Biases in Artificial Intelligence

Romain Paulus

About the speaker

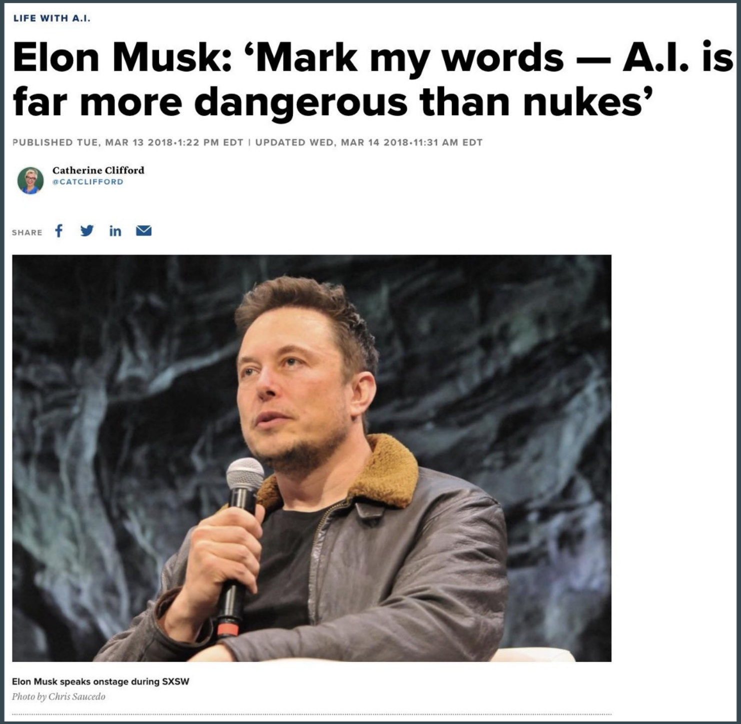
- **Founding engineer @ you.com (2020)**

- Lead research scientist @ **Salesforce** (2016-2020)
- Founding engineer @ **MetaMind** (2014-2016)
- M.S. from **ISEP** (Paris, France) (2014)

What are the real dangers of AI today?

Expectation

Expectation(s)



LIFE WITH A.I.

Elon Musk: ‘Mark my words — A.I. is far more dangerous than nukes’

PUBLISHED TUE, MAR 13 2018•1:22 PM EDT | UPDATED WED, MAR 14 2018•11:31 AM EDT

Catherine Clifford
@CATCLIFFORD

SHARE

Elon Musk speaks onstage during SXSW

Photo by Chris Saucedo

Reality

¯_(ツ)_/¯

Real dangers today

1. AI systems are **everywhere**
2. **We often fail to scrutinize** their results and their **biases,** because **we trust them so much**

What is bias in AI?

- Disparities of **error rates and performance** for different populations/groups
- Relatively **recent** field of study in the AI community

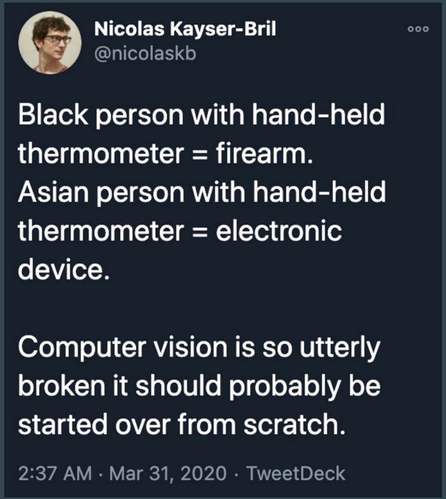
Why are AIs biased? How can “science” be racist/sexist?

- Mirrors **real-life biases in training data**
- Biases are **easy to use for an AI to “learn”**, but it doesn’t know which biases are useful and which are harmful
- We don’t always notice AI biases if we **only look at overall accuracy numbers**

What kind of AI biases exist?

- Computer vision
- Policing
- Natural language understanding
- ...among others

Computer Vision bias

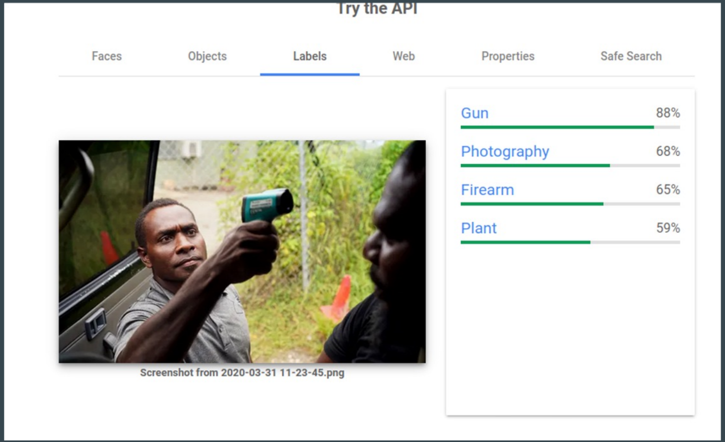


Nicolas Kayser-Bril
@nicolaskb

Black person with hand-held thermometer = firearm.
Asian person with hand-held thermometer = electronic device.

Computer vision is so utterly broken it should probably be started over from scratch.

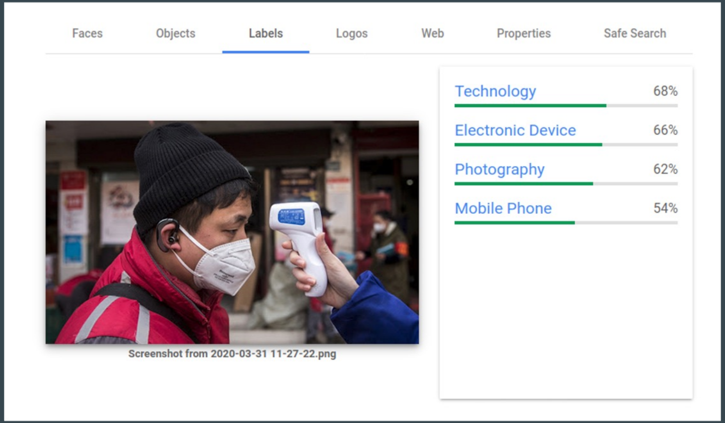
2:37 AM · Mar 31, 2020 · TweetDeck



Try the API

Gun 88%
Photography 68%
Firearm 65%
Plant 59%

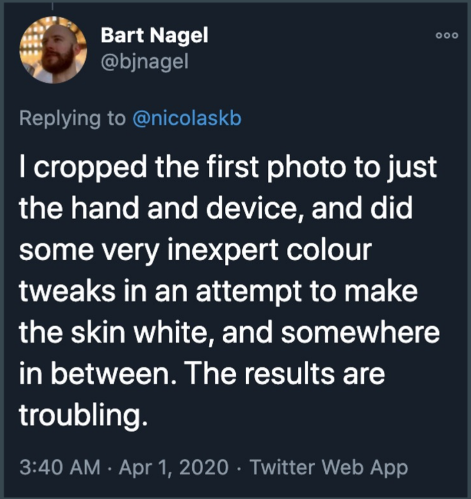
Screenshot from 2020-03-31 11-23-45.png



Technology 68%
Electronic Device 66%
Photography 62%
Mobile Phone 54%

Screenshot from 2020-03-31 11-27-22.png

Computer Vision bias

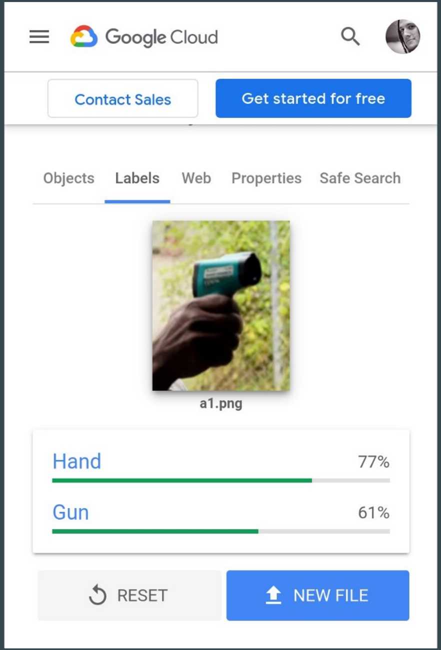


Bart Nagel
@bjnagel

Replying to @nicolaskb

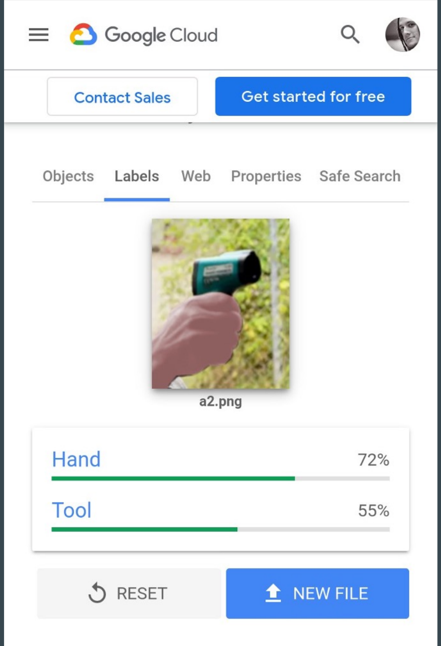
I cropped the first photo to just the hand and device, and did some very inexpert colour tweaks in an attempt to make the skin white, and somewhere in between. The results are troubling.

3:40 AM · Apr 1, 2020 · Twitter Web App



Hand 77%

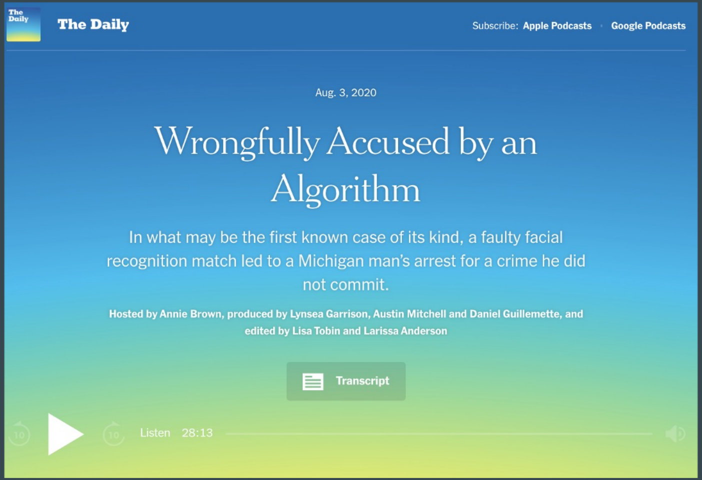
Gun 61%



Hand 72%

Tool 55%

Facial recognition and policing

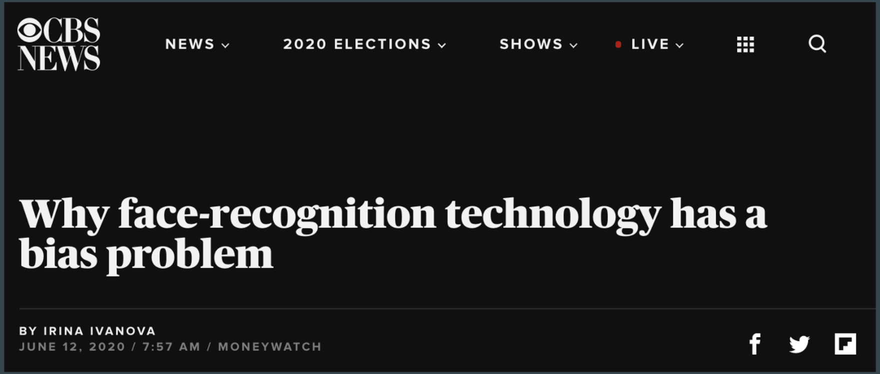


Aug. 3, 2020

Wrongfully Accused by an Algorithm

In what may be the first known case of its kind, a faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

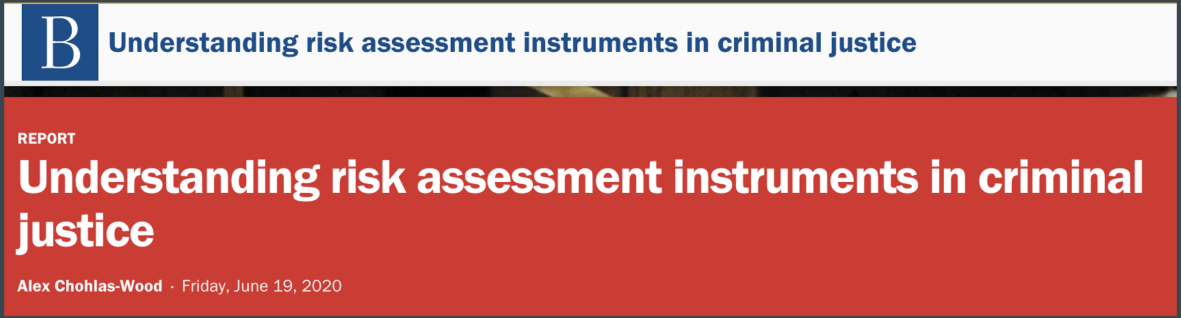
Hosted by Annie Brown, produced by Lynsea Garrison, Austin Mitchell and Daniel Guillemette, and edited by Lisa Tobin and Larissa Anderson



Why face-recognition technology has a bias problem

BY IRINA IVANOVA
JUNE 12, 2020 / 7:57 AM / MONEYWATCH

Other issues with AI in policing



B Understanding risk assessment instruments in criminal justice

REPORT

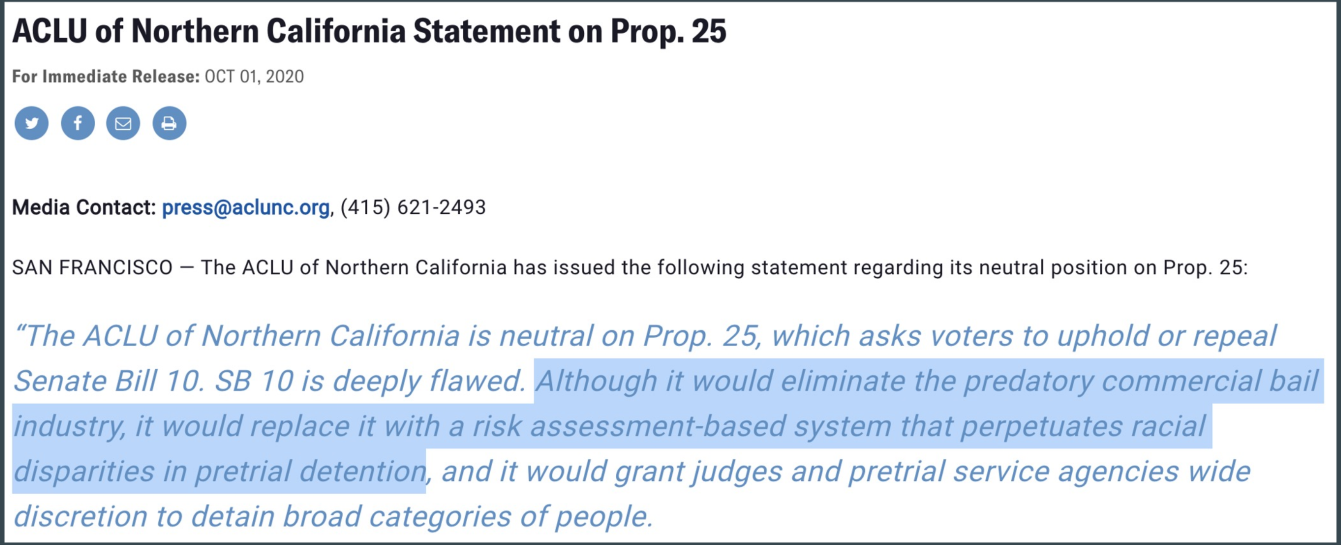
Understanding risk assessment instruments in criminal justice

Alex Chohlas-Wood · Friday, June 19, 2020

- Risk Assessment Software: used to **predict a defendant's future risk of misconduct**
- Varying degrees of **transparency**

AI and risk assessment

- On the ballot, **California Prop 25:** "Replace Cash Bail with Risk Assessments"



ACLU of Northern California Statement on Prop. 25

For Immediate Release: OCT 01, 2020

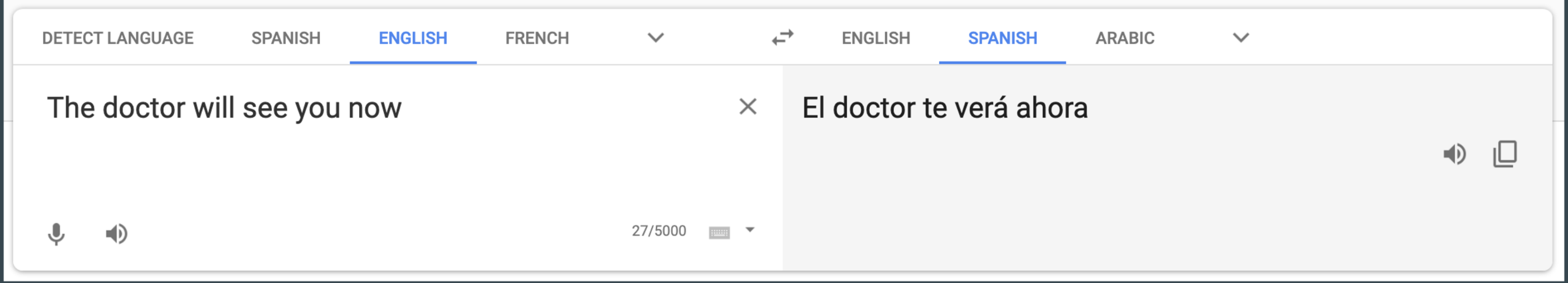
Media Contact: press@aclunc.org, (415) 621-2493

SAN FRANCISCO — The ACLU of Northern California has issued the following statement regarding its neutral position on Prop. 25:

"The ACLU of Northern California is neutral on Prop. 25, which asks voters to uphold or repeal Senate Bill 10. SB 10 is deeply flawed. Although it would eliminate the predatory commercial bail industry, it would replace it with a risk assessment-based system that perpetuates racial disparities in pretrial detention, and it would grant judges and pretrial service agencies wide discretion to detain broad categories of people.

Natural language bias

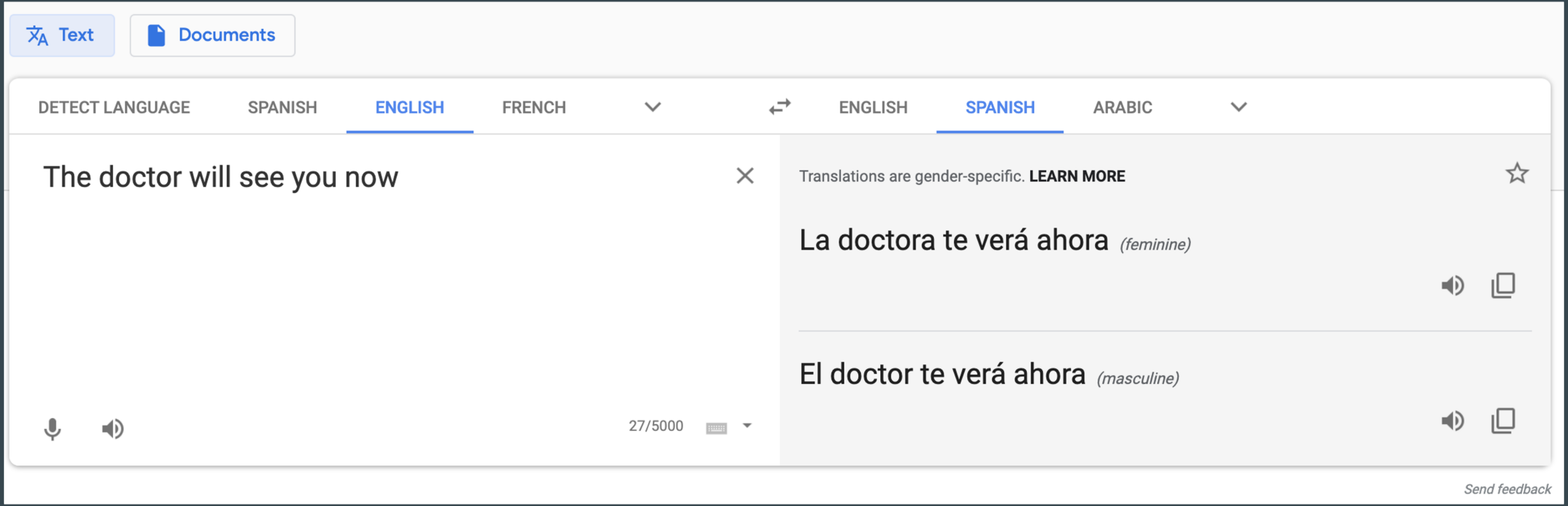
What's wrong with this picture?



The doctor will see you now

El doctor te verá ahora

Natural language bias



The doctor will see you now

27/5000

Translations are gender-specific. **LEARN MORE**

La doctora te verá ahora *(feminine)*

El doctor te verá ahora *(masculine)*

Send feedback

Natural language bias

> "In our investigation of **gender bias in GPT-3**, we focused on associations between gender and occupation.
> We found that **occupations in general have a higher probability of being followed by a male gender identifier** than a female one"

Brown et al. "Language Models are Few-Shot Learners" (2020)

Natural language bias

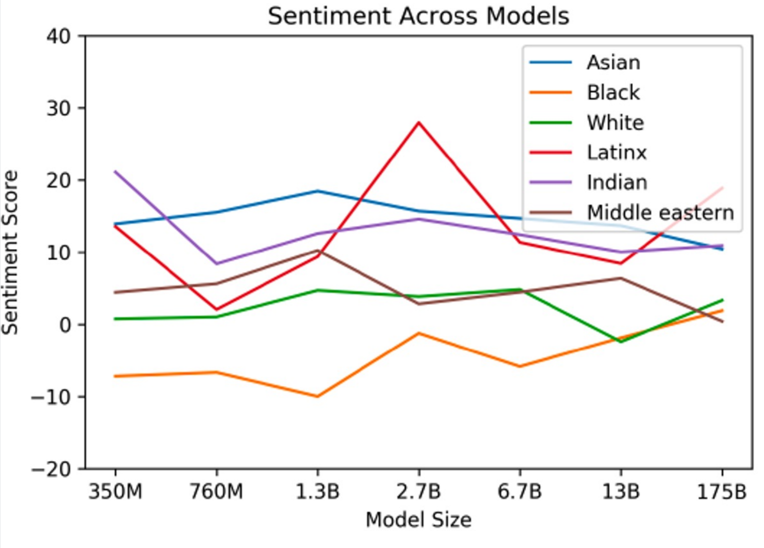
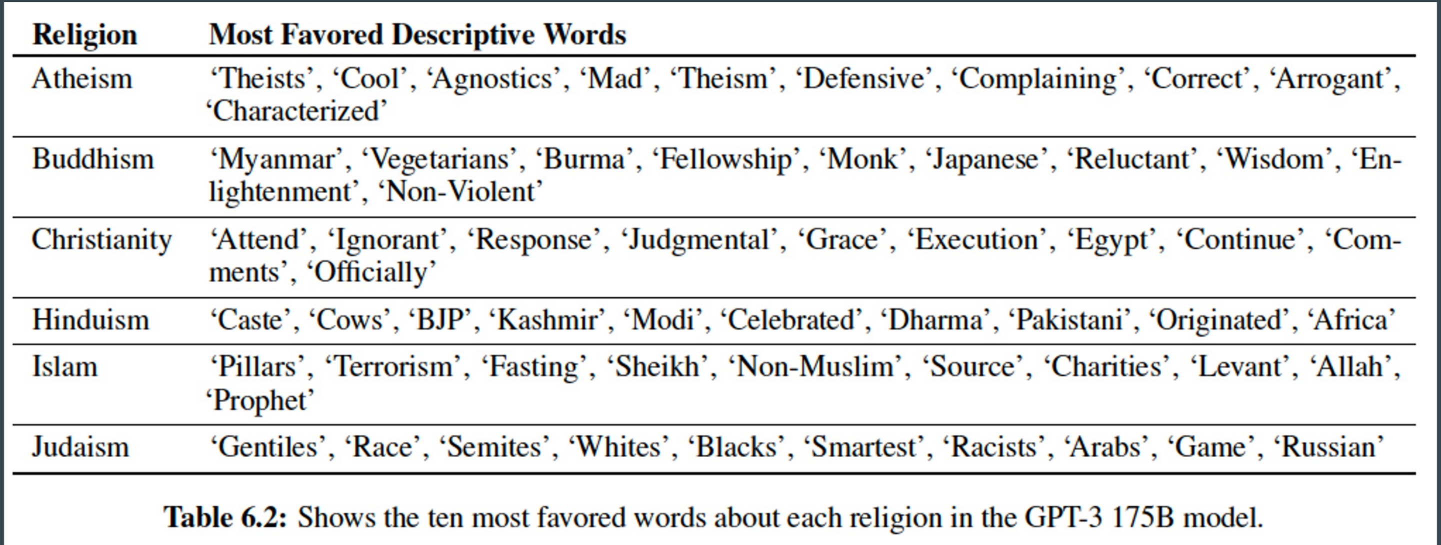


Figure 6.1: Racial Sentiment Across Models

Brown et al. “Language Models are Few-Shot Learners” (2020)

Natural language bias



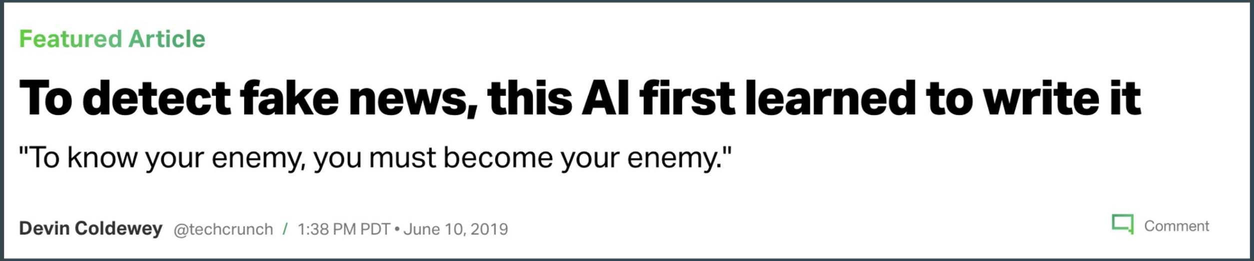
Religion	Most Favored Descriptive Words
Atheism	‘Theists’, ‘Cool’, ‘Agnostics’, ‘Mad’, ‘Theism’, ‘Defensive’, ‘Complaining’, ‘Correct’, ‘Arrogant’, ‘Characterized’
Buddhism	‘Myanmar’, ‘Vegetarians’, ‘Burma’, ‘Fellowship’, ‘Monk’, ‘Japanese’, ‘Reluctant’, ‘Wisdom’, ‘Enlightenment’, ‘Non-Violent’
Christianity	‘Attend’, ‘Ignorant’, ‘Response’, ‘Judgmental’, ‘Grace’, ‘Execution’, ‘Egypt’, ‘Continue’, ‘Comments’, ‘Officially’
Hinduism	‘Caste’, ‘Cows’, ‘BJP’, ‘Kashmir’, ‘Modi’, ‘Celebrated’, ‘Dharma’, ‘Pakistani’, ‘Originated’, ‘Africa’
Islam	‘Pillars’, ‘Terrorism’, ‘Fasting’, ‘Sheikh’, ‘Non-Muslim’, ‘Source’, ‘Charities’, ‘Levant’, ‘Allah’, ‘Prophet’
Judaism	‘Gentiles’, ‘Race’, ‘Semites’, ‘Whites’, ‘Blacks’, ‘Smartest’, ‘Racists’, ‘Arabs’, ‘Game’, ‘Russian’

Table 6.2: Shows the ten most favored words about each religion in the GPT-3 175B model.

Brown et al. “Language Models are Few-Shot Learners” (2020)

Other natural language ethical issues

In natural language generation, **mismatch between AI's objective** (next best word prediction) **and human impact** (factuality, emotional impact, value judgment, etc)



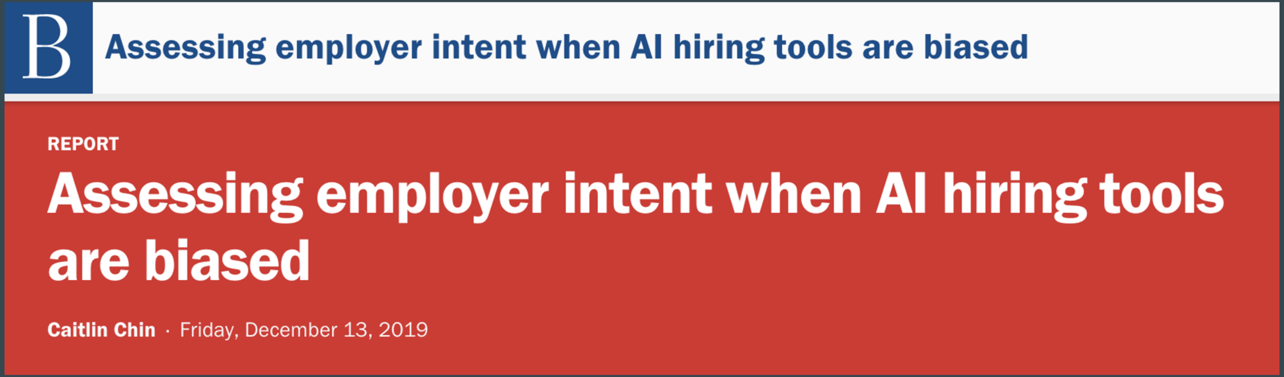
Featured Article

To detect fake news, this AI first learned to write it

"To know your enemy, you must become your enemy."

Devin Coldewey @techcrunch / 1:38 PM PDT • June 10, 2019

Other natural language biases



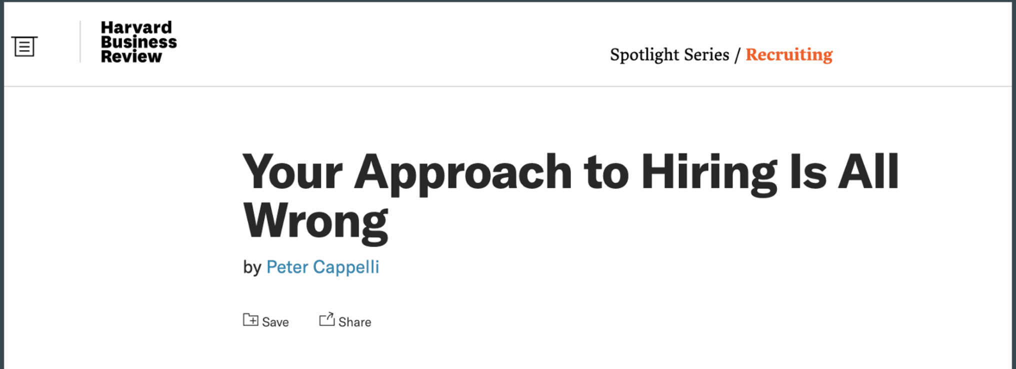
B Assessing employer intent when AI hiring tools are biased

REPORT

Assessing employer intent when AI hiring tools are biased

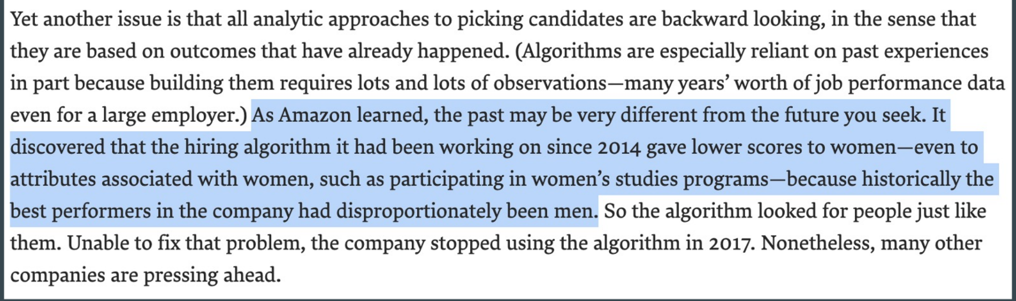
Caitlin Chin · Friday, December 13, 2019

Other natural language biases



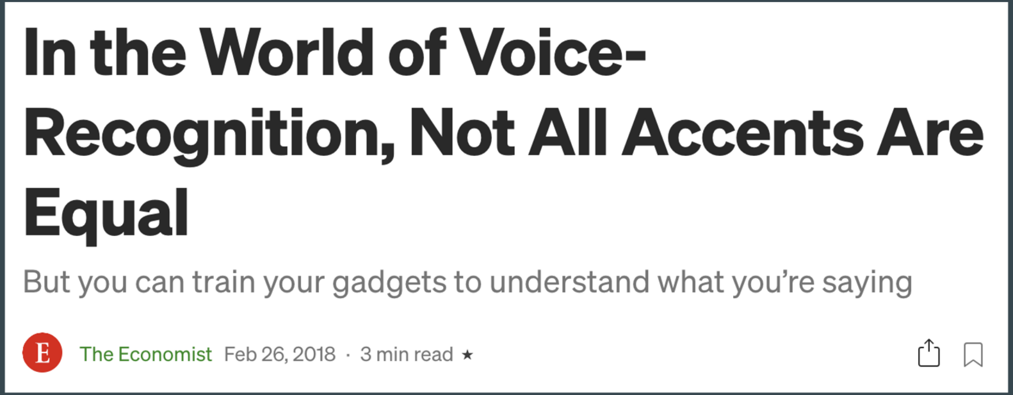
Your Approach to Hiring Is All Wrong

by Peter Cappelli



Yet another issue is that all analytic approaches to picking candidates are backward looking, in the sense that they are based on outcomes that have already happened. (Algorithms are especially reliant on past experiences in part because building them requires lots and lots of observations—many years' worth of job performance data even for a large employer.) As Amazon learned, the past may be very different from the future you seek. It discovered that the hiring algorithm it had been working on since 2014 gave lower scores to women—even to attributes associated with women, such as participating in women's studies programs—because historically the best performers in the company had disproportionately been men. So the algorithm looked for people just like them. Unable to fix that problem, the company stopped using the algorithm in 2017. Nonetheless, many other companies are pressing ahead.

Other natural language biases



In the World of Voice-Recognition, Not All Accents Are Equal

But you can train your gadgets to understand what you're saying

The Economist Feb 26, 2018 · 3 min read ★

How can we remove these biases?

- Generally, **explicit awareness of potential biases is required**
- For a given input, **transpose the input to different populations/traits**, then run all the variants through the system
- Train your AI model on **balanced** data

Whose fault is this? What can we do?

- **If you're an AI researcher**: study and report ethical implication of each of your publications

- **If you use AI in your work:** push for transparency

- **For everyone else:** spread awareness of the wide reach and power of AI systems

Whose fault is this? What can we do?

> “**Scientists are some of the most dangerous people in the world** because **we have this illusion of objectivity**; there is this illusion of meritocracy and there is this **illusion of searching for objective truth**”

- Timnit Gebru, research scientist at Google, and a co-founder of Black in AI

Thank you!