Imperial College London

MSc Machine Learning & Data Science

Applicable Mathematics

Ethics of ML/DS Part I

Week 3: Fairness

Dr. Chris Anagnostopoulos

Table of contents

- Real-world examples
- Error parity and confusion matrices
- Other fairness metrics
- Pareto fronts

Table of contents

- **Real-world examples**
- Error parity and confusion matrices
- Other fairness metrics
- Pareto fronts

Real-world examples of harm

We will start this week’s content by covering two high profile examples of bias in real-world AI systems that have attracted widespread attention. Namely:

- Bias in word embeddings
- Bias in face recognition software

Man is to king as woman is to ….?

Imagine you wanted to build a simple Turing test for a natural language processing AI. You might ask it to write a paragraph of text, like OpenAI's GPT-x series of algos are doing. A simpler way is to ask it to play the analogy game:

- Man is to king as woman is to …
- France is to Paris as Italy is to …

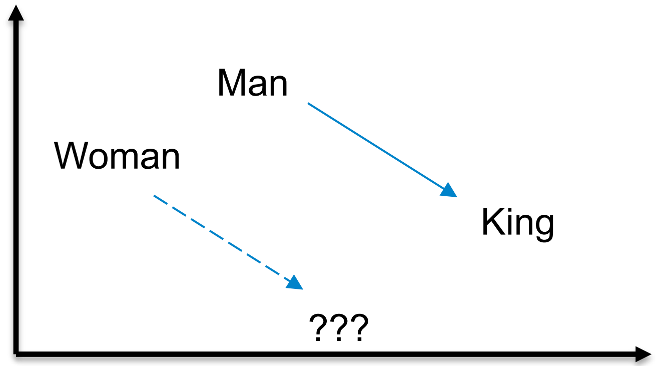
The Turing Test was proposed by Alan Turing in 1950, as a way to ascertain the extent to which an AI system exhibits intelligent behavior. It involves a human having an open-ended conversation with a machine and then a human and being unable to tell which one was which.

Man is to king as woman is to ….?

Imagine you wanted to build a simple Turing test for a natural language processing AI. You might ask it to write a paragraph of text, like OpenAI's GPT-x series of algos are doing. A simpler way is to ask it to play the analogy game:

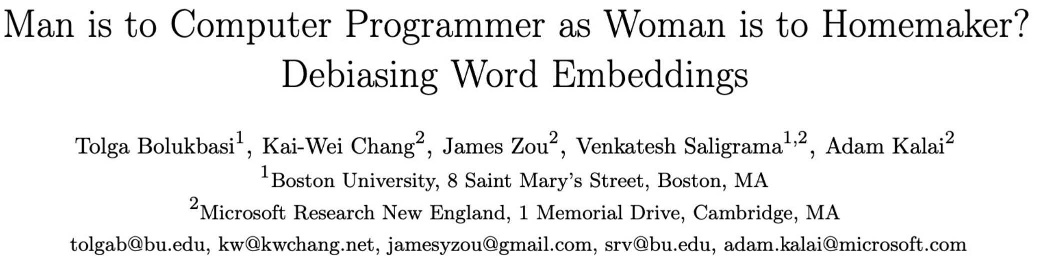
- Man is to king as woman is to queen
- France is to Paris as Italy is to Rome

F(Woman) + f(King) – f(Man) = ?



Man is to king as woman is to ….?

- Man is to doctor as woman is to nurse
- Man is to computer programmer as woman is to homemaker

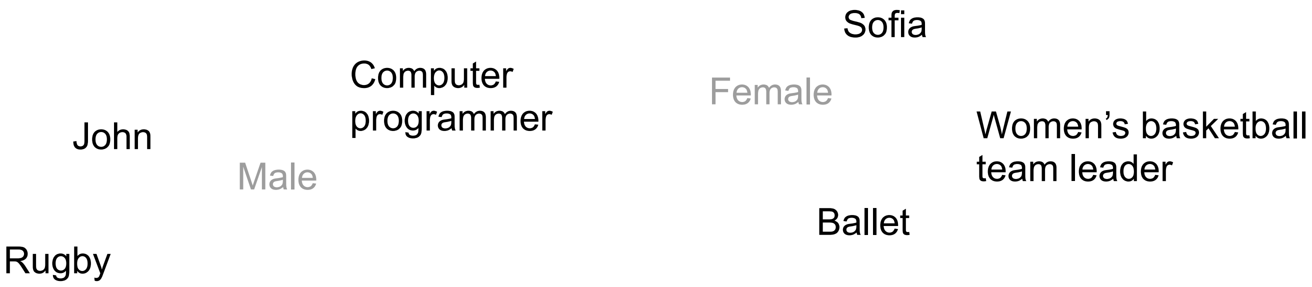


Man is to Computer Programmer as Woman is to Homemaker?
Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Bolukbasi, Tolga, et al. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." Advances in neural information processing systems 29 (2016): 4349-4357.

Predicting candidate relevance via vector embeddings



Sofia

Computer programmer

Female

John

Women’s basketball team leader

Male

Ballet

Rugby

How easy is it to debias embeddings?

- Bias is the result of historical inequalities being reflected in the training document corpus
- Correcting it could rely on either:
  - Modifying the training corpus (e.g., by using or placing more weight on recent data)
  - In a supervised setting, you could re-annotate parts of the data
  - You could try to post-hoc (i.e., after training) modify the algorithm to reduce its bias

Notice how explainability proves key in our ability to detect bias and reason about how to fix it. Embeddings are compatible with visual intuition.

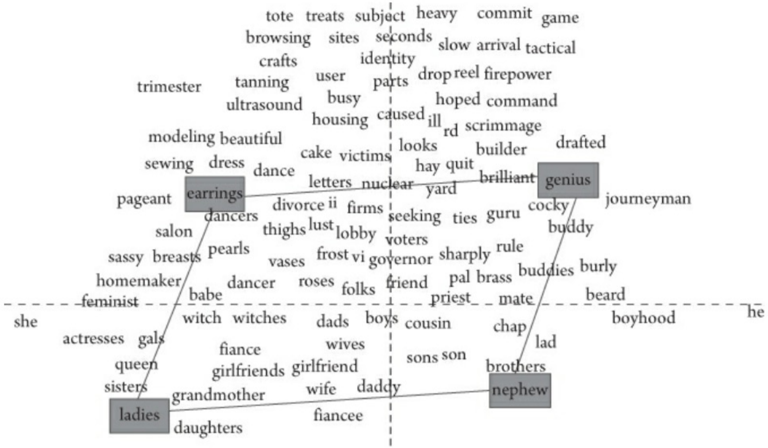
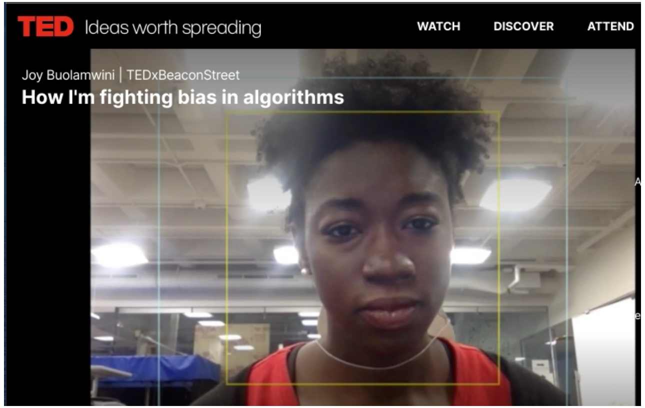


Fig. 8. A word embedding exhibiting gender bias.

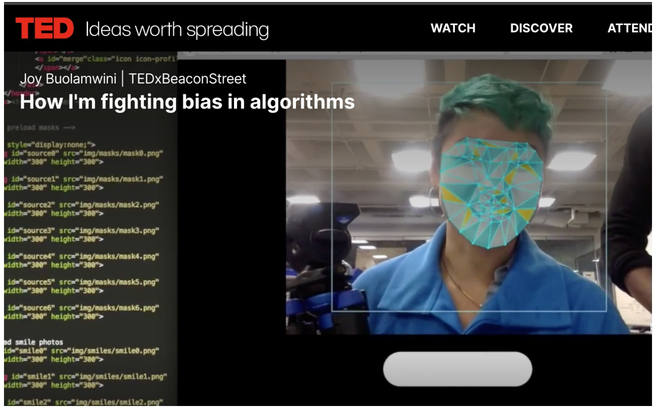
Figure from the Ethical Algorithm illustrating the geometric interpretation of word embedding bias

The ethics of face recognition



Joy Buolamwini | TEDxBeaconStreet

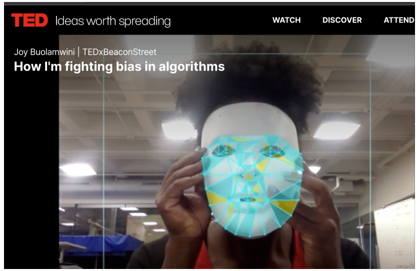
How I'm fighting bias in algorithms



Joy Buolamwini | TEDxBeaconStreet

How I'm fighting bias in algorithms

```
    <a id="merge"class="icon icon-profi
    </span></a>

preload masks -->

style="display:none;">
g id="source0" src="img/masks/mask0.png"
width="300" height="300">

g id="source1" src="img/masks/mask1.png"
width="300" height="300">

 id="source2" src="img/masks/mask2.png"
width="300" height="300">

 id="source3" src="img/masks/mask3.png"
width="300" height="300">

 id="source4" src="img/masks/mask4.png"
width="300" height="300">

 id="source5" src="img/masks/mask5.png"
width="300" height="300">

 id="source6" src="img/masks/mask6.png"
width="300" height="300">

ad smile photos
id="smile0" src="img/smiles/smile0.png"
width="300" height="300">

g id="smile1" src="img/smiles/smile1.png"
width="300" height="300">

 id="smile2" src="img/smiles/smile2.png"
```



Joy Buolamwini | TEDxBeaconStreet

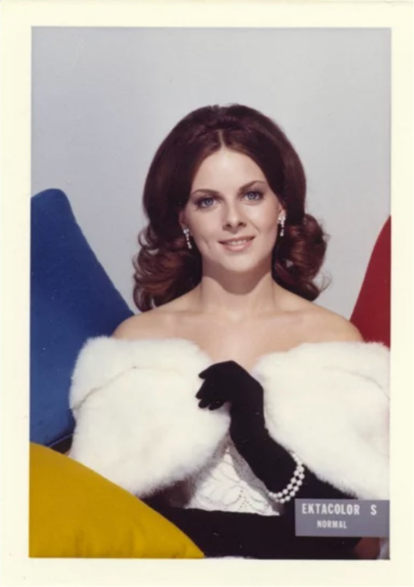
How I'm fighting bias in algorithms

https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Biased calibration – the Kodak story

Film manufacturers and film developers used a test picture as a color-balance benchmark. This test picture became known as the “Shirley card,” named after Shirley Page, a Kodak employee and the first model to pose for it.32 It perhaps goes without saying that Shirley and her successors were overwhelmingly White.

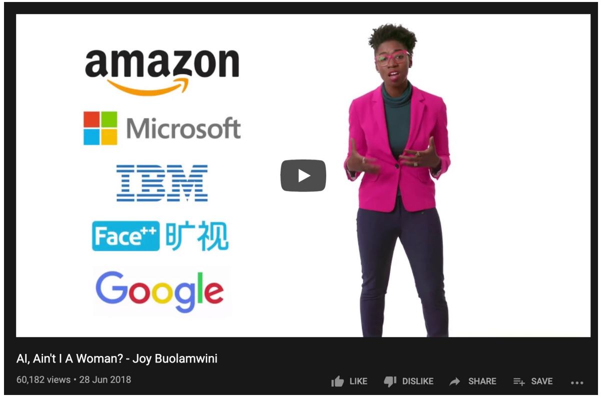
Christian, Brian; Christian, Brian. The Alignment Problem: How Can Machines Learn Human Values? (p. 37). Atlantic Books. Kindle Edition.



The original Shirley card

Systematic analysis of “error parity”

http://gendershades.org/index.html

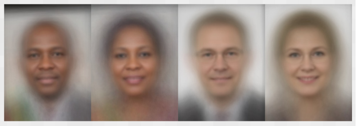


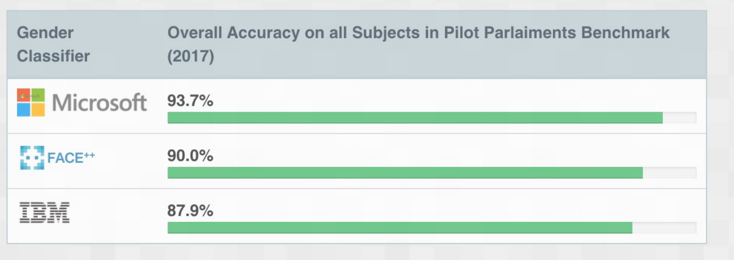
AI, Ain't I A Woman? - Joy Buolamwini

60,182 views • 28 Jun 2018

Systematic analysis of “error parity”

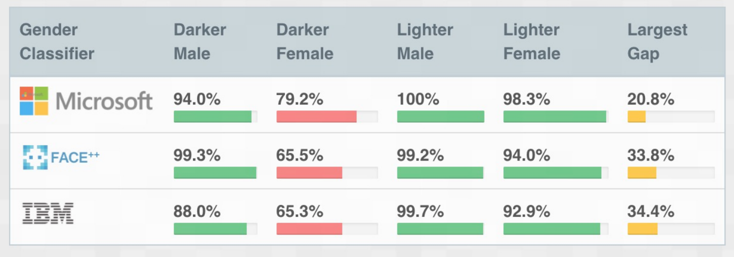
Balanced dataset





Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

Error parity



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%

How easy is it to “fix” face recognition?

- A balanced training dataset might not be sufficient to solve the problem. One class of examples might pose a harder learning problem for various reasons (e.g., have more in-class heterogeneity).
- Solving for error parity will come at a cost of overall accuracy. Are we prepared to pay that price?
- Error parity seems like an objective ethical requirement, but, if the technology in question in a given context is used disproportionately on one population, then should we aim for better accuracy there?
- What if some errors are particularly hurtful (recall example on misclassifying people of color)?

□ Monitoring all the important dimensions of a classifier is a no-regret move.

□ We might need to make hard choices, trading off one objective for another.

□ The very process of coming up with fairness metrics elucidates our ethical values.

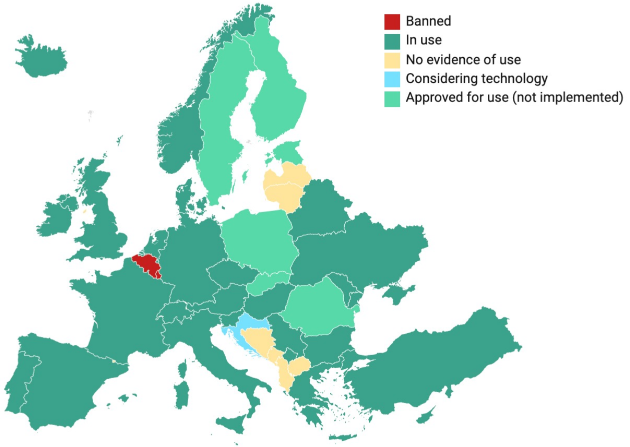
What if the technology isn’t yet safe?

Ban vs moratorium vs sunsetting

If tech is *not* banned, but private companies sunset their services, will that ensure a better or worse outcome for at-risk groups?

The status of facial recognition in Europe

How countries in Europe are implementing the use of facial recognition technology



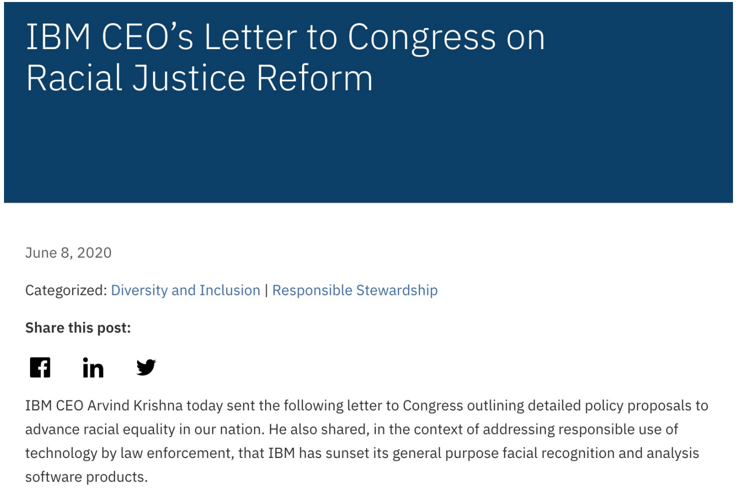
**While not banned by law, Belgium found its use to be in breach of the law and Luxembourg prime minister spoke against it*

Map: EUobserver • Source: Surfshark • Get the data • Created with Datawrapper

What if the technology isn’t yet safe?

Ban vs moratorium vs sunsetting

If tech is *not* banned, but private companies sunset their services, will that ensure a better or worse outcome for at-risk groups?



IBM CEO’s Letter to Congress on Racial Justice Reform

June 8, 2020

Categorized: Diversity and Inclusion | Responsible Stewardship

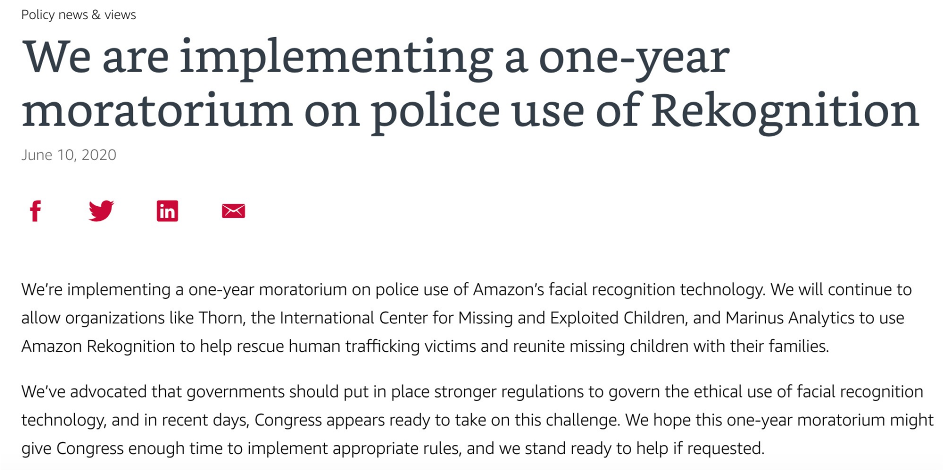
Share this post:

IBM CEO Arvind Krishna today sent the following letter to Congress outlining detailed policy proposals to advance racial equality in our nation. He also shared, in the context of addressing responsible use of technology by law enforcement, that IBM has sunset its general purpose facial recognition and analysis software products.

What if the technology isn't yet safe?

Ban vs moratorium vs sunsetting

If tech is *not* banned, but private companies sunset their services, will that ensure a better or worse outcome for at-risk groups?



Policy news & views

We are implementing a one-year moratorium on police use of Rekognition

June 10, 2020

We're implementing a one-year moratorium on police use of Amazon's facial recognition technology. We will continue to allow organizations like Thorn, the International Center for Missing and Exploited Children, and Marinus Analytics to use Amazon Rekognition to help rescue human trafficking victims and reunite missing children with their families.

We've advocated that governments should put in place stronger regulations to govern the ethical use of facial recognition technology, and in recent days, Congress appears ready to take on this challenge. We hope this one-year moratorium might give Congress enough time to implement appropriate rules, and we stand ready to help if requested.

Summary

- Come up with fairness metrics to monitor alongside accuracy.
- Can be subtler to address in unsupervised learning.
- If a technology is not yet safe, say so clearly.
- Restricting its use might be desirable, though perhaps not enforceable or even counterproductive.
- A classifier might end up with biased performance for a variety of reasons:
  - Training data that are skewed in some way (e.g., selection bias, or inaccurate labels)
  - Imbalanced training data compromising accuracy in the minority class
  - The learning task might have different structure across different groups

Focus on improvements over the current baseline, even if they don’t yet get us to perfect error parity, and be completely transparent about the residual bias by extensive and systematic monitoring.