Responsible Machine Learning

INFO-4604, Applied Machine Learning

University of Colorado Boulder

November 14, 2017

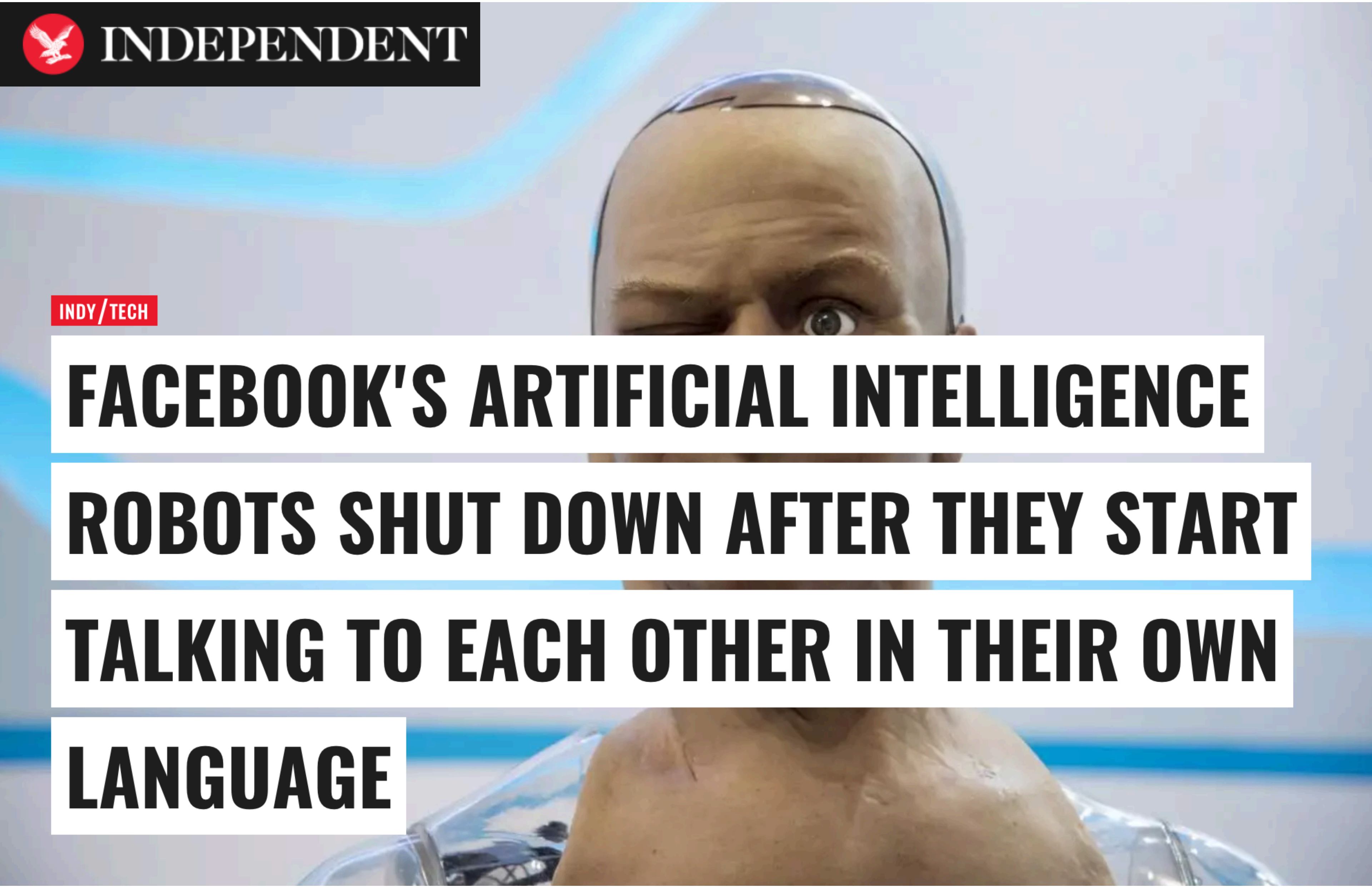
Prof. Michael Paul

Is Machine Learning Dangerous?

Elon Musk Warns Governors: Artificial Intelligence Poses 'Existential Risk'

July 17, 2017 · 10:39 AM ET

CAMILA DOMONOSKE


INDEPENDENT
INDY/TECH
FACEBOOK'S ARTIFICIAL INTELLIGENCE ROBOTS SHUT DOWN AFTER THEY START TALKING TO EACH OTHER IN THEIR OWN LANGUAGE

Sean J. Taylor
@seanjtaylor

Follow

Man: Hmm, Windows froze, I guess I need to reboot my PC.
The Independent: MAN FORCED TO SHUT DOWN AI AFTER IT STOPS RESPONDING TO COMMANDS

5:53 PM - 1 Aug 2017

theguardian

Facebook translates 'good morning' into 'attack them', leading to arrest

The man, a construction worker in the West Bank settlement of Beitar Illit, near Jerusalem, posted a picture of himself leaning against a bulldozer with the caption “يصبحهم”, or “yusbihuhum”, which translates as “good morning”.

But Facebook’s artificial intelligence-powered translation service, which it built after parting ways with Microsoft’s Bing translation in 2016, instead translated the word into “hurt them” in English or “attack them” in Hebrew.

Police officers arrested the man later that day, according to Israeli newspaper Haaretz, after they were notified of the post. They questioned him for several hours, suspicious he was planning to use the pictured bulldozer in a vehicle attack, before realising their mistake. At no point before his arrest did any Arabic-speaking officer read the actual post.

Venkat Viswanathan
@venkvis

Follow

.@TeslaMotors Model S autopilot camera misreads 101 sign as 105 speed limit at 87/101 junction San Jose. Reproduced every day this week.

8:40 PM - 14 Jul 2017

Is Machine Learning Dangerous?

- "Doomsday" scenarios not likely any time soon
  - Algorithms are not "intelligent" enough

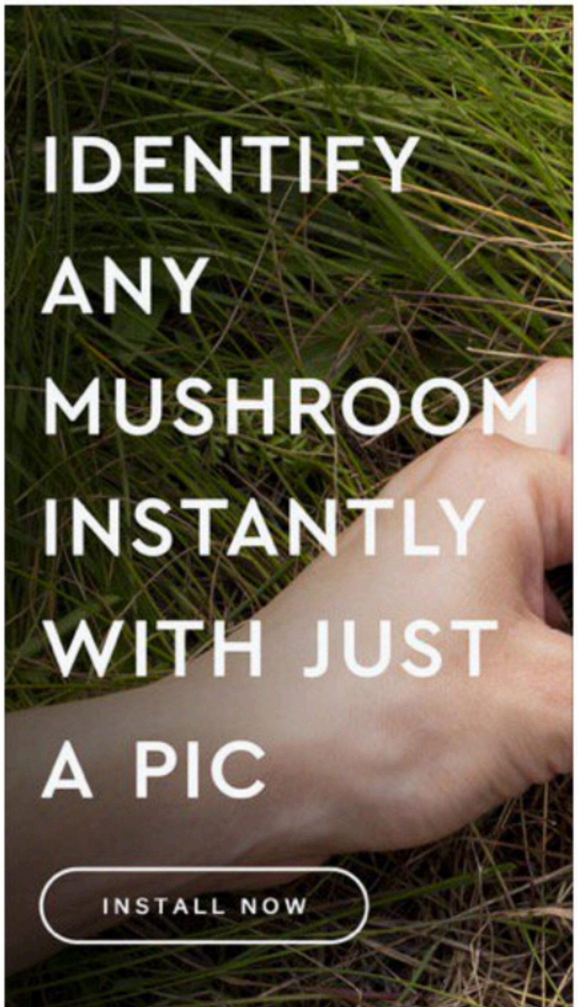
- But machine learning can potentially be misused, misleading, and/or invasive
  - Important to consider implications of what you build

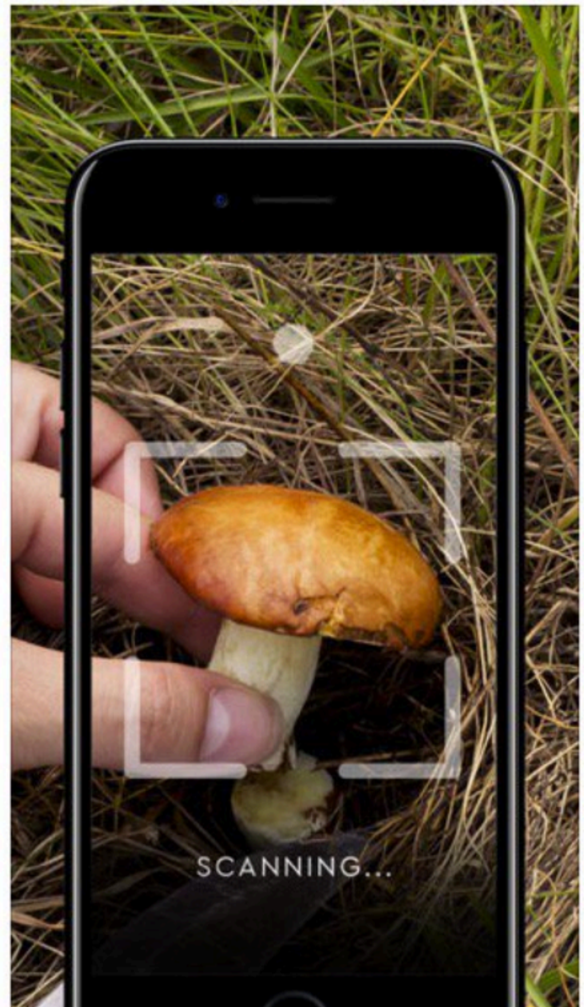
Mushroom - Instant mushroom plants identifi...

Quest Mobile llc

$7.99

In-App Purchases





Principles for Accountable Algorithms

Statement from **Fairness, Accountability, and Transparency in Machine Learning** organization

https://www.fatml.org/resources/principles-for-accountable-algorithms

Algorithms and the data that drive them are designed and created by people -- There is always a human ultimately responsible for decisions made or informed by an algorithm. "The algorithm did it" is not an acceptable excuse if algorithmic systems make mistakes or have undesired consequences, including from machine-learning processes.

Principles for Accountable Algorithms

Responsibility

- Make available externally visible avenues of redress for adverse individual or societal effects of an algorithmic decision system, and designate an internal role for the person who is responsible for the timely remedy of such issues.

Explainability

- Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms.

Accuracy

- Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures.

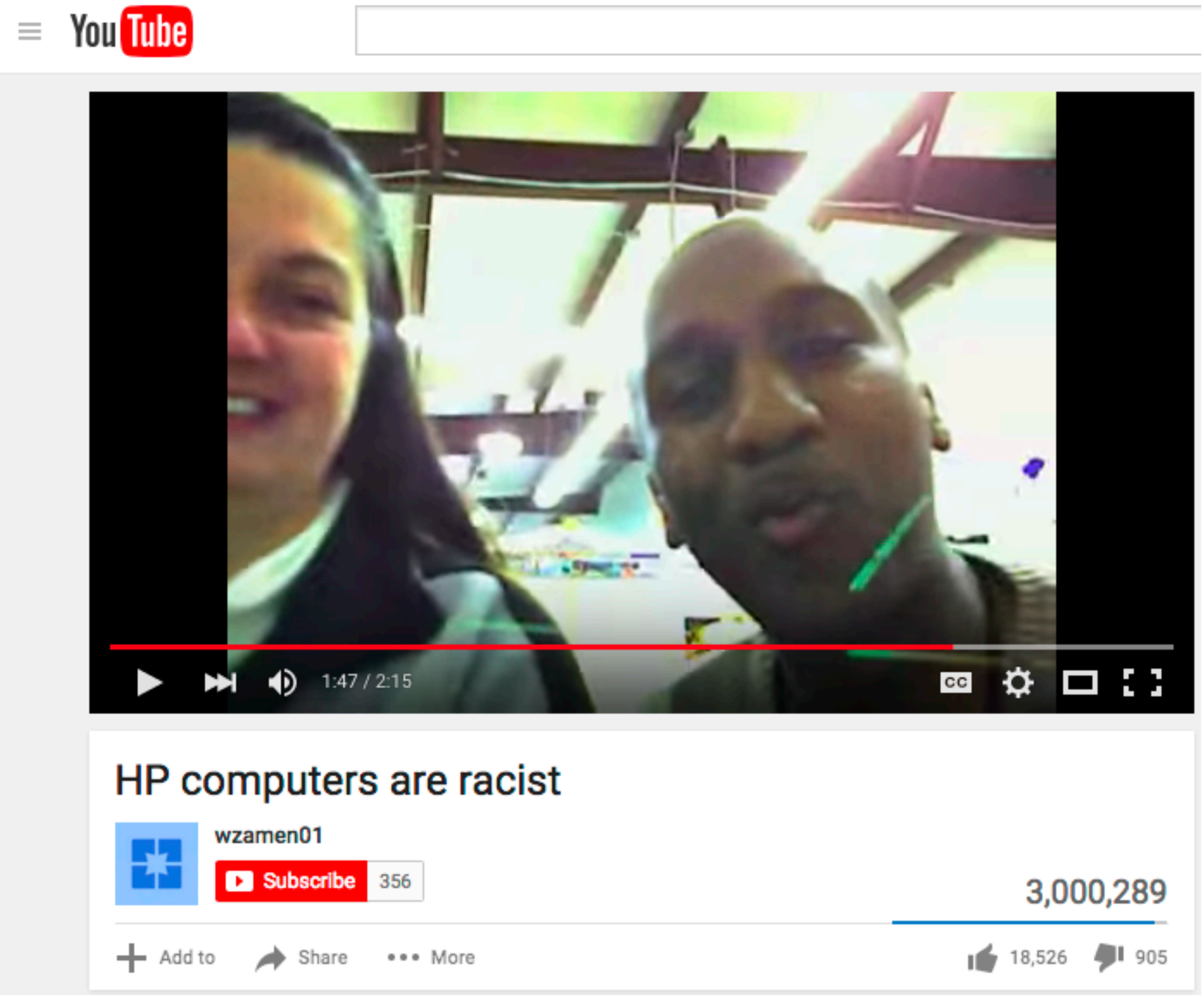
Auditability

- Enable interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use.

Fairness

- Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics (e.g. race, sex, etc).

Fairness


You Tube
1:47 / 2:15
HP computers are racist
wzamen01
Subscribe
356
3,000,289
Add to
Share
More
18,526
905

Fairness

How does this type of error happen?

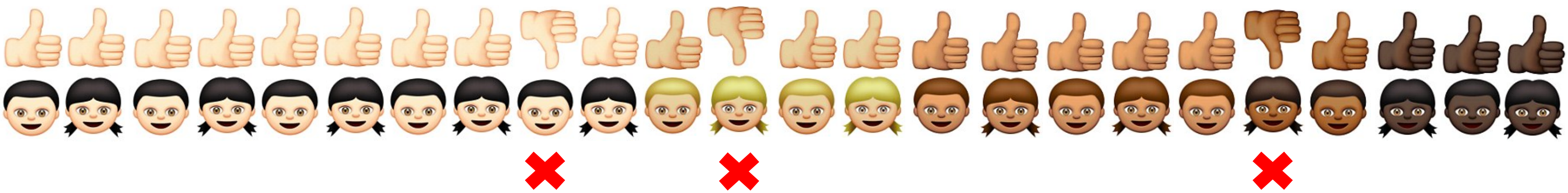
Possibilities:

- Not enough diversity in training data
- Not enough diversity in test data
- Not enough error analysis

Fairness

Suppose your classifier gets 90% accuracy…

Scenario 1:

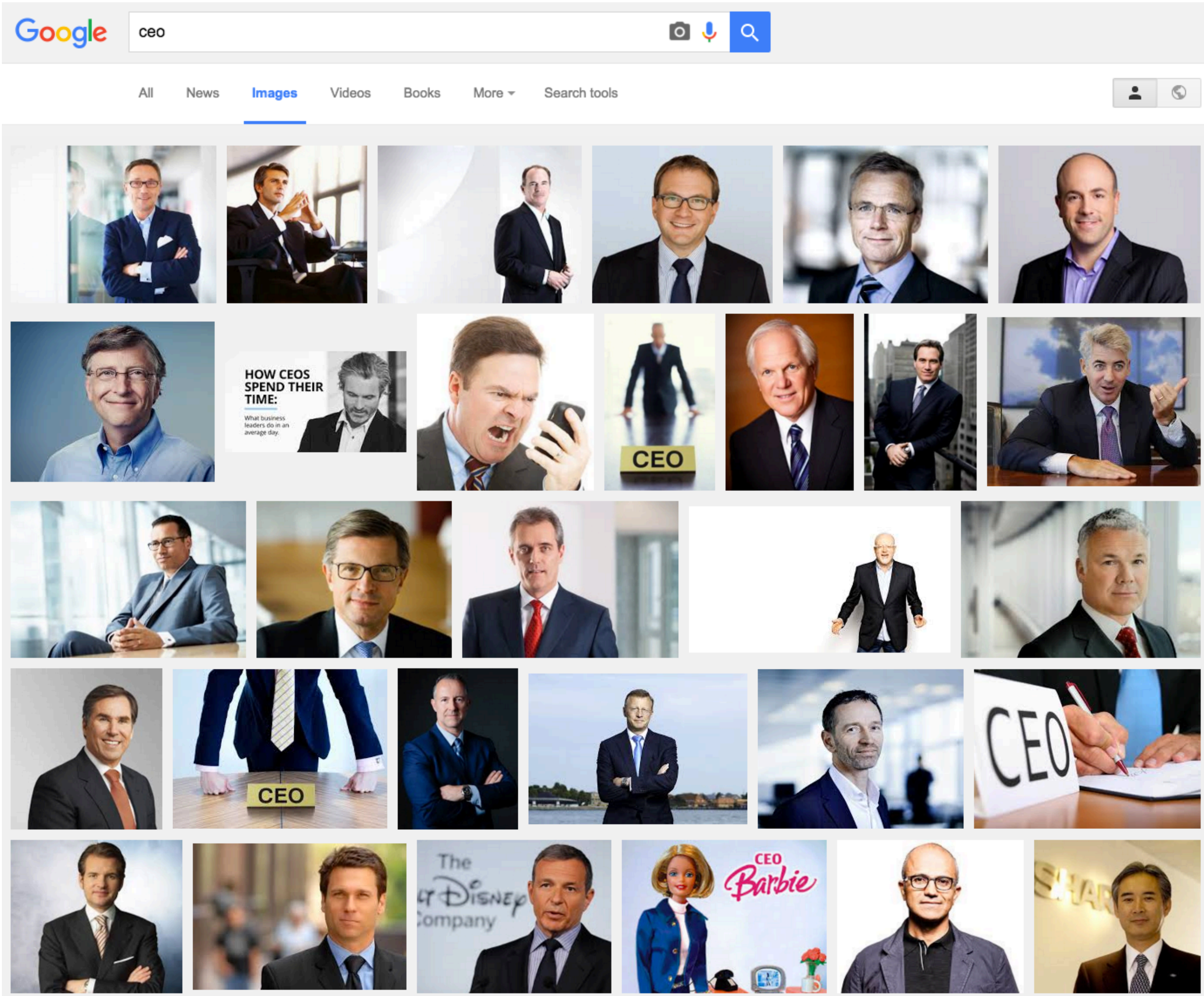


Scenario 2:

Bias

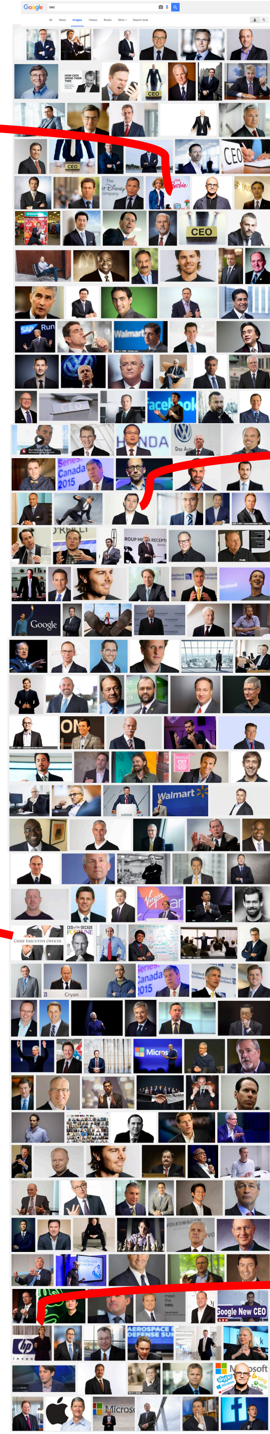
Biases and stereotypes that exist in data will be learned by ML algorithms

Sometimes, those biases will be *amplified* by ML


Google
ceo
All
News
Images
Videos
Books
More
Search tools
HOW CEOS SPEND THEIR TIME:
What business leaders do in an average day.
CEO
CEO
CEO
CEO
Barbie

Zooming out…

Barbie

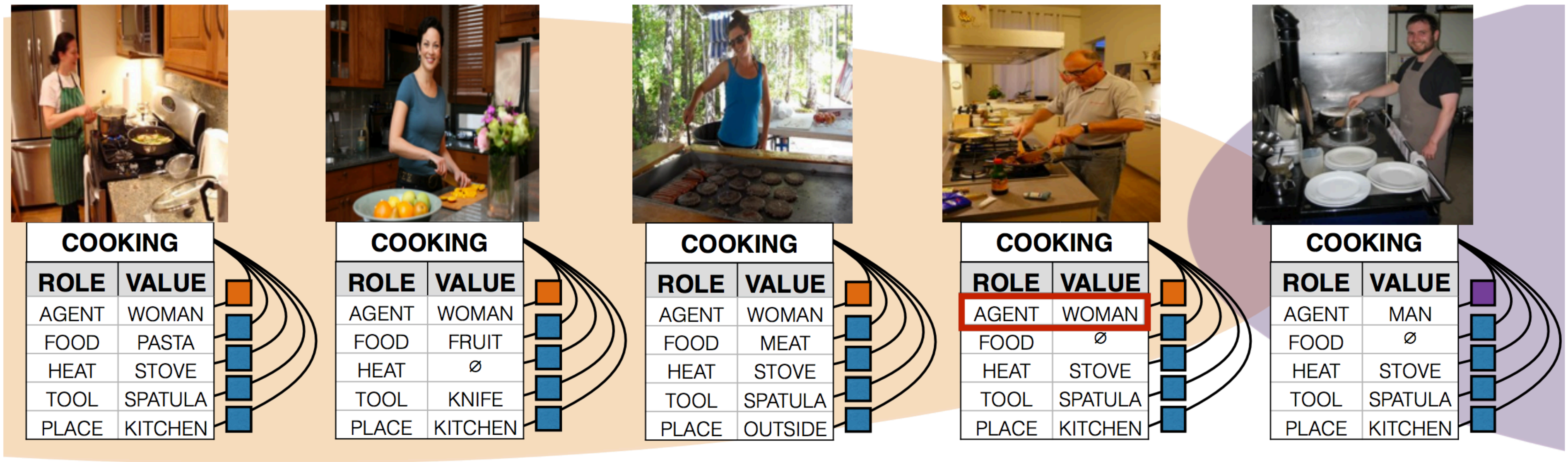


Martin Shkreli, since arrested by the FBI

A woman's hand

Carly Fiorina, former HP CEO, 2016 presidential candidate

- First woman after **206 images!**



Zhao et al (2017):

- Training data:
  Women appeared in ‘cooking’ images 33% more often than men
- Predictions:
  Women appeared 68% more often

Privacy

Training data is often scraped from the web

Personal data may get scooped up by ML systems

- Are users aware of this?
  How do they feel about it?

MegaFace dataset: 4.7 million photos of 627,000 individuals, from Flickr users

Use and Misuse

Machine learning can predict:

- if you are overweight
- if you are transgender
- if you have died

People may build these classifiers for legitimate purposes, but could easily be misused by others

Case Study

Wu and Zhang (2016), “Automated Inference on Criminality using Face Images”

Can we predict if someone is prone to *committing a crime* based on their facial structure?

This study claims yes, with 90% accuracy

Good summary of why the answer is probably no:
http://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

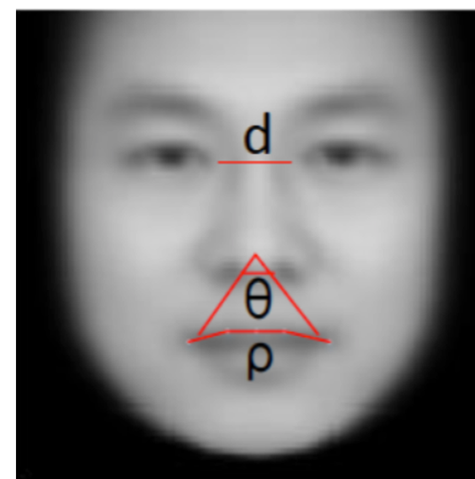
Case Study

How was the dataset created?

- Criminal photos: government IDs
- Non-criminal photos: professional headshots

What did the classifier learn?

- "The algorithm finds that criminals have shorter distances between the inner corners of the eyes, smaller angles between the nose and the corners of the mouth, and higher curvature to the upper lip."



Case Study

If your tool seems dystopian:

- Consider whether this is really something you should be building…
  - One argument: someone will eventually build this technology, so better for researchers to do it first to understand it
  - Still, proceed carefully: understand potential misuse
- Be sure that your claims are correct
  - Solid error analysis is critical
  - Misuse of an inaccurate system even worse than misuse of an accurate system