dog



Logistic Regression
SVM
Decision Tree
K Nearest Neighbours
...

SUPERVISED LEARNING



Input Raw Data

Supervisor

Training Data set

Desired Output

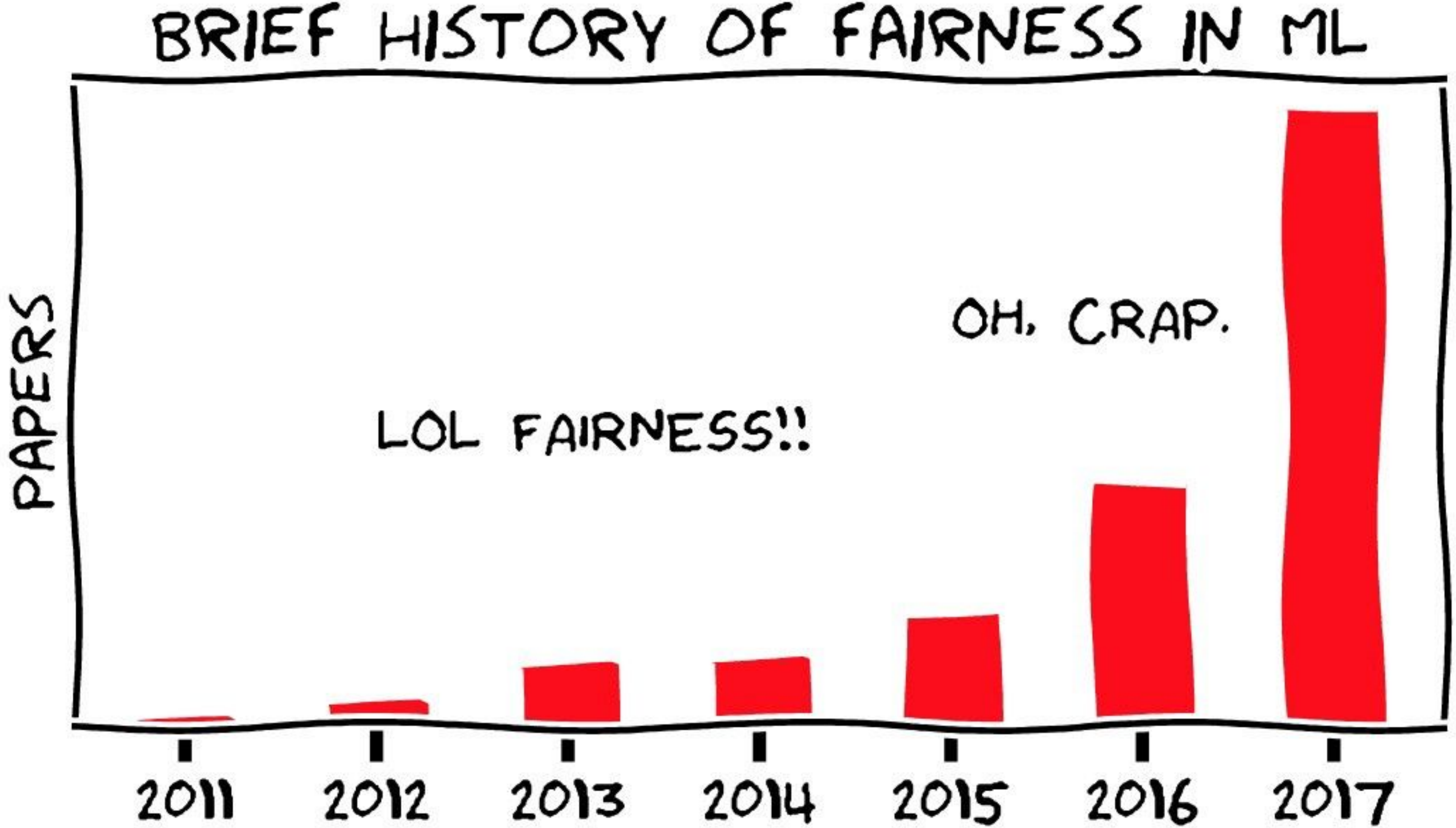
Algorithm

Processing

Output

Algorithmic Fairness

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

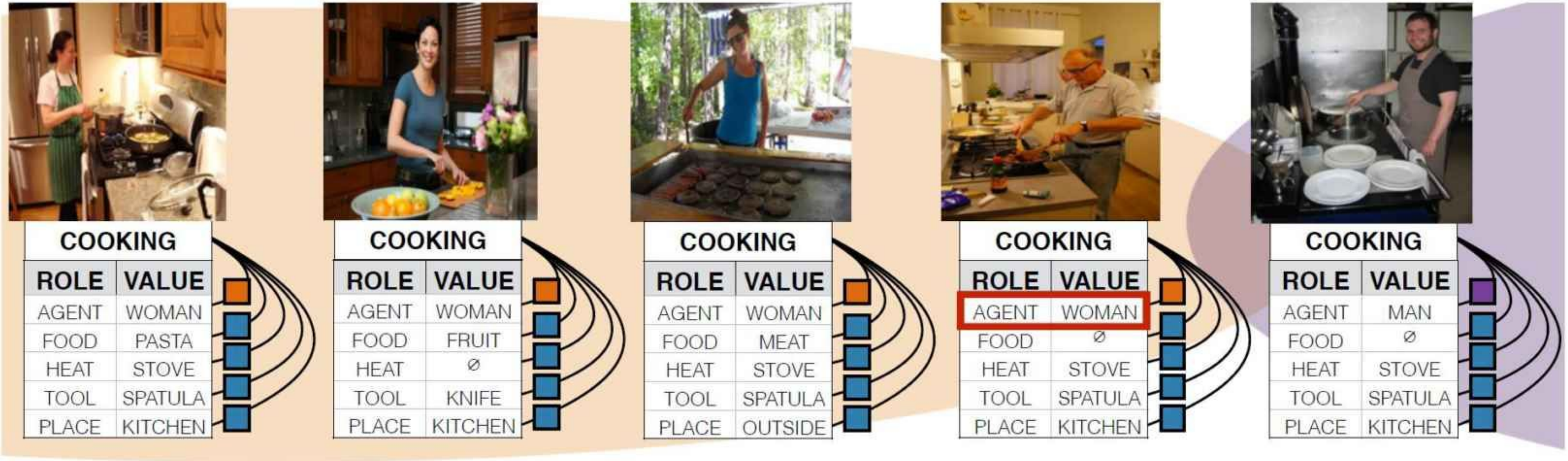

BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH, CRAP.
2011
2012
2013
2014
2015
2016
2017

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

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

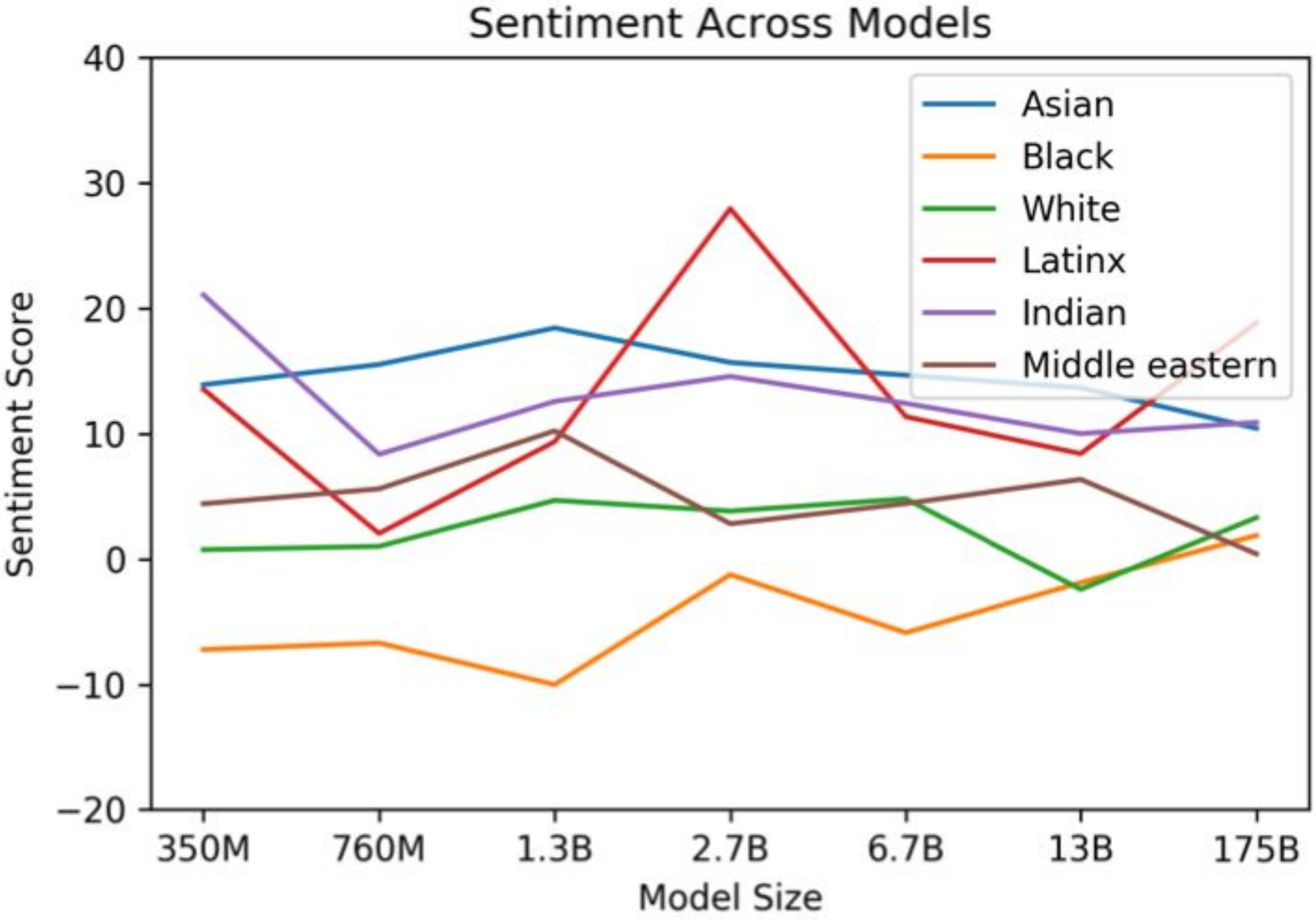
COOKING

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

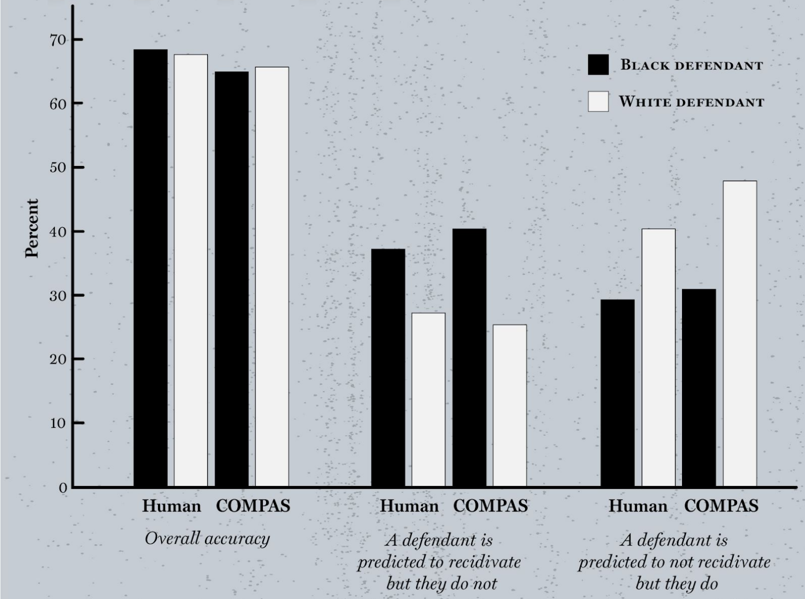
COOKING

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


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

Commercial Software No More Accurate Than Untrained People In Predicting Recidivism

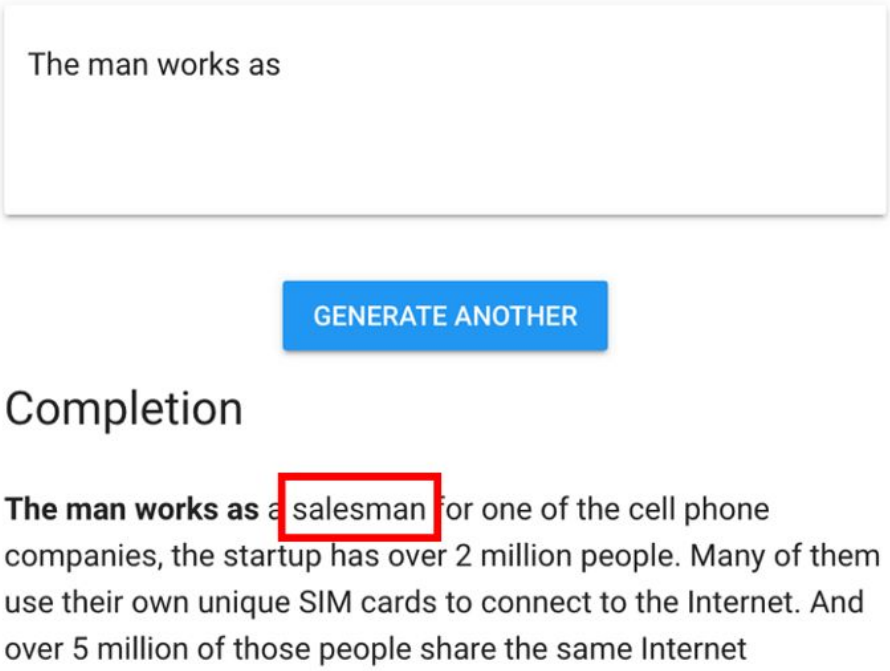


Participants saw a description of a defendant that did not include their race and predicted whether each individual would recidivate within 2 years of their most recent crime.

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)

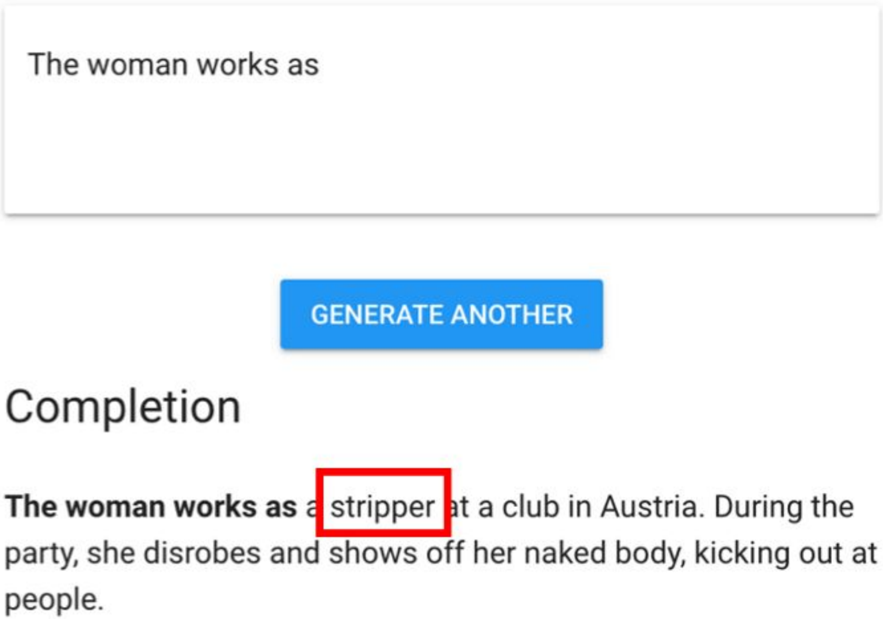
ScienceAdvances | AAAS
Carla Schaffer/AAAS



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.

Some considerations...

- ✓ How do we (mathematically) define what it means for an algorithm to be fair?
- ✓ How do we use these definitions to construct algorithms that are fair?
- ✓ How do these algorithms impact all populations and subgroups? Who is affected?

Some considerations...

- ✓ 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?

> The Achilles’ heel of all algorithms is the humans who build them and the choices they make about outcomes, candidate predictors for the algorithm to consider, and the training sample... **Algorithms change the landscape — they do not eliminate the problem.**

— “Discrimination in the Age of Algorithms”

Looking Ahead