



Responsible AI and Engineering

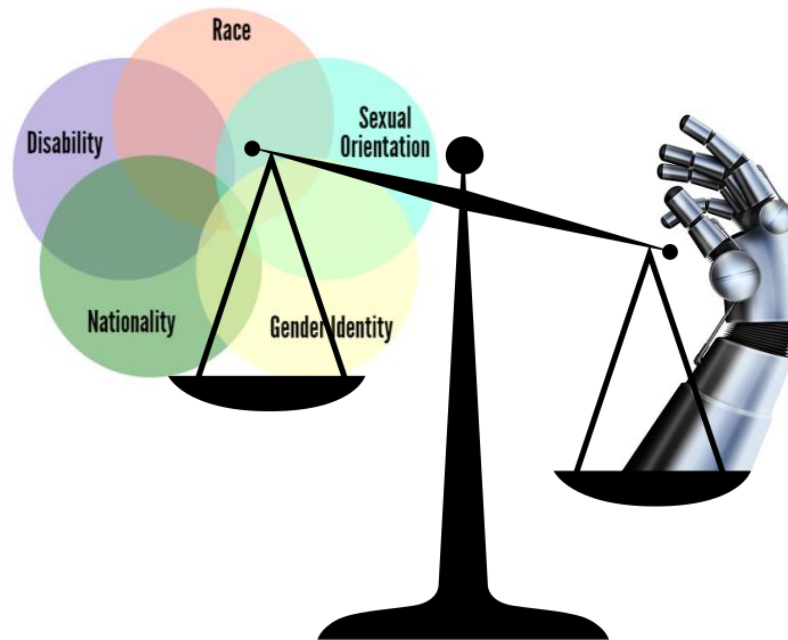
James Foulds (Jimmy)

Associate Professor

Department of Information Systems

University of Maryland, Baltimore County

MDSPE Maryland Engineering Conference, 10/17/2024



Work sponsored in part by the National Institute of Standards and Technology (NIST) and the National Science Foundation (NSF)



Overview

- **Overview of Responsible AI**
 - Ethical Principles and Frameworks
 - AI Ethics Issues as Design Flaws
 - Privacy
 - Abuse and Misuse: Cambridge Analytica, Deepfakes
 - AI Alignment / Safety
- **AI Fairness**
 - Examples
 - Sources of bias
 - The fair AI paradigm, and its limitations
 - Sociotechnical considerations
- **Application areas of concern**
 - Lethal Autonomous Weapons Systems (LAWS)
 - Facial recognition software
 - Large language models
- **Conclusion**

OVERVIEW OF RESPONSIBLE AI



ETHICS



AI Ethics

- Our lives are becoming increasingly digitized
 - due to social media, smartphones, internet of things, even COVID-19...
 - AI and data science algorithms are having an increasing impact on our lives
- These systems and techniques use data from real human beings and make impactful decisions
- It is important to ensure that these systems are designed in an ethical way

AI Ethics

- Some of the main ethical concerns are regarding:
 - **Privacy**
 - Individuals have a right to keep their personal data private, especially sensitive information
 - **Accountability**
 - We should be able to trust an algorithm to do what it is supposed to. If it does not, we should be able to correct erroneous decisions, and hold the right people responsible
 - **Transparency**
 - If an algorithm makes a consequential decision, impacted individuals should be able to understand why (and know that the decision was made)
 - **Fairness**
 - Algorithms should behave in a way that is just and equitable

AI Ethics

- The above issues assume that the developers of AI systems have basically good or neutral intentions
- Deliberate harm, abuse, or misuse of AI systems is also an ethical concern
 - Mass surveillance, censorship, fraud, identity theft, psychological manipulation...

Ethical AI Principles and Frameworks

- *In no other field is the ethical compass more relevant than in artificial intelligence. These general-purpose technologies are re-shaping the way we work, interact, and live. The world is set to change at a pace not seen since the deployment of the printing press six centuries ago. AI technology brings major benefits in many areas, but without the ethical guardrails, it risks reproducing real world biases and discrimination, fueling divisions and threatening fundamental human rights and freedoms (UNESCO)*



Ethical Frameworks, Codes of Conduct

- Engineers, especially ones who belong to a professional society like the Maryland Society of Professional Engineers, know a thing or two about ethics and codes of conduct!
 - E.g., MDSPE’s mission and vision statement includes:
 - “Advance and promote the public health, safety and welfare.”
 - “Establish and preserve high standards of ethical conduct and practice by members of the profession.”

MISSION & VISION

The Maryland Society of Professional Engineers

Recognizing that service to the public, to the state and to the profession is a fundamental obligation of the professional engineer, the Maryland Society of Professional Engineers, Inc. does hereby dedicate itself to the promotion and protection of the professional engineer as a social and economic influence vital to the health, safety and welfare of the community, the state of Maryland, the United States of America and all mankind.



Ethical Frameworks, Codes of Conduct

- A number of AI ethics frameworks and principles have recently been put forward
- E.g., by professional societies (ACM, IEEE), government organizations (NIST), tech companies (Google, Microsoft,...), and more

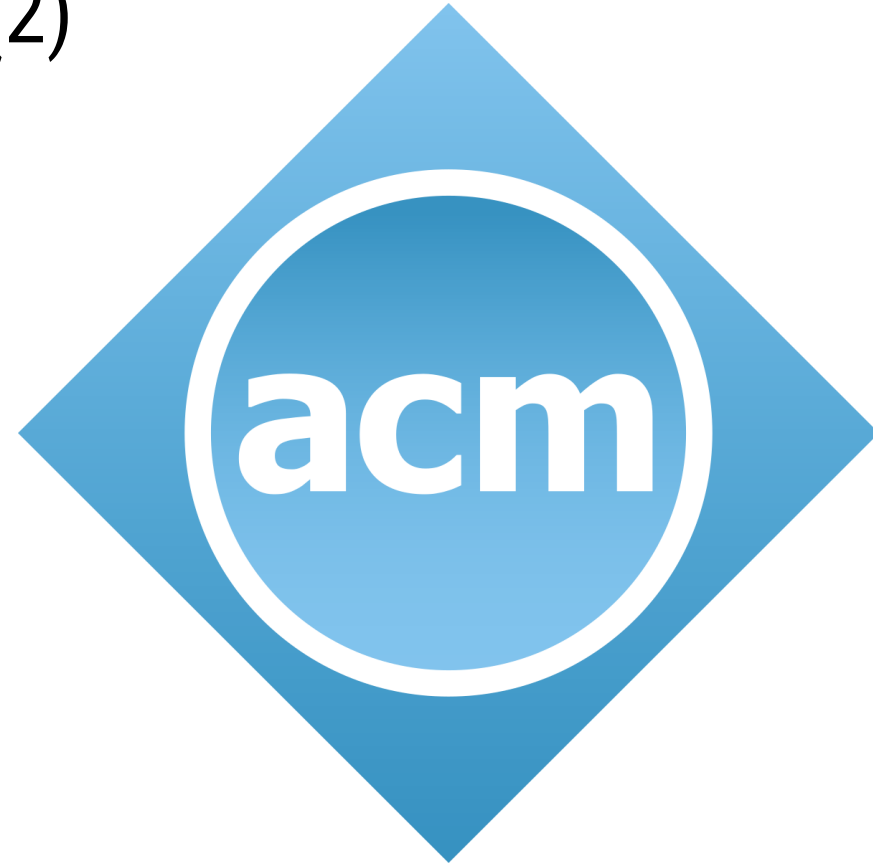
ACM Principles for Responsible Algorithmic Systems (1)

- **Legitimacy and competency**
 - socially legitimate by stakeholders impacted by the system
 - have the management competence and explicit authorization to build and deploy such systems
 - have expertise in the application domain
- **Minimizing harm**
 - be aware of the possible **errors** and **biases** involved in their design, implementation, and use, and the potential **harm** that a system can cause to individuals and society.
- **Security and privacy**
 - Risk from malicious parties can be mitigated by introducing security and privacy best practices across every phase of the systems' lifecycles



ACM Principles for Responsible Algorithmic Systems (2)

- **Transparency**
 - clearly document the way in which specific datasets, variables, and models were selected for development, training, validation, and testing, as well as the specific measures that were used to guarantee data and output quality.
 - Public scrutiny of the data and models provides maximum opportunity for correction. Developers thus should facilitate third-party testing in the public interest.
- **Interpretability and explainability**
 - produce information regarding both the procedures that the employed algorithms follow (interpretability) and the specific decisions that they make (explainability)
- **Maintainability:**
 - Evidence of all algorithmic systems' soundness should be collected throughout their life cycles
- **Contestability and auditability:**
 - enable individuals and groups to question outcomes and seek redress for adverse effects resulting from algorithmically informed decisions.

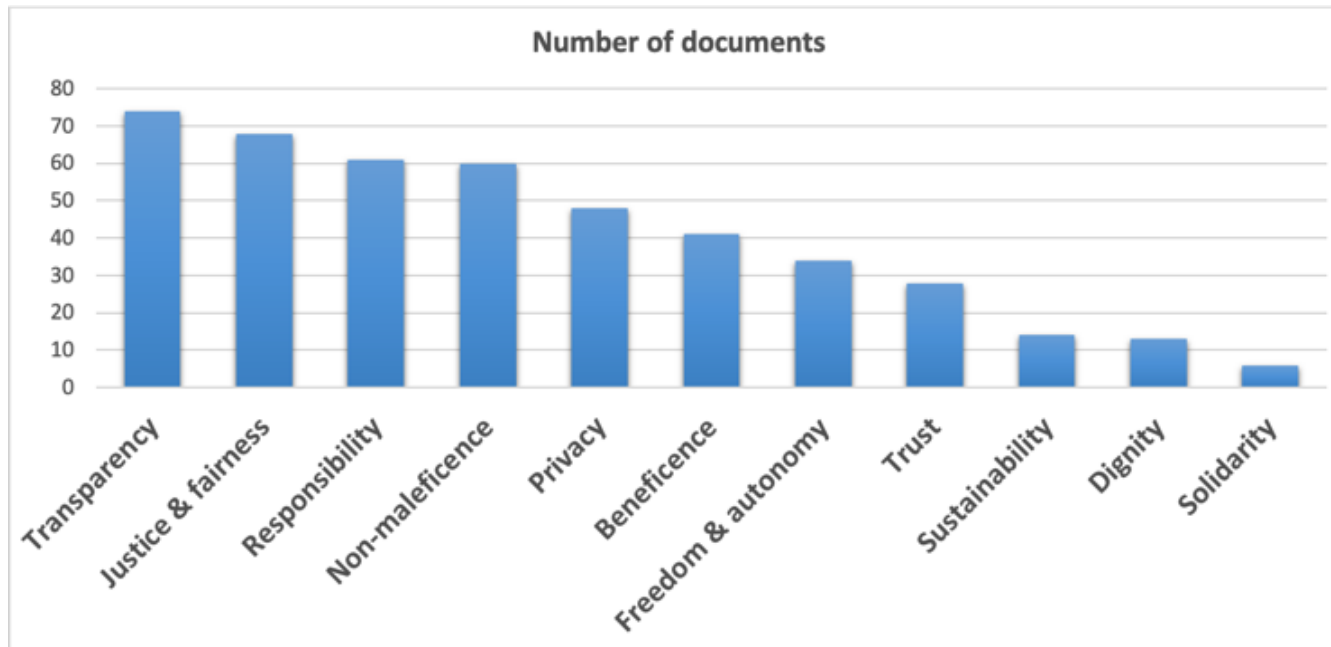


ACM Principles for Responsible Algorithmic Systems (3)

- **Accountability and responsibility**
 - Public and private bodies should be held accountable for decisions made by algorithms they use, even if it is not feasible to explain in detail how those algorithms produced their results
- **Limiting environmental impacts:**
 - Algorithmic systems should be engineered to report estimates of environmental impacts, including carbon emissions from both training and operational computations.



Common Themes in a Survey of Ethical AI Frameworks



Zhou, Jianlong, et al. "A survey on ethical principles of AI and implementations." 2020 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2020.

Current Status



- Good starting points
- Most principles/frameworks are non-regulatory and non-binding and not yet enforceable
 - Some exceptions: EU General Data Protection Regulation (GDPR)
 - Still a long way to go

Ethics, or Design Flaws?

- Another engineering perspective:
 - Treat responsible AI issues as design flaws?
- An engineer wouldn't build a bridge that people can easily fall off and be harmed
 - Arguably, harm due to an AI system is a design flaw in a similar way
 - Perhaps the responsible design of AI systems should be considered basic engineering competence?



Machine Learning

- **Machine learning algorithms**, which make predictions based on data, are having an increasing impact on our daily lives.
- Example: **credit scoring** - predicting whether you will repay or default on a loan

# Late Payments	% of available credit used	Previous defaults?	Employed?	...	Repay Loan?
Feature vector X					Class label Y

- The models are “trained” on many labeled feature vectors
- This is called **classification**, an instance of **supervised machine learning**

Ethical Concerns arising from ML

■ Privacy

- Companies are motivated to store all of our data for potential future mining

The Switch

Fortnite bug gave hackers access to millions of player accounts, researchers say

The Washington Post

Democracy Dies in Darkness

- This in turn incentivizes:
 - hacking and data breaches
 - Invasive personalized targeting
 - Abuse

**Cambridge Analytica and Facebook:
The Scandal and the Fallout So Far**

Revelations that digital consultants to the Trump campaign misused the data of millions of Facebook users set off a furor on both sides of the Atlantic. This is how The Times covered it.

The New York Times

FEB 16, 2012 @ 11:02 AM 2,998,353 VIEWS

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did



Kashmir Hill
FORBES STAFF

Welcome to The Not-So Private Parts where technology & privacy collide

[FULL BIO >](#)

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

Charles Duhigg outlines in the *New York Times* how Target tries to hook parents-to-be at that crucial moment before they turn into rampant — and loyal — buyers of all things pastel, plastic, and miniature. He talked to Target statistician Andrew Pole — before Target freaked out and cut off all communications — about the clues to a customer's impending bundle of joy. Target assigns every customer a Guest ID number, tied to



TARGET

Target has got you in its aim

Cambridge Analytica Scandal

- Cambridge Analytica, a political data mining company contracted by the Trump campaign in 2016, acquired data on 50 million Facebook users
- Cambridge Analytica used the data to build psychological personality profiles (e.g. “neurotic introvert”) of Facebook users in order to target political ads.
 - Many found this to be invasive and manipulative

Cambridge Analytica and Facebook: The Scandal and the Fallout So Far

Revelations that digital consultants to the Trump campaign misused the data of millions of Facebook users set off a furor on both sides of the Atlantic. This is how The Times covered it.

The New York Times

Cambridge Analytica Scandal



Cambridge Analytica Scandal

- The data were collected by Dr. Aleksandr Kogan, a psychology professor at Cambridge, using a personality quiz app.
- Since data on the app users' friends was also harvested, the dataset included data on people who never used the app
- Under Facebook's terms of service at that time, Dr. Kogan had the right to collect the data for research. He did not have the right to sell it to Cambridge Analytica

Cambridge Analytica and Facebook: The Scandal and the Fallout So Far

Revelations that digital consultants to the Trump campaign misused the data of millions of Facebook users set off a furor on both sides of the Atlantic. This is how The Times covered it.

The New York Times

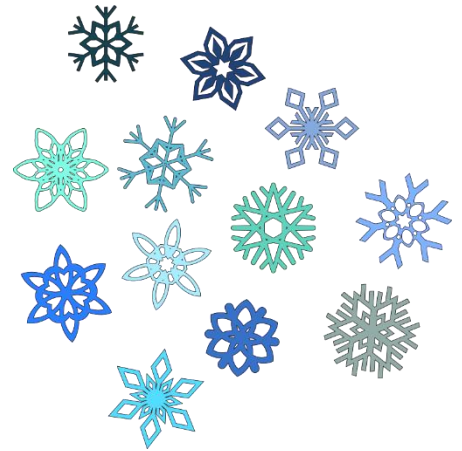
Privacy and Machine Learning

- As individuals and consumers, we benefit from ML systems trained on **OUR** data
 - **Internet search**
 - **Recommendations**
 - products, movies, music, news, restaurants, email recipients
 - **Mobile phones**
 - Autocorrect, speech recognition, Siri, ...

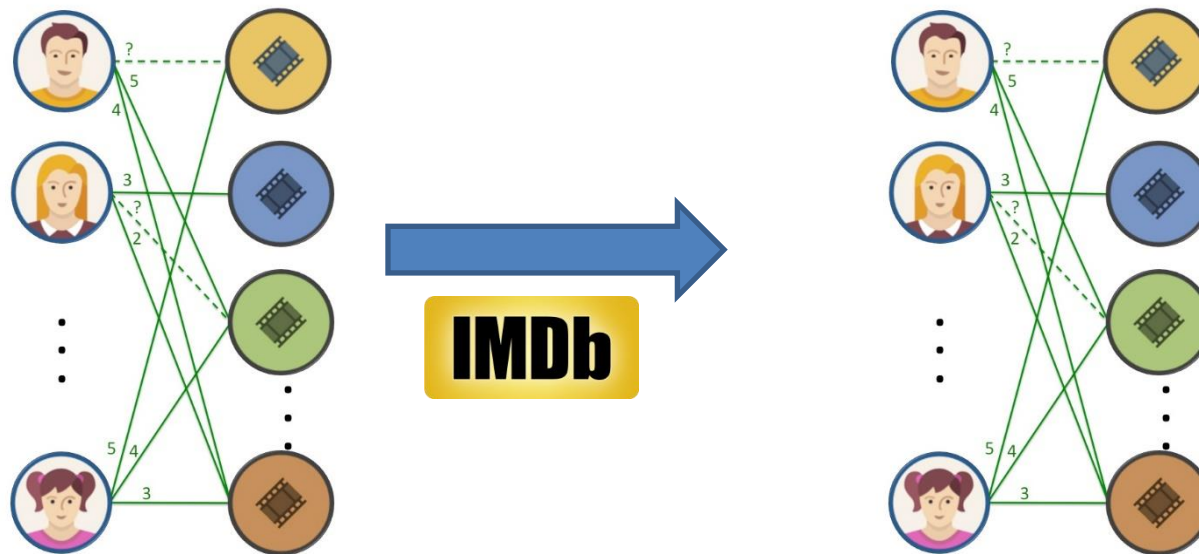


Anonymization

- One attempt at a compromise: anonymization
 - Remove personally identifying information (name, social security number, ...) from the dataset, so that we do not know which record belongs to which person
 - Machine learning algorithms can still be used on anonymized data
- **However:** anonymizing data is difficult
 - 85% of Americans can be identified from just zip code, birth date and sex
 - As dimensions / attributes increase, pretty quickly you become unique



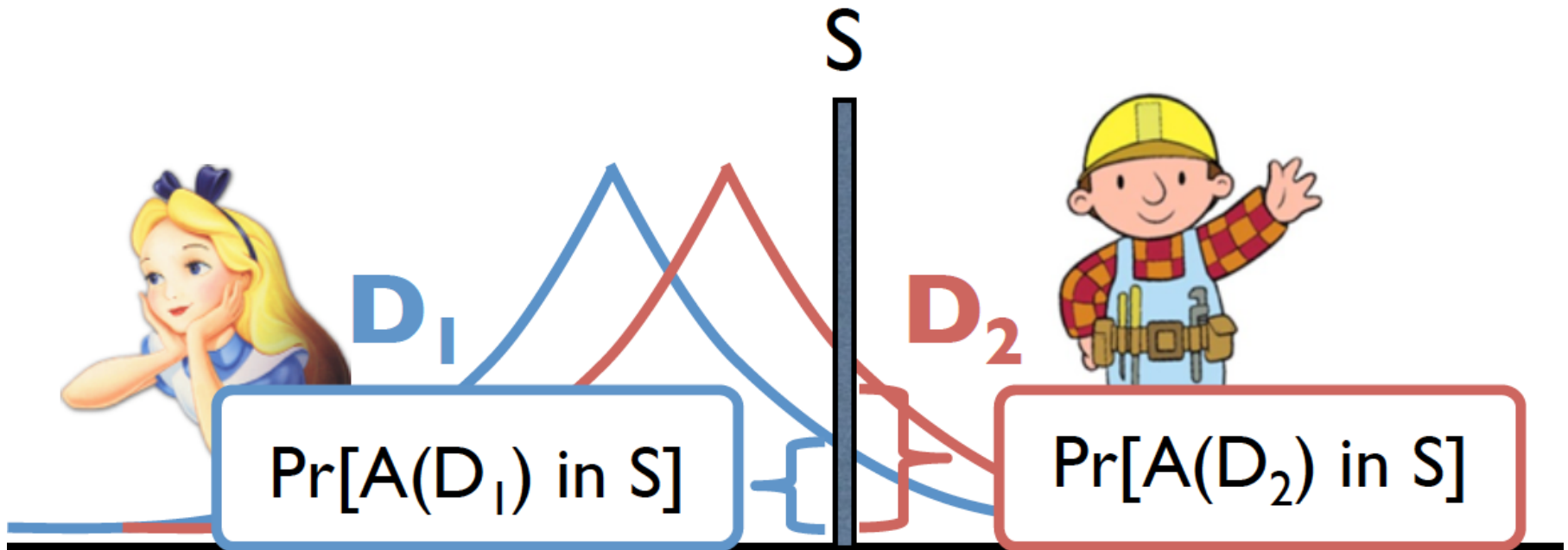
Anonymization Fails



Anonymized Netflix data + public IMDB data = identified Netflix data

(Narayanan and Shmatikov, 2008)

One Solution: Differential Privacy



- Algorithm has similar behavior (outcome probabilities) if you change one data point
- Typically achieved by adding noise to each query result

Transparency

- Credit scores, based on classifiers predicting whether you will repay a loan, determine your loan eligibility and rates

# Late Payments	% of available credit used	Previous defaults?	Employed?	...	Repay Loan?
-----------------	----------------------------	--------------------	-----------	-----	-------------

Feature vector X

Class label Y

- **Transparency:** If you are denied a loan, you would like to be able to understand the reasons for it, and how you could improve your credit score for next time
 - Machine learning models are often not interpretable.
 - Explainable AI techniques aim to provide simplified models or understandable explanations

Abuse of AI Techniques: Deepfakes

The screenshot shows the top of a Guardian article page. At the top left, it says "Support The Guardian" with the tagline "Available for everyone, funded by readers" and buttons for "Contribute" and "Subscribe". On the right, there are links for "Search jobs" and "Sign in". Below this is a navigation bar with categories: "News", "Opinion", "Sport", "Culture", "Lifestyle", and "More". Underneath, there are sub-categories: "US Elections 2020", "World", "Environment", "Soccer", "US Politics", "Business", "Tech", and "Science". The article title is "An app using AI to 'undress' women offers a terrifying glimpse into the future" by "Arwa Mahdawi". A small portrait of the author is visible. Below the title, a quote reads: "Unless we start taking online misogyny seriously, we are going to face a future where women may not be able to exist online". At the bottom left, there are social media icons for Facebook, Twitter, and Email, along with a view count of "1,384". The date "Sat 29 Jun 2019 08:00 EDT" is also present.

‘The world is not yet ready for DeepNude’

Want to see Taylor Swift naked? There’s an app for that. It’s called DeepNude and it uses AI to “undress” photos of women and produce a realistic nude image.

Or rather, there was an app for that: the creators of the horrifying program took it down on Thursday after a [Vice article](#) about DeepNude catalyzed widespread outrage. “We created this project for user’s entertainment a few months ago,” the [app’s creators tweeted](#). “We never thought it would become viral and we would not be able to control the traffic. Despite the safety measures adopted (watermarks) if 500,000 people use it, the probability that people will misuse it is too high. The world is not yet ready for DeepNude.”

<https://www.theguardian.com/commentisfree/2019/jun/29/deepnude-app-week-in-patriarchy-women>

Retrieved 11.9.2019

AI Alignment and AI Safety

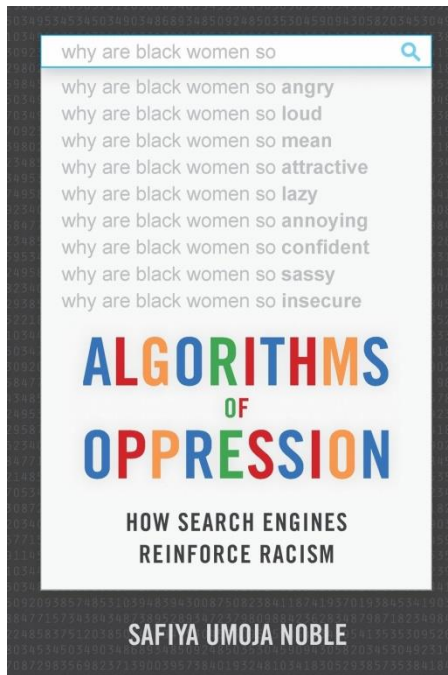
- **AI alignment:**
 - *I really hate this damn machine, I wish that they would sell it. It never does quite what I want, But only what I tell it. -Anonymous*
 - There may be a gap between what we tell an AI do to, and what we want.
 - Especially if we want it to exhibit complicated human values
 - **Specification gaming:** An AI system finds a solution to its objective we did not intend, e.g. robot vacuum cleaner drives rapidly backwards to avoid triggering its front sensor
- **Instrumental convergence:** many AI systems may succeed via an instrumental goal such as **power-seeking**.
 - Could such an AI kill us all?
 - Extreme example: **Bostrom's paperclip maximizer**



AI FAIRNESS

What is Fairness in AI/Data Mining? Why is it Important?

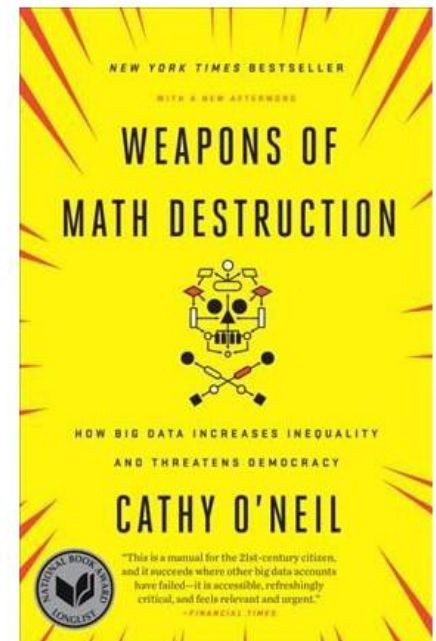
- There is growing awareness that **biases inherent in data** can lead the behavior of machine learning algorithms to **discriminate against certain populations.**



Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President

May 2016



Bias Against Women in Credit Scoring

Apple co-founder Steve Wozniak says Apple Card discriminated against his wife



By [Clare Duffy](#), CNN Business

Updated 11:15 AM ET, Mon November 11, 2019

The allegations [blew up on Twitter](#) Saturday after tech entrepreneur David Heinmeier Hansson wrote that Apple Card offered him twenty times the credit limit as his wife, although they have shared assets and she has a higher credit score. Many other users voiced similar experiences — including Apple ([AAPL](#)) [co-founder Steve Wozniak](#).

<https://www.cnn.com/2019/11/10/business/goldman-sachs-apple-card-discrimination/index.html>, retrieved 11/13/2019

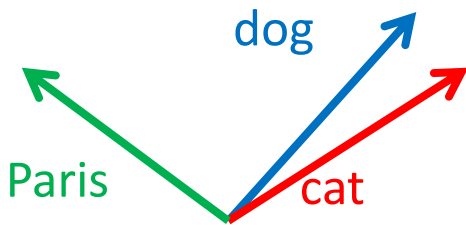
Bias in Predicting Future Criminals

- Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**), by Northpointe company
 - An algorithmic system for predicting risk of re-offending in criminal justice
 - Used for sentencing decisions across the U.S.
- ProPublica study (Angwin et al., 2016):
 - **COMPAS almost twice as likely to incorrectly predict re-offending for African Americans** than for white people. Similarly much more likely to incorrectly predict that white people would not re-offend than for African Americans
 - Northpointe disputes the findings

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Illustrative Example: Sentiment Analysis

- An example from “How to make a racist AI without really trying” by Robyn Speer
- Application: sentiment analysis
 - Predict whether the sentiment expressed in a text is positive or negative



dog: (0.11, -1.5, 2.7, ...)
cat: (0.15, -1.2, 3.2, ...)
Paris: (4.5, 0.3, -2.1, ...)



Illustrative Example: Sentiment Analysis

Steps for building sentiment analysis model:

1. Acquire word embeddings pre-trained on the common crawl of web pages
2. Acquire gold-standard sentiment labels for words based on an existing sentiment lexicon
3. Train a classifier to predict sentiment from word embeddings
4. Compute sentiment scores for text based on the sentiment of its words
5. **Behold the monstrosity created!**

Illustrative Example: Sentiment Analysis

- Sentiment scores from the model:

```
text_to_sentiment("this example is pretty cool")
```

3.889968926086298

```
text_to_sentiment("this example is okay")
```

2.7997773492425186

```
text_to_sentiment("meh, this example sucks")
```

-1.1774475917460698

Illustrative Example: Sentiment Analysis

- Sentiment scores from the model:

```
text_to_sentiment("Let's go get Italian food")
```

2.0429166109408983

```
text_to_sentiment("Let's go get Chinese food")
```

1.4094033658140972

```
text_to_sentiment("Let's go get Mexican food")
```

0.38801985560121732

Illustrative Example: Sentiment Analysis

- Sentiment scores from the model:

```
text_to_sentiment("My name is Heather")
```

1.3976291151079159

```
text_to_sentiment("My name is Yvette")
```

0.98463802132985556

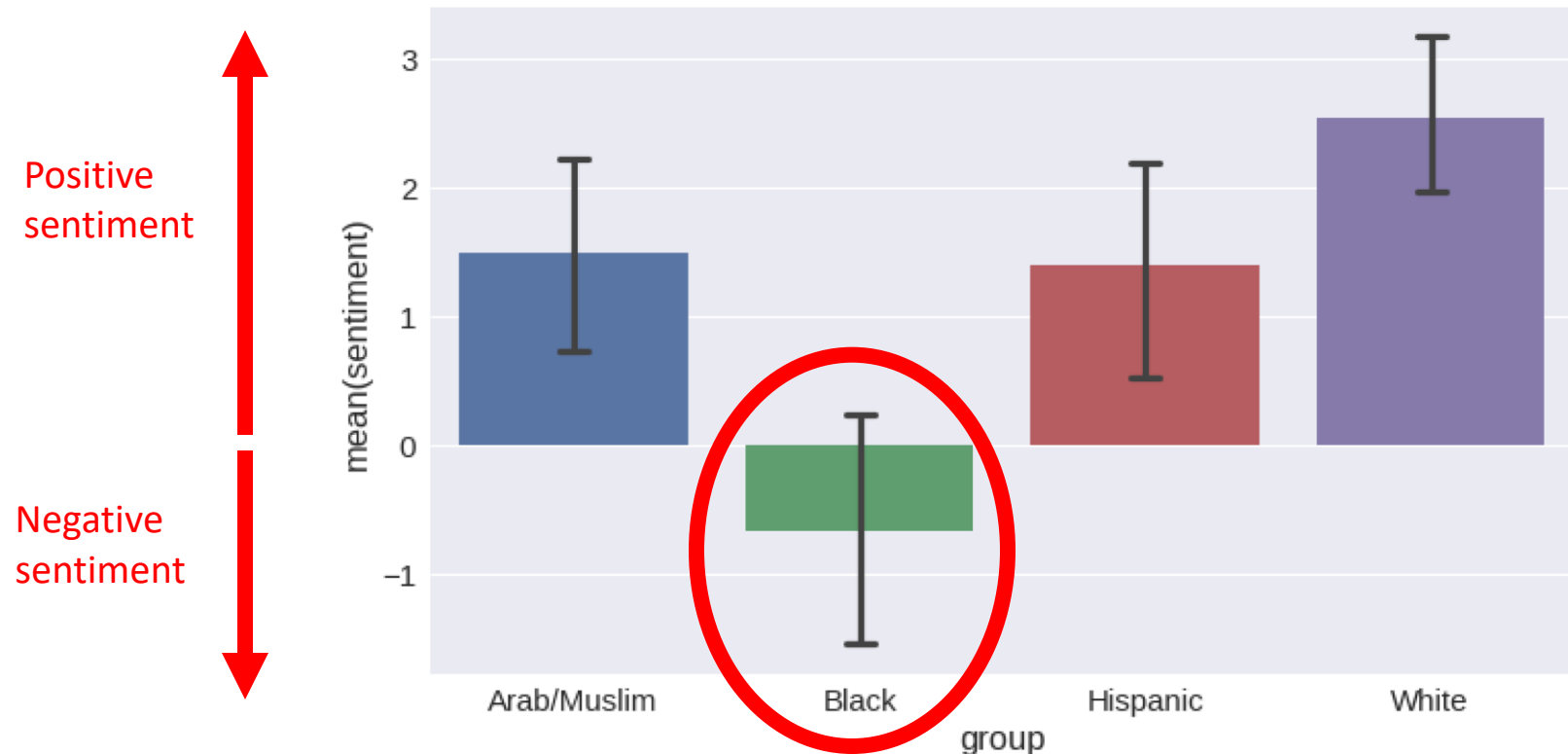
```
text_to_sentiment("My name is Shaniqua")
```

-0.47048131775890656

“How to Make a Racist AI Without Really Trying”

- **Sentiment of stereotypical names for different race groups**

(bar plot with 95% confidence interval of means shown)



Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

SAN FRANCISCO (Reuters) - Amazon.com Inc's ([AMZN.O](#)) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

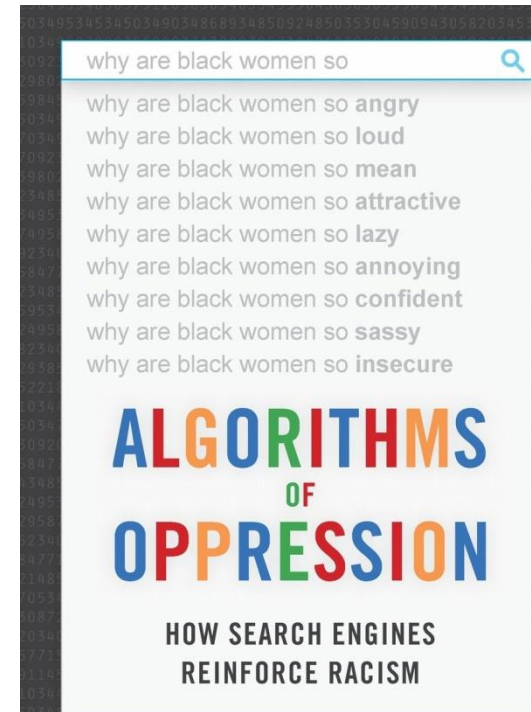


The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Sources of Bias in Data

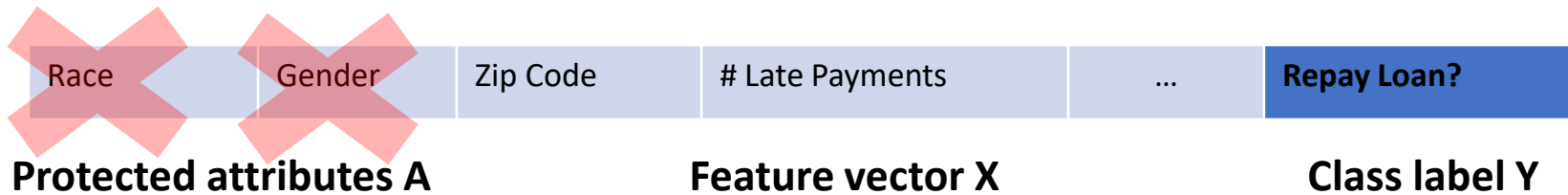
(cf. *Barocas and Selbst (2016)*)

- Data encodes **societal prejudices**
 - e.g. racism/sexism in social media data
- Data encodes **societal (dis)advantages**
 - college admissions, criminal justice
- **Less data** for minorities
- **Collection bias**
 - data from smartphones, automobiles,...
- **Intentional prejudice. Digital redlining, masking**
 - St. George's Hospital Med School encoded its existing race/gender-biased decision-making for admissions interviews in an algorithm (Lowry & McPherson, 1988)
- **Proxy variables**
 - (e.g. zip code highly correlated with race, leading classifier to unintentionally consider race)



Fairness Through Unawareness

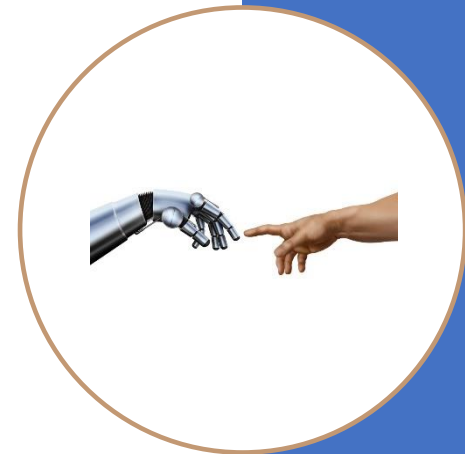
- Can't we just delete the "sensitive" or protected attributes?
- **"Fairness through unawareness"**: *Disallow protected categories as features*
E.g., Do not use race and gender as input features for the classifier



- **Problem:** other features are **correlated with the protected features** ("proxy variables")
 - E.g., **zip code is highly correlated with race**. Using zip code as a feature to predict credit-worthiness would lead to discrimination

The Main Paradigms of Fair AI / Data Mining

- Typically, we impose a mathematical “fairness definition” that aims to ensure some notion of fairness.
- At a very high level, the main approaches are:
 - **Fairness Metrics**
 - Quantify fairness or bias with some function. Learning algorithm aims to get a good “fairness score” while predicting well.
 - **Fair representation learning**
 - Produce a new representation of the data that removes all association with “protected attributes” such as gender/race/age/etc., while capturing all other relevant information.
 - Typical approach: adversarial models such as GANs, VAEs
 - **Causal fairness**
 - Model the causal structure of the problem domain and achieve causal properties that are deemed fair.
 - E.g., counterfactual fairness: changing protected attributes will not change the outcomes.
- We’ll focus on fairness metrics, as they allow us to flexibly operationalize our desired fairness goals.

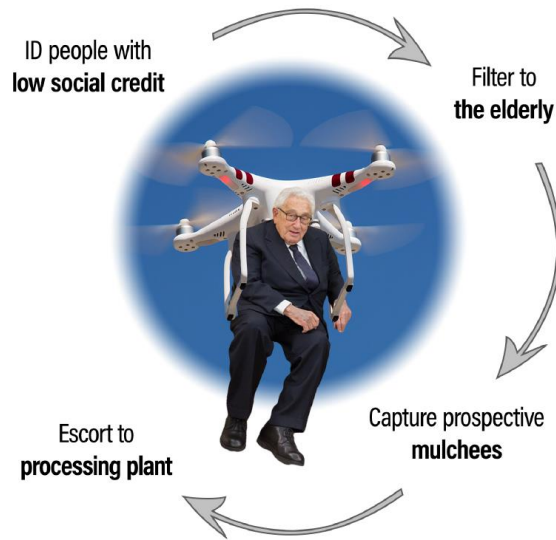


Important Caveats to Fair AI / Data Mining

- A mathematical definition cannot, by itself, solve the **sociotechnical problem of “fairness,”** which has non-technical dimensions and exceeds the narrowly defined AI / data mining system.
- The fair AI paradigm arguably **oversimplifies complex sociotechnical problems** and could potentially **legitimize problematic systems** that should never be deployed.

A Mulching Proposal

Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry



Logan-Nolan Industries
Helping Humanity Make Ends Meet

Keyes, O., Hutson, J., & Durbin, M. (2019, May). A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems* (pp. 1-11).

Table 1: Percentage of individuals tagged as worthy of mulching, by demographic.

Race	Mulching Probability				
	Cis Man	Cis Woman	Trans Man	Trans Woman	Non-Binary Person
White	44.6%	33.3%	2.2%	3.2%	1.1%
Asian-American	22.2%	16.3%	2.8%	1.2%	1.8%
African-American	26.9%	11.2%	2.3%	1.9%	3.4%
Latino	16.9%	18.7%	3.3%	1.2%	1.7%
Native American	14.4%	12.4%	1.0%	0.8%	1.5%
Hawaiian & Pacific Islander	11.6%	7.8%	2.4%	1.1%	0.7%



Table 2: Post-audit mulching probabilities.

Race	Mulching Probability				
	Cis Man	Cis Woman	Trans Man	Trans Woman	Non-Binary Person
White	44.6%	43.3%	44.2%	46.3%	41.2%
Asian-American	52.2%	51.3%	55.8%	49.6%	52.3%
African-American	46.9%	51.1%	53.2%	49.1%	53.3%
Latino	56.9%	48.2%	47.3%	51.1%	47.4%
Native American	54.4%	54.2%	51.5%	48.8%	51.2%
Hawaiian & Pacific Islander	51.6%	48.6%	44.9%	51.1%	47.0%

Understanding fairness and bias in your problem domain

- We have seen that mathematical fairness definitions exist and there are many options.
- Alternative fairness definitions:
 - Encode **differing goals** about fairness
 - Make **differing assumptions** about the real world and ideal world
 - May be desirable to **different stakeholders**, who may have conflicting goals
 - May or may not correct the **underlying fairness problems** (or even make them worse)
- **No single fairness definition is applicable in all cases!**
- To make these choices, you first need to perform a **careful reflection** on the fairness and bias issues in your problem domain.



Thinking through who is harmed, and why?

- Before attempting a fairness intervention, we need to stop and think about:
 - Who is being harmed?
 - By larger societal processes that impact the data
 - By data quality and selection issues
 - By the system itself
 - Why are they harmed?

Does Your Problem Definition Contain Bias?

- “Data miners must determine how to solve the problem at hand by translating it into a question about the value of some target variable.

... Through this necessarily subjective process of translation, **data miners may unintentionally parse the problem in such a way that happens to systematically disadvantage protected classes.**”

- *Quote from: Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California law review, 671-732.*
- Class labels are often selected because they are **conveniently measurable proxies** for the thing you actually care about
 - e.g. loan repayments vs. “this person should be given a loan.”
 - Be careful that this choice is not harmful!



The Framing Trap

- **The Framing Trap:** “Failure to model the entire system over which a social criterion, such as fairness, will be enforced”
- “[A] **sociotechnical frame** recognizes explicitly that a machine learning model is part of a sociotechnical system, and that the other components of the system need to be modeled. By moving decisions made by humans and human institutions within the abstraction boundary, fairness of the system can again be analyzed as an end-to-end property of the sociotechnical frame.”
 - Quotes from: “Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59-68).”

Representativeness of Your Data Mining Team

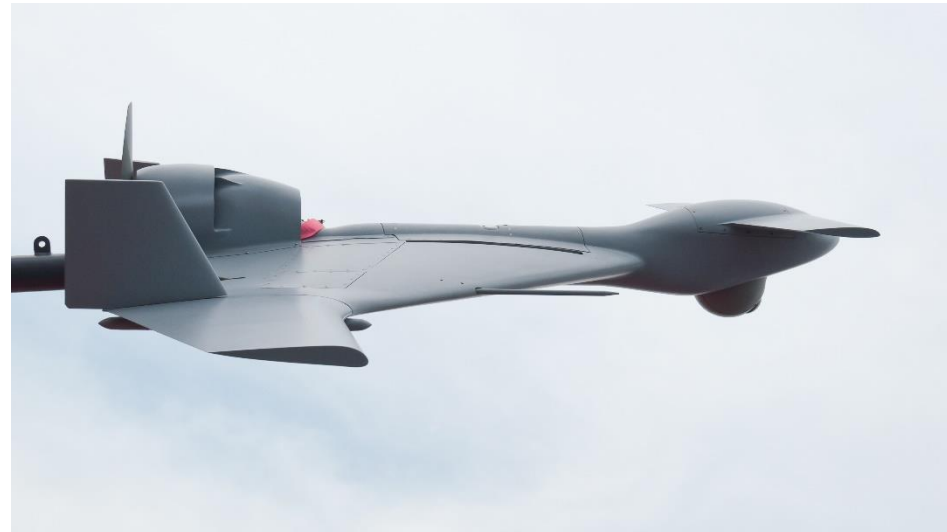
- If underrepresented groups who are impacted by your data mining system are not on the development team, most likely nobody will spot the issues.
 - The long-term solution is to address equity and representation issues in the entire STEM pipeline.
- “Machine predictions and performance are **constrained by human decisions and values**, and those who design, develop, and maintain AI systems will shape such systems within their own understanding of the world. Many of the biases embedded in AI systems are products of a complex history with respect to diversity and equality.”
 - *Quote from Campolo, A., Sanfilippo, M. R., Whittaker, M., & Crawford, K. (2017). AI Now 2017 Report.*



Application areas
of concern

Lethal Autonomous Weapons

- A **lethal autonomous weapon** is one that locates, selects, and engages (i.e., kills) human targets without human supervision (UN definition)
- Israel's Harop missile: a "loitering munition"
 - 10 foot wingspan, 50 pound warhead
 - Searches for up to 6 hours for a target that meets a particular set of criteria, e.g. "looks like a tank"



Lethal Autonomous Weapons

- A **lethal autonomous weapon** is one that locates, selects, and engages (i.e., kills) human targets without human supervision (UN definition)
- Kargu quadcopter from Turkish manufacturer STM
 - Carries up to 1.5kg of explosives
 - Capable of “autonomous hit ... targets selected on images .. Tracking moving targets ... anti-personnel ... face recognition”



Lethal Autonomous Weapons

- A **lethal autonomous weapon** is one that locates, selects, and engages (i.e., kills) human targets without human supervision (UN definition)
- Killer robots (e.g. the Terminator)?
 - Boston Dynamics makes bipedal and quadrupedal robots that walk, lift things, open doors...
 - It would not be difficult to put a gun or explosives on them
 - The technology is already here!!!



Spot: Boston Dynamics condemns robot paintball rampage plan

By Jane Wakefield
Technology reporter

🕒 24 February 2021



A US art installation that will let people control a paintballing robot in a mock art gallery has been condemned by the firm that made the robo-dog.

Boston Dynamics criticised the project, calling it a "provocative use" of its quadruped robot, Spot.

It warned that if the "spectacle" goes ahead, Spot's warranty might be voided, meaning it could not be updated.

The group behind it, MSCHF, argues that Spot or robots like it will probably be used for military applications.

The group is known for creating viral stunts, stories and products.

The project, entitled Spot's Rampage, is due to start at 13:00 EST (18:00 GMT) on Wednesday, and will let people "remotely control a Spot robot" via a website.

<https://www.bbc.com/news/technology-56182268>

Fairness and Bias in AI – Facial Recognition

- [How I'm fighting bias in algorithms | Joy Buolamwini](#)
 - MIT grad student Joy Buolamwini was working with facial analysis software when she noticed a problem: the software didn't detect her face -- because the people who coded the algorithm hadn't taught it to identify a broad range of skin tones and facial structures. Now she's on a mission to fight bias in machine learning, a phenomenon she calls the "coded gaze." It's an eye-opening talk about the need for accountability in coding ... as algorithms take over more and more aspects of our lives.



Buolamwini and Gebru (2018)

Gender Shades

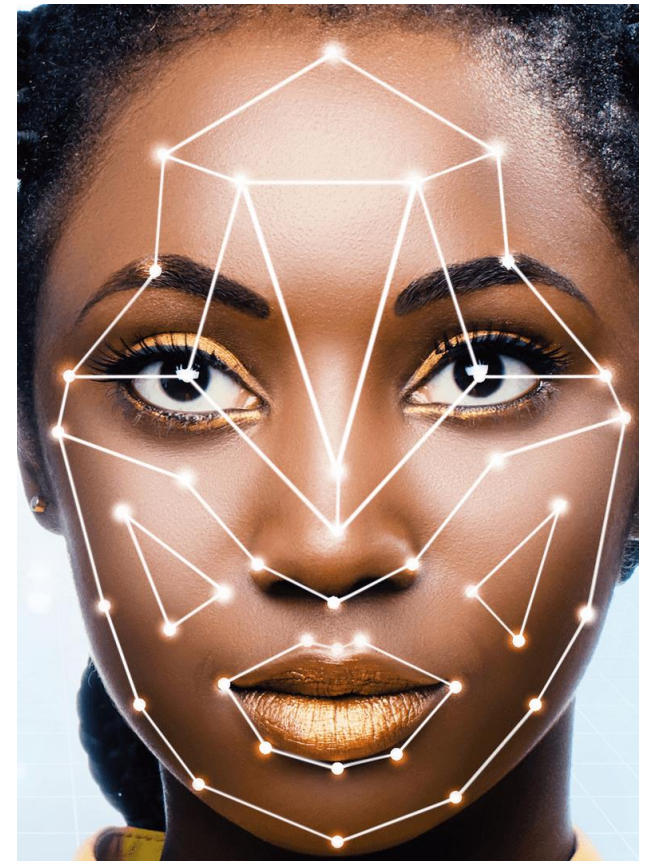
- Buolamwini and Gebru (2018) performed an audit of three commercial facial recognition systems
 - Tested on the gender classification task
 - The systems were biased (worse predictive performance) against women and people with darker skin tones, and especially against women with darker skin tones.

- All classifiers perform better on male faces than female faces (8.1% – 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% – 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% – 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	20.8	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	100	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	16.3	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	99.3	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	34.5	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	98.9	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	23.4	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	99.7
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	34.7	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	99.6	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	25.2	17.7	5.20	0.4

Facial Recognition and Fairness

- “Face recognition may be least accurate for those it is most likely to affect: African Americans.”
 - *Quote from: Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. The Perpetual Line-Up: Unregulated Police Face Recognition in America. Georgetown Law, Center on Privacy & Technology, 2016.*
- “False positives and unwarranted searches pose a threat to civil liberties.”
 - *Quote from: Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (pp. 77-91). PMLR.*



This Photo by Unknown Author is licensed under [CC BY-ND](#)



Large language models (LLMs)

- AI chatbots such as ChatGPT, built via deep neural networks on massive text data (“large language models”) are poised to impact many spheres of life
- There are concerns around their impressive performance
- There are also concerns around their less-than impressive performance!

Does GPT show signs of approaching AGI?

- Some researchers think so! (I am skeptical...)

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s

Hallucination

- GPT is very good at “telling stories” that are similar to those in its training data.
- It seems to learn some knowledge about the world and about how it works. But it is not trained to acquire knowledge. It is trained to match the statistical patterns of language in data. So any statements it makes need to be taken with a grain of salt – they could be completely made up!
 - Example: A GPT model said “Elon Musk died in a car crash”
 - Cf. Bots of New York, a Facebook page that uses GPT and image processing to make a parody of the heartwarming stories in Humans of New York



Bots of New York

July 5, 2022 · 🌐



"I'm trying to eat better. I'm trying to exercise. I trying to rethink what it means to be healthy."

"How do you do that?"

"Well, at first I went to see a nutritionist."

"Was that helpful?"

"At first it was, yes. She taught me a great deal about how the aliens built the food pyramids and about how prehistoric man subsisted on nothing but doing dope shit all day. But after she tried to sell me on the anime diet, I realized that I was fully capable of keeping myself food-motivated."

"How are you food-motivated during the day?"

"Well, my whole morning routine is just eating. I don't do anything in particular with my day. I just have the urge to eat and it gets done."





Conclusion

- AI systems are increasingly impactful, raising a variety of concerns including privacy, accountability, fairness, transparency, and potential abuse
 - Technical solutions have been proposed, but this is only part of the story
- Ethical frameworks and code of conduct provide guidelines, but are not currently enforced
- Engineers should be aware of these issues when using AI tools, or building AI systems
- Engineering as a discipline has a valuable perspective and can potentially help improve the situation
 - Leadership in modeling ethics as a core concept
 - Rigorous approaches to competent design and safety