AI Ethics

CS229: Machine Learning

Carlos Guestrin
Stanford University

The Ethics of AI

- Thus far, we focused on methods and techniques

- But, the systems we build impact people, everyday
- The ethics of AI focuses on the principles and methods to help ensure our systems reflect our values
  - There are social, political and legal implications
    - But, we'll focus on methods for the next two lectures
- Much more too learn
  - See CS281 – Ethics of AI in Spring 2022

Are Emily and Greg More Employable than Lakisha and Jamal? [Bertrand & Mullainathan '03]

ML-based system for recruiting

- Could decrease this bias…

- But, could also amplify biases…

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

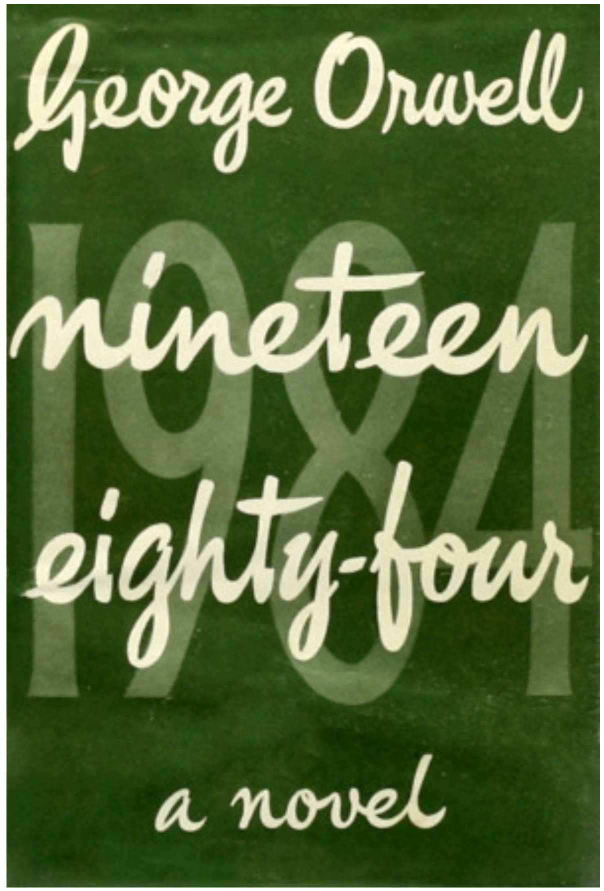
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Ethical Concerns of Artificial Intelligence

The most challenging ethical questions in AI are bound by nuanced complex tradeoffs

Privacy and Survaillance


George Orwell
1984
nineteen
eighty-four
a novel

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE

BRIAN BARRETT LILY HAY NEWMAN SECURITY SEP 3, 2021 12:50 PM

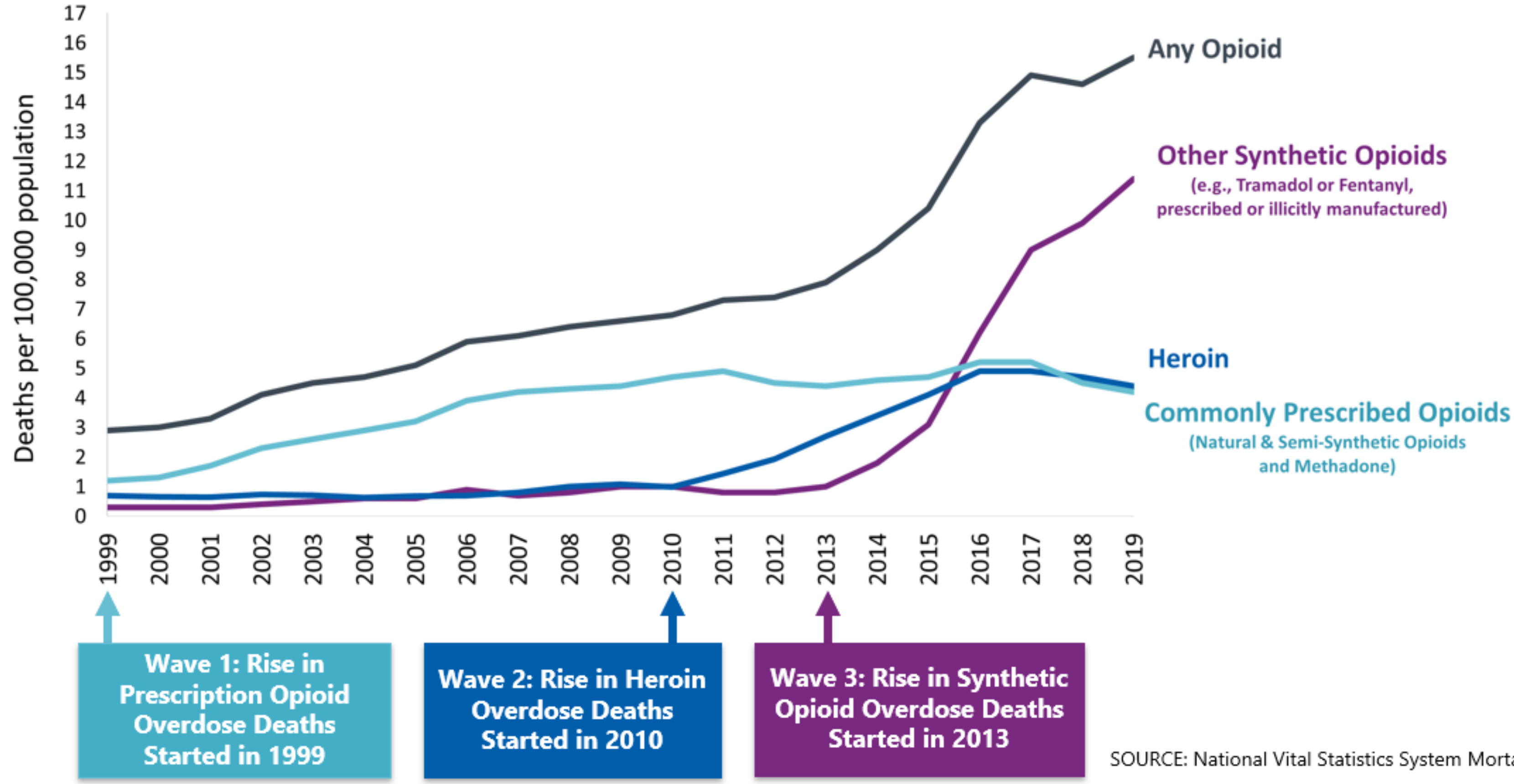
Apple Backs Down on Its Controversial Photo-Scanning Plans

A sustained backlash against a new system to look for child sexual abuse materials on user devices has led the company to hit pause.

Privacy advocates and security researchers are cautiously optimistic about the pause. PHOTOGRAPH: JUSTIN SULLIVAN/GETTY IMAGES

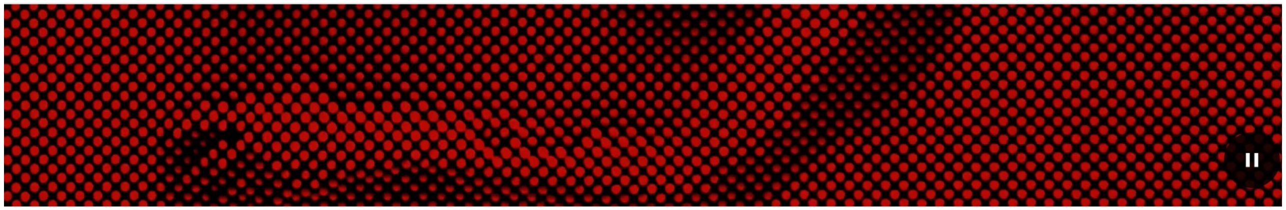
Opacity of Predictions

Three Waves of the Rise in Opioid Overdose Deaths



SOURCE: National Vital Statistics System Mortality File.

WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY SIGN IN SUBSCRIBE



VIDEO: SAM CANNON

MAIA SZALAVITZ BACKCHANNEL AUG 11, 2021 6:00 AM

The Pain Was Unbearable. So Why Did Doctors Turn Her Away?

A sweeping drug addiction risk algorithm has become central to how the US handles the opioid crisis. It may only be making the crisis worse.

The AI Database →

APPLICATION: ETHICS, PREDICTION, REGULATION

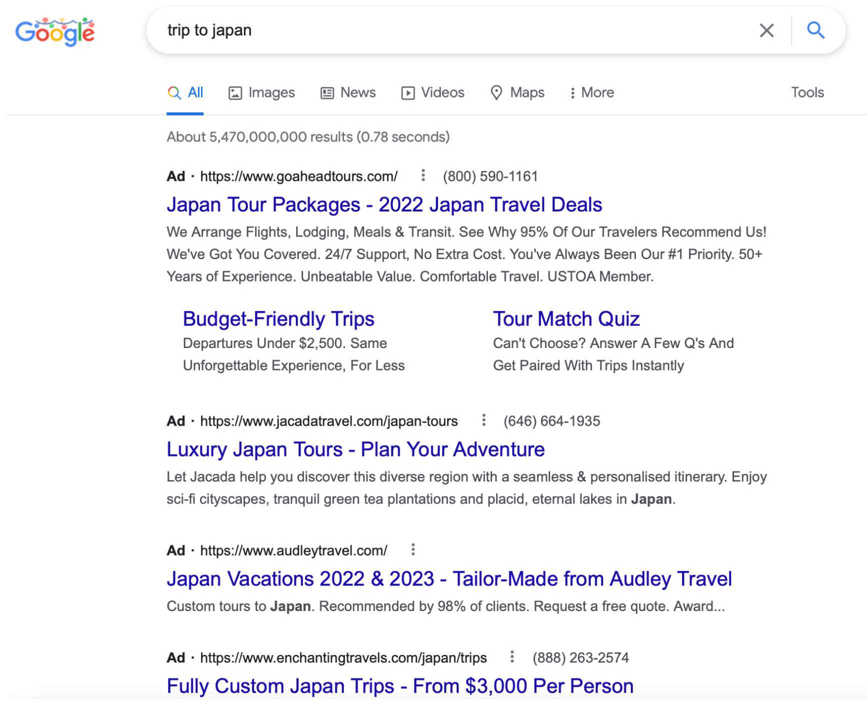
SECTOR: HEALTH CARE, PUBLIC SAFETY

ONE EVENING IN July of 2020, a woman named Kathryn went to the hospital in excruciating pain.

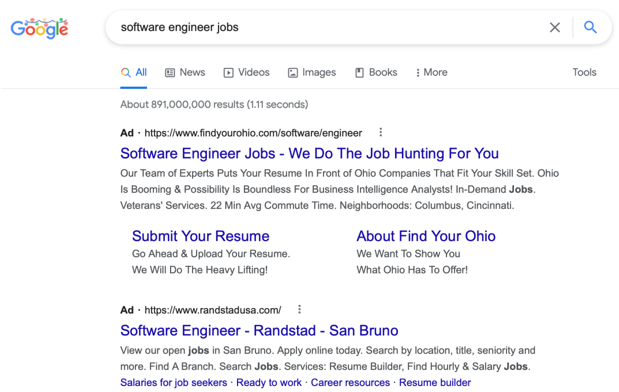
A 32-year-old psychology grad student in Michigan, Kathryn lived with endometriosis, an agonizing condition that causes uterine-like cells to abnormally develop in the wrong

Biased Decisions

Ads can be annoying…



Ads can represent opportunity…



- Ads targeted (using ML) based on predicted features of users…
- Some users don't get the "opportunity" of the ad…

Manipulation of Behavior

"It will be almost as convenient to search for some bit of truth concealed in nature as it will be to find it hidden away in an immense multitude of bound volumes."

- Denis Diderot, 1755

NEWS & POLITICS CULTURE FOOD **salon** SCIENCE & HEALTH LIFE STORIES VIDEO

Search...

EXPLAINER

How "engagement" makes you vulnerable to manipulation and misinformation on social media

Algorithms that rank and recommend posts based on "likes," shares and comments tend to amplify low-quality content

By FILIPPO MENCZER PUBLISHED SEPTEMBER 18, 2021 9:00PM (EDT)

Automation and Employment

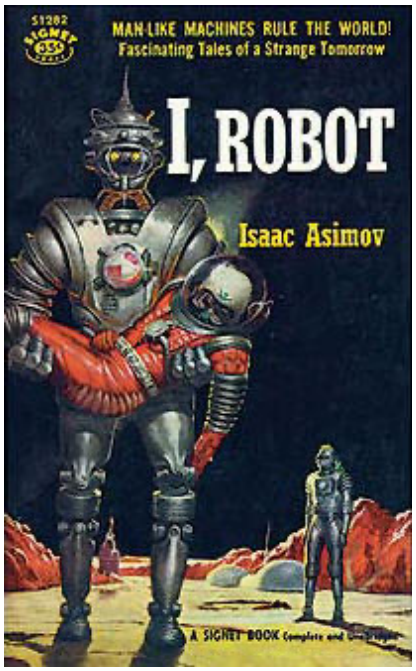
TIME   **SPOTLIGHT STORY** UKRAINIAN WOMEN ARE MOBILIZING BEYOND THE BATTLEFIELD   SIGN IN   SUBSCR

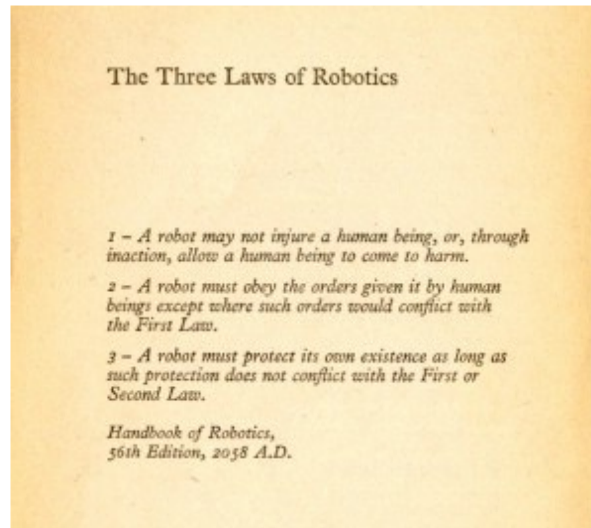
I Worked at an Amazon Fulfillment Center; They Treat Workers Like Robots


2078
1348
1348
1374
1375

https://www.youtube.com/watch?v=4sEVX4mPuto

Decisions by Proxy





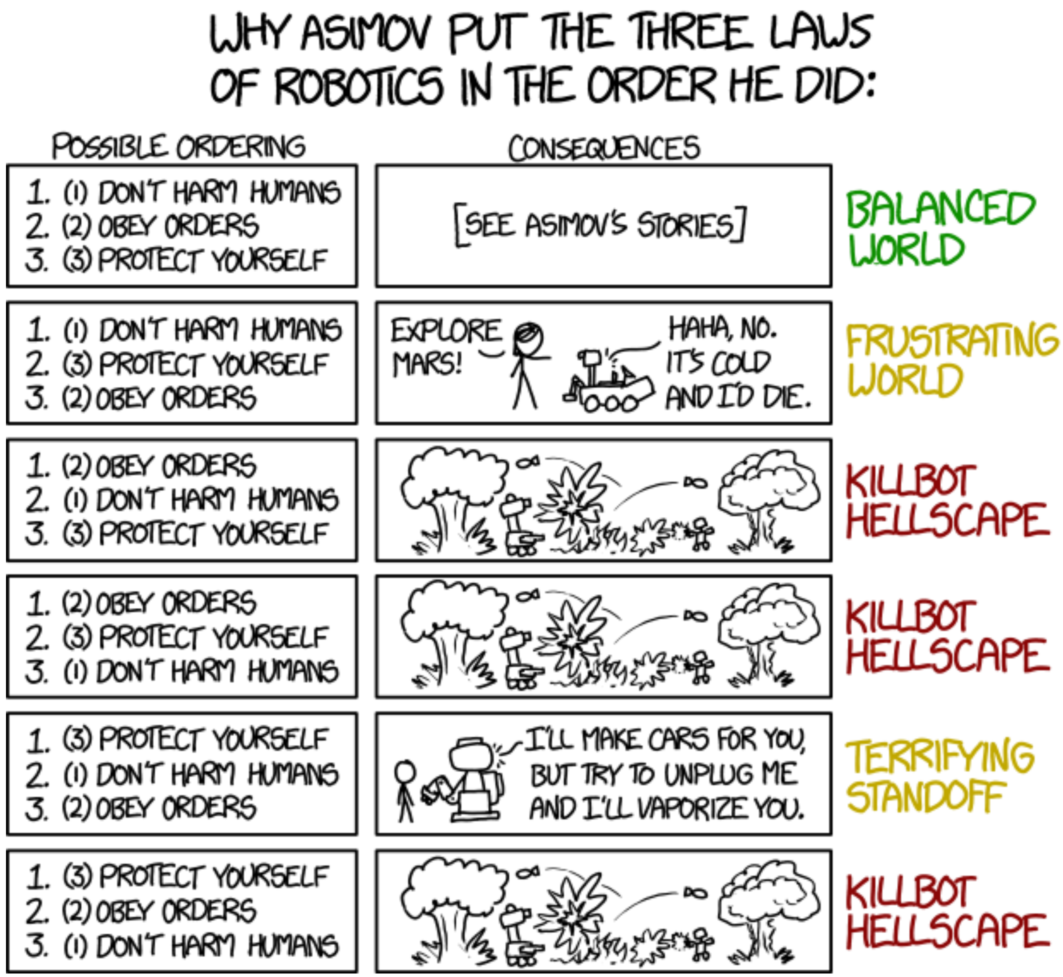
The Three Laws of Robotics

1 – A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

2 – A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

3 – A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

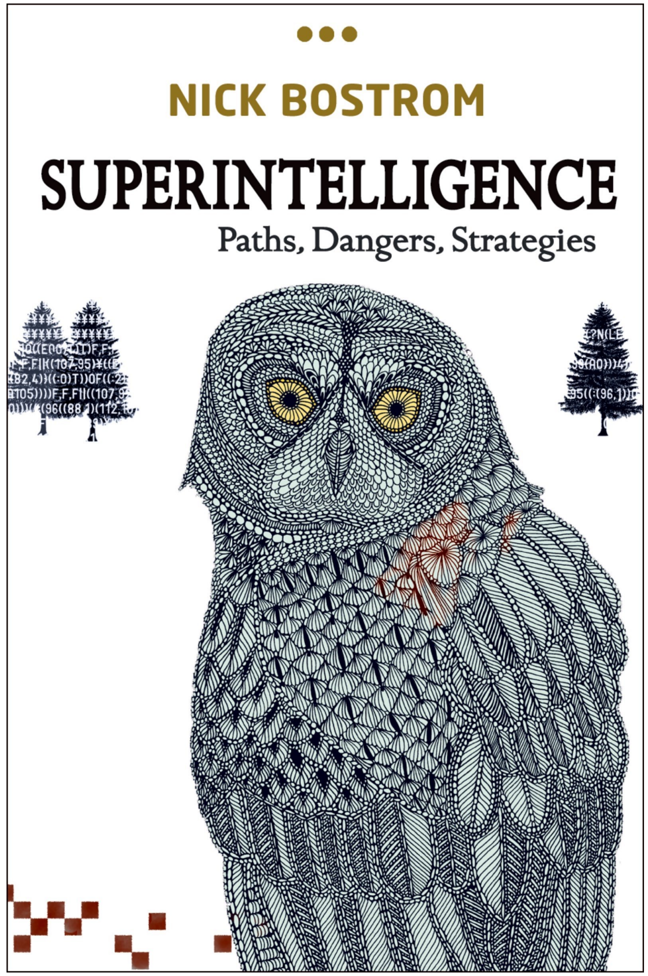
Handbook of Robotics,
56th Edition, 2058 A.D.



https://xkcd.com/1613/

https://www.youtube.com/watch?v=Mme2Aya_6Bc

Existential Risk


NICK BOSTROM
SUPERINTELLIGENCE
Paths, Dangers, Strategies

Focus of Next 2 Lectures

- Fairness and algorithmic bias
- Explainability
- Privacy

AI Ethics:
Fairness & Algorithmic Bias

CS229: Machine Learning
Carlos Guestrin
Stanford University

Regulated Domains in the US

- **Credit** (Equal Credit Opportunity Act)
- **Education** (Civil Rights Act of 1964; Education Amendments of 1972)
- **Employment** (Civil Rights Act of 1964)
- **Housing** (Fair Housing Act)
- **‘Public Accommodation’** (Civil Rights Act of 1964)

Legally-Recognized Protected Classes in the US

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964);**National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967);**Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Sources of Bias

Sources of Bias: Human Bias

- Data reflects human decisions and biases
- Example: ML for Hiring decisions
  - Data from previous hiring decisions perpetuates existing biases

    If managers are biased $\Rightarrow$ ML could also be biased

  - Could reduce bias by measuring employee success
    - Harder to measure and institutional biases can impact success

Sources of Bias: Negative Feedback Loops

- Data collected in biased fashion
  - Negative feedback loop: future observations confirm predictions and reduce further contradicting evidence
- Example: Allocation of police attention based on prevalence of crime

Neighborhoods

	1	2
$t=0$	ε crime	$\varepsilon/2$ crime
	more police	less police
$t=1$	10ε	$\varepsilon/10$

Sources of Bias: Sample Size Disparity

- Models for minority group may be less accurate, if less data is used
- Example: Race representation in medical studies

if study composition

⇒ less likely to be able to evaluate treatment.
less likely to detect side effects for hispanics/blacks

Sources of Bias: Unreliable Data

- If data from minority groups is less reliable or less informative
  - Models may be less accurate for minority groups
  - (Beneficial) interventions may less available to minority groups
- Examples:
  - Inaccurate census in predominantly minority neighborhoods
  - Medical interventions with limited diagnostic tools

Sources of Bias: Proxies

- Even if sensitive attributes (e.g., gender or race) are not used by model, there may be other proxy features that are correlated with sensitive attributes
- Example: Redlining in loan and insurance applications
  - https://www.npr.org/sections/thetwo-way/2016/10/19/498536077/interactive-redlining-map-zooms-in-on-americas-history-of-discrimination
  - https://www.npr.org/2017/05/03/526655831/a-forgotten-history-of-how-the-u-s-government-segregated-america

Mitigating Bias at Every Stage

- Problem definition
- Data collection
- Model development ← many ML papers only focus here
- Model evaluation
- Use of predictions in practice
- Feedback loops

How do we measure fairness?

Consider a loan application…

- x – features of applicant (address, credit history,…)
- c – sensitive features of applicant (gender, race,…)
- d – decision (loan approved or denied) $d(x, c) \in \{0, 1\}$
- y – (hidden) true target in decision (will this person pay the loan)

- **Shorthand probability notation:** $P(y \mid x, c) = P_c(y \mid x)$
- **"Perfect" predictor:** $d = y$

Fairness through Unawareness

- Definition: ignore sensitive features

  $d(x,c) = d(x)$

- Desirable properties: Intuitive, simple, some legal support.
- Criticisms: Proxies !!

  x is correlated with c, e.g., zipcode & race

Three Important Fairness Criteria

- Independence
- Separation
- Sufficiency

All these criteria are achievable…

- Techniques include:
  - Pre-processing
  - Changing training procedure
  - Post-processing

e.g., discussed in CS281

1. Independence

- Definition: Decision d independent of sensitive features **c**

$$d \perp c \Rightarrow \forall i,j \quad P_{c=i}(d=1) = P_{c=j}(d=1)$$

- A.k.a. **demographic parity**: Probability of loan approved is the same across sensitive attributes

$$P_{Black}(\text{loan yes}) = P_{White}(\text{loan yes})$$

Note: fraction of applicants need not be ~~does~~ the same for all races.

$$P(c=\text{White}) \neq P(c=\text{Black})$$

$d \perp c$

Independence: Desirable Properties

- Simple
- Some legal support "4/5 rule"
- In some settings, can increase representation, e.g., in admissions

if before: $P_{Black}(d=1) << P_{White}(d=1)$

Now: $P_{Black}(d=1) = P_{White}(d=1)$

$d \perp c$

Independence: Shortcomings

- Ignores possible correlations between y and **c**
  - Precludes perfect predictor d=y

if $P_{c=0}(y=1) \neq P_{c=1}(y=1)$

- Laziness: quality of decision doesn't need to be uniformly good between groups

For $c=0$ $d=y$

$c=1$ random d, as long as $P_{c=0}(d=1) = P_{c=1}(d=1)$

2. Separation

$d \perp c \mid y$

- Definition: decision d and sensitive features **c** conditionally independent given true target y

$c - y - d$

$\forall c, d \quad \forall i, j \qquad P_{c=i}(d \mid y) = P_{c=j}(d \mid y)$

Variant of Separation: False negative rate parity

- Probability of loan denied for a deserving applicant is the same across sensitive attributes

FNR $P(d=0 \mid y=1)$

FNR Parity $\forall i, j \quad P_{c=i}(d=0 \mid y=1) = P_{c=j}(d=0 \mid y=1)$

Separation: Confusion Matrix Interpretation (Equalized Odds, Equal Opportunity)

- Separation: $P_{c=i}(d \mid y) = P_{c=j}(d \mid y)$
- Confusion matrix:

$y \backslash d$	0	1
0	TN	FP
1	FN	TP

all entries same for all groups

- Variants:

FNR parity, FPR parity

Equal opportunity: TPR parity $P_{c=i}(d=1 \mid y=1) = P_{c=j}(d=1 \mid y=1)$

Separation: Desirable Properties

- Optimality compatibility

  $d = y$ is allowed

- Incentivize to reduce errors equally across groups

Separation: Shortcomings

- Can amplify disparities

	C=0	C=1
Num candidates	100	100
Num qualified y=1	58	2

if 30 slots, equal opportunity: ⇒ 29 from C=0
1 from C=1

if job is well paid, more children from C=0 will have access to better education, even more qualified candidates from C=0

3. Sufficiency

$y \perp c \mid d$

$c - d - y$

- Definition: decision variable d is sufficient to predict target y, independently of sensitive features **c**

$\forall y,d \;\; \forall i,j \quad P_{c=i}(y \mid d) = P_{c=j}(y \mid d)$

- Equivalently, predictive rate parity:
  - Positive predictive rate:

    $P_{c=i}(y=1 \mid d=1) = P_{c=j}(y=1 \mid d=1)$

  - Negative predictive rate:

    $P_{c=i}(y=0 \mid d=0) = P_{c=j}(y=0 \mid d=0)$

decision d is consistent with goals of employer / bank

Sufficiency: Desirable Properties

- Optimality compatibility:

$d = y$ is allowed

- Equal chance of success, given acceptance:

$$P_{Black}(y=1 \mid d=1) = P_{White}(y=1 \mid d=1)$$

Sufficiency: Shortcomings

- Also can amplify disparities

Same example as separation

All these criteria are achievable…

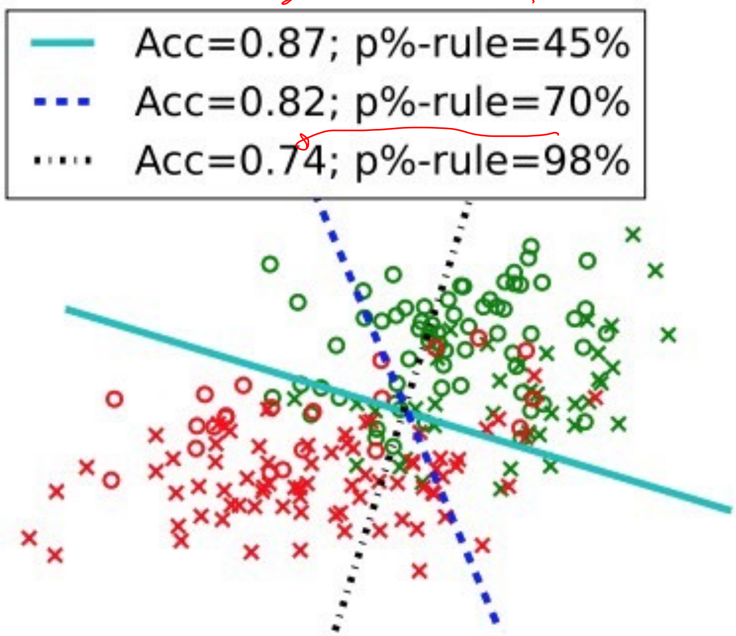
- Techniques include:
  - Pre-processing
  - Changing training procedure
  - Post-processing

Trade-offs are Inevitable

Tradeoff Between Fairness and Accuracy

Tradeoff Between Group-Specific Performance and Average-Case Performance

independence



Accuracy vs demographic parity [Zafar et al. AISTATS2017]

Impossibility Result

- Independence, Separation & Sufficiency are reasonable criteria
- **Theorem:** Any two of these is mutually exclusive!!
  - Except for degenerate cases

XOR Independence

XOR

Separation Sufficiency

XOR

Trade-offs are necessary!

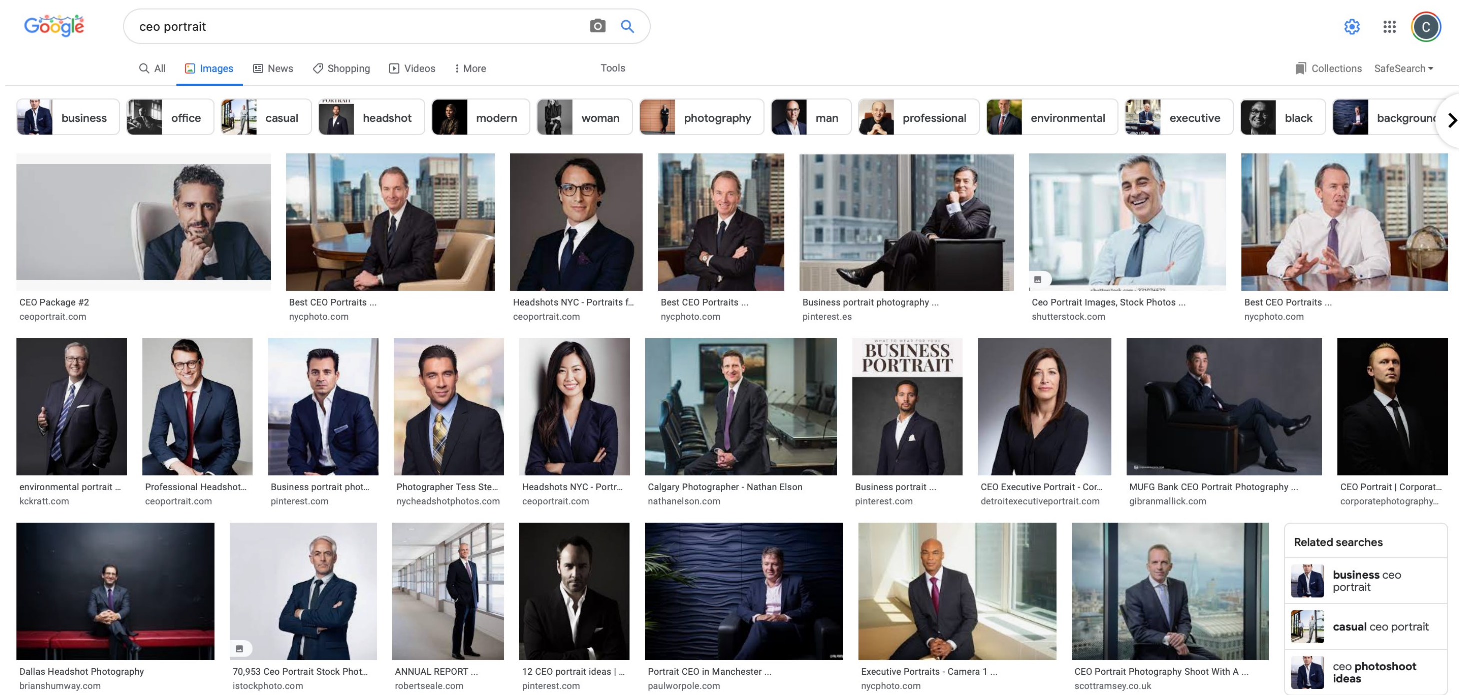
- Choose a criteria, instead of others?
  - Which one?
- Choose a balance between criteria?

- Very general issue in fairness and ML

What are we teaching our models?

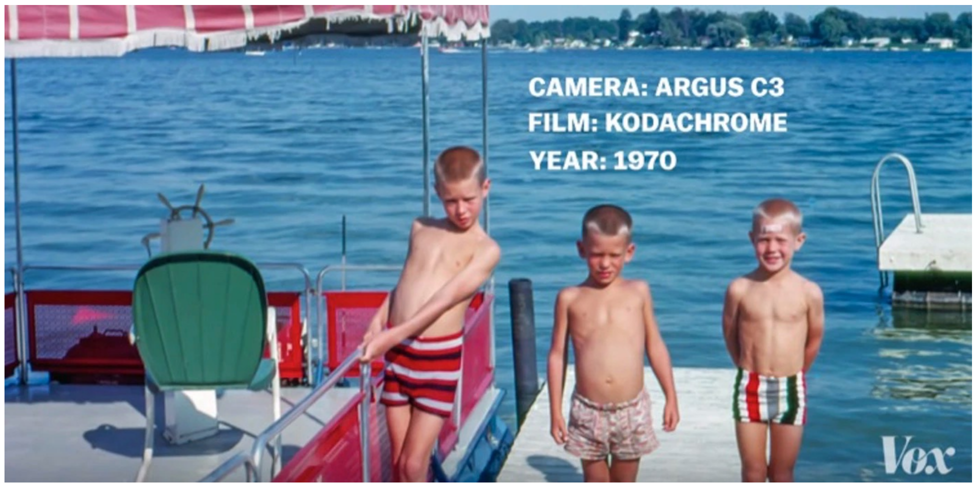

PAW
PATROL

ML perpetuates stereotypes…



The choice of data defines decisions of ML model

Source: www.vox.com/2015/9/18/9348821/photography-race-bias

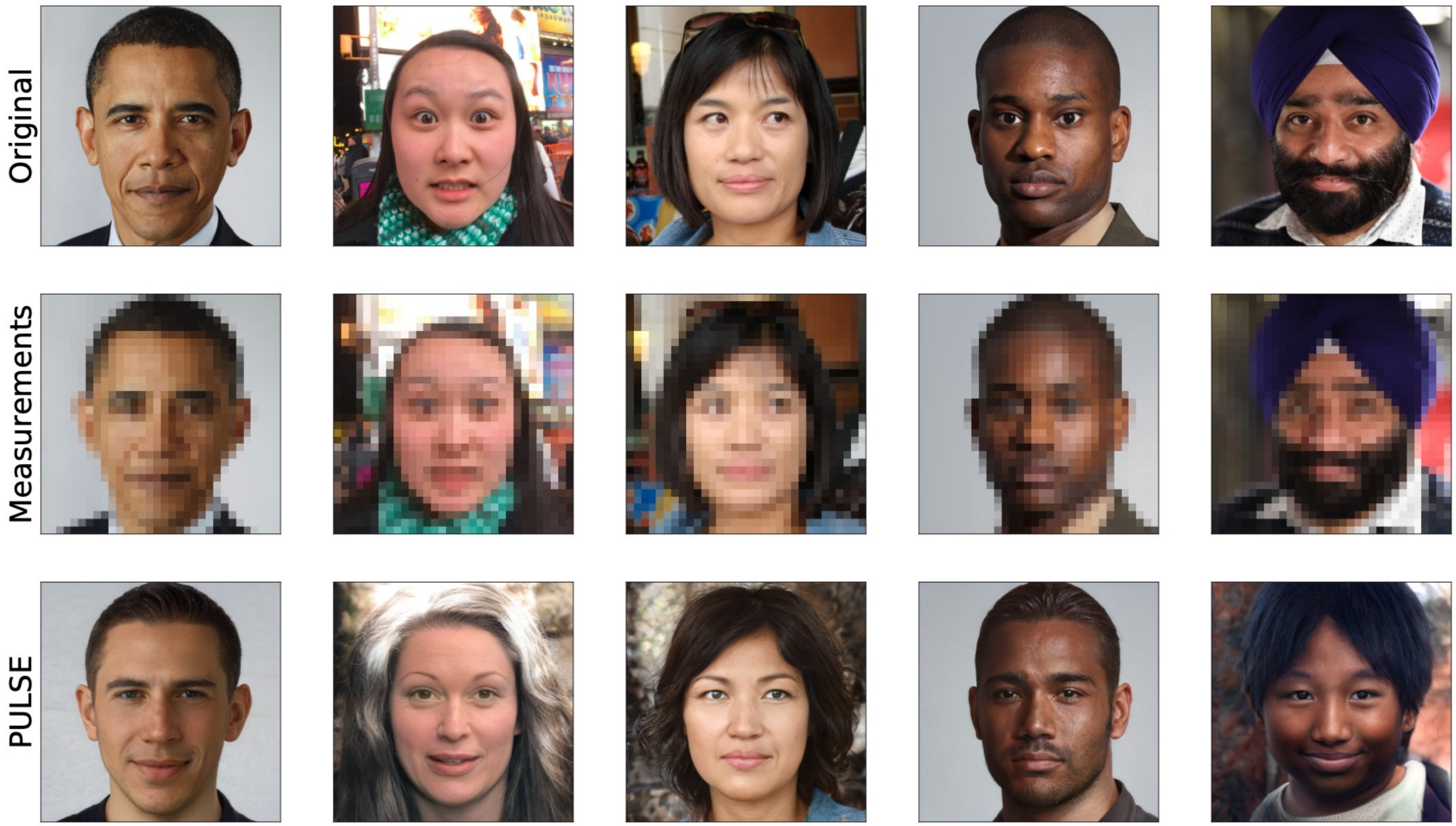


Source: www.vox.com/2015/9/18/9348821/photography-race-bias

Source: www.vox.com/2015/9/18/9348821/photography-race-bias

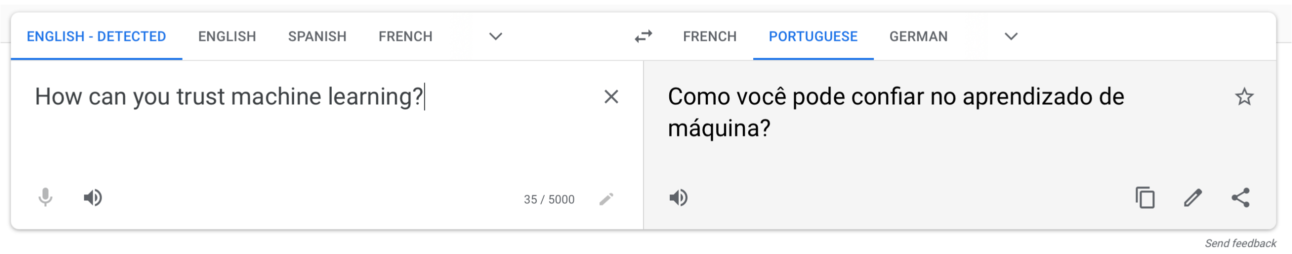
Source: www.vox.com/2015/9/18/9348821/photography-race-bias

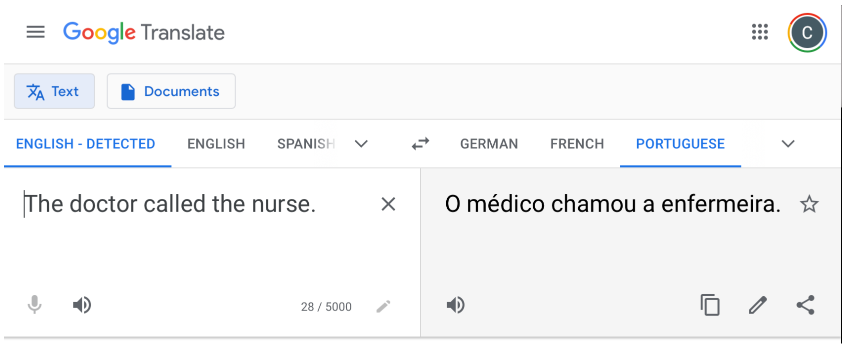
These biases show up in ML…

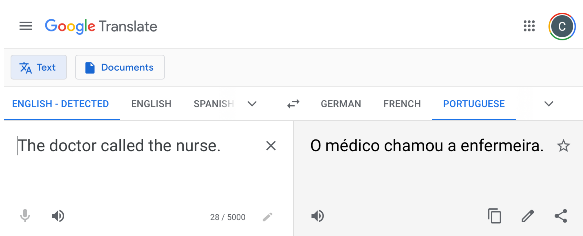


And, it's not just about diversity or coverage in the data we collect…

- Must ensure all development decisions reflect values we want the model to exhibit

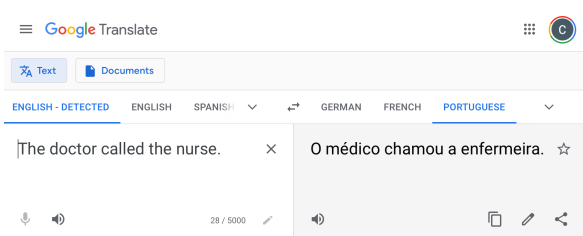

ENGLISH - DETECTED
ENGLISH
SPANISH
FRENCH
FRENCH
PORTUGUESE
GERMAN
How can you trust machine learning?
Como você pode confiar no aprendizado de máquina?
35 / 5000
Send feedback


Google Translate
Text
Documents
ENGLISH - DETECTED
ENGLISH
SPANISH
GERMAN
FRENCH
PORTUGUESE
The doctor called the nurse.
28 / 5000
O médico chamou a enfermeira.
C



If >50% of doctors are male in the dataset,
all instances of “doctor” translated to male form

Even with infinite and representative data, this issue will not be resolved



If >50% of doctors are male in the dataset, all instances of “doctor” translated to male form

Even with infinite and representative data,
this issue will not be resolved

AI Ethics is about considering the consequences of every decision we make in the ML system