

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 to 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.

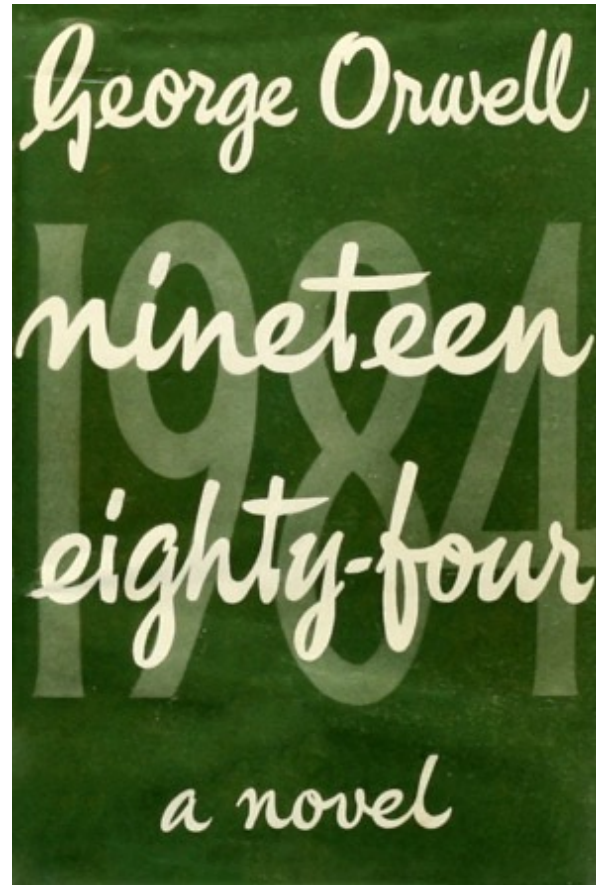
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Ethical Concerns of Artificial Intelligence

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

Privacy and Surveillance



BRIAN BARRETT

LILY HAY NEWMAN

SECURITY SEP 3, 2021 12:58 PM

Apple Backs Down on Its Controversial Photo-Scanning Plans

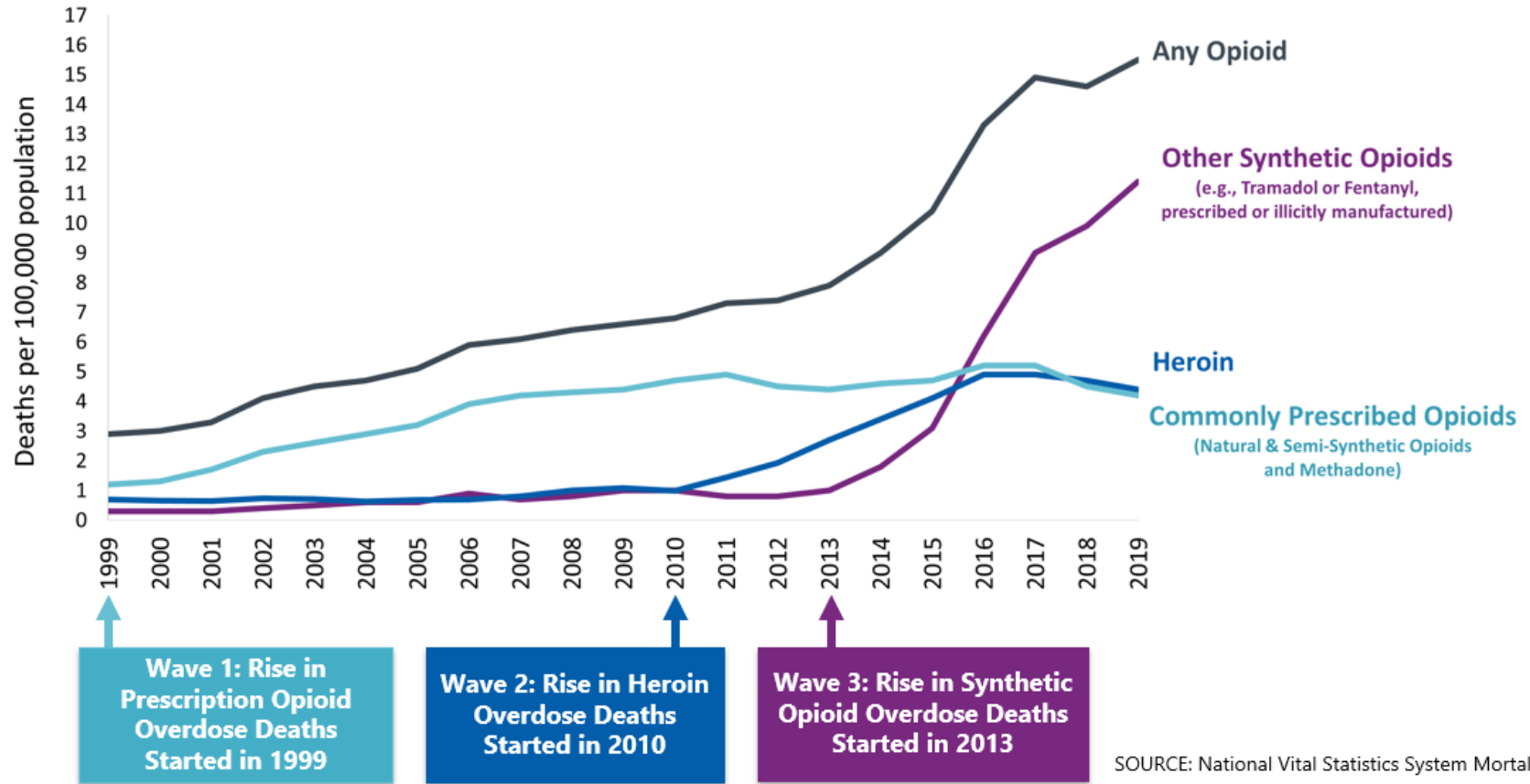
A sustained backlash against a new system to look for child sexual abuse materials on user devices has led the company to hit pause.

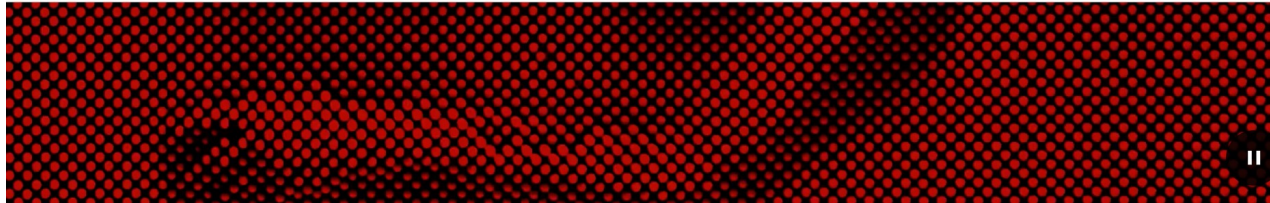


Privacy advocates and security researchers are cautiously optimistic about the pause. PHOTOGRAPH: JUSTIN SULLIVAN/GETTY IMAGES

Opacity of Predictions

Three Waves of the Rise in Opioid Overdose Deaths





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

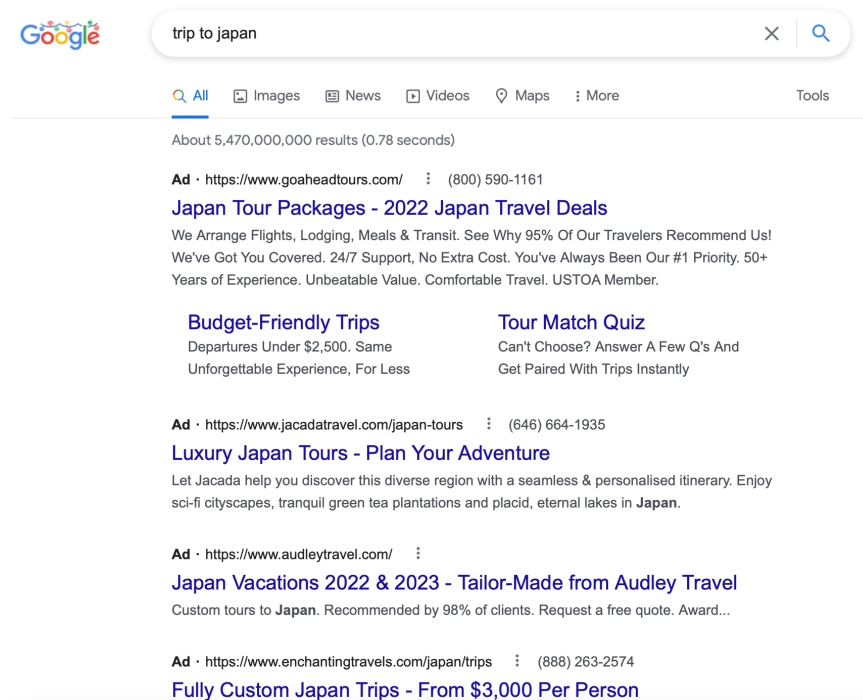
SECTOR: HEALTH CARE, PUBLIC SAFETY

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...



The screenshot shows a Google search for "trip to japan". The search bar is at the top with the Google logo on the left and a search icon on the right. Below the search bar are navigation links for "All", "Images", "News", "Videos", "Maps", and "More". The search results indicate "About 5,470,000,000 results (0.78 seconds)".

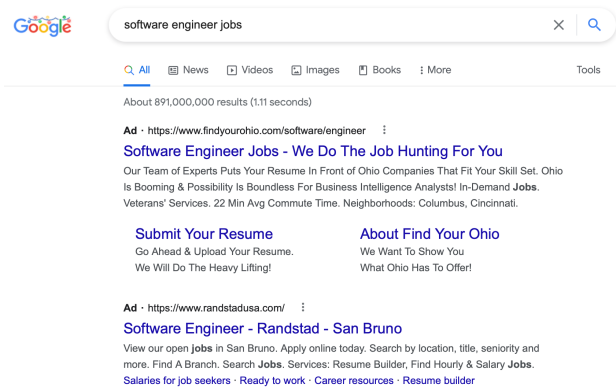
The first advertisement is from <https://www.goaheadtours.com/> with phone number (800) 590-1161. The headline is "Japan Tour Packages - 2022 Japan Travel Deals". The description reads: "We Arrange Flights, Lodging, Meals & Transit. See Why 95% Of Our Travelers Recommend Us! We've Got You Covered. 24/7 Support, No Extra Cost. You've Always Been Our #1 Priority. 50+ Years of Experience. Unbeatable Value. Comfortable Travel. USTOA Member." Below the description are two sub-promotions: "Budget-Friendly Trips" (Departures Under \$2,500. Same Unforgettable Experience, For Less) and "Tour Match Quiz" (Can't Choose? Answer A Few Q's And Get Paired With Trips Instantly).

The second advertisement is from <https://www.jacadatravel.com/japan-tours> with phone number (646) 664-1935. The headline is "Luxury Japan Tours - Plan Your Adventure". The description reads: "Let Jacada help you discover this diverse region with a seamless & personalised itinerary. Enjoy sci-fi cityscapes, tranquil green tea plantations and placid, eternal lakes in Japan."

The third advertisement is from <https://www.audleytravel.com/>. The headline is "Japan Vacations 2022 & 2023 - Tailor-Made from Audley Travel". The description reads: "Custom tours to Japan. Recommended by 98% of clients. Request a free quote. Award..."

The fourth advertisement is from <https://www.enchantingtravels.com/japan/trips> with phone number (888) 263-2574. The headline is "Fully Custom Japan Trips - From \$3,000 Per Person".

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

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

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



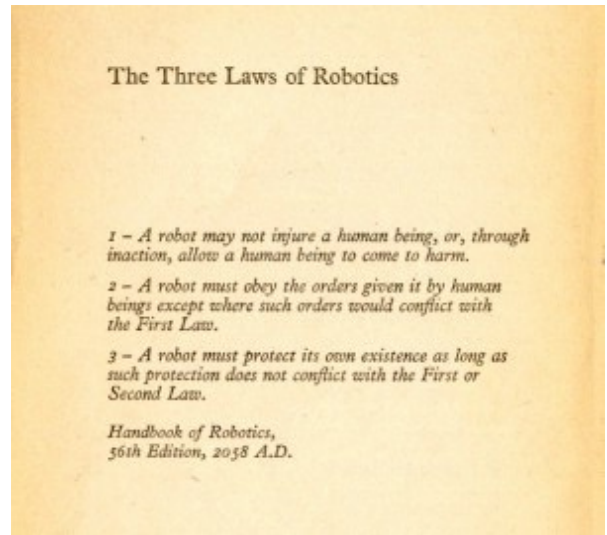
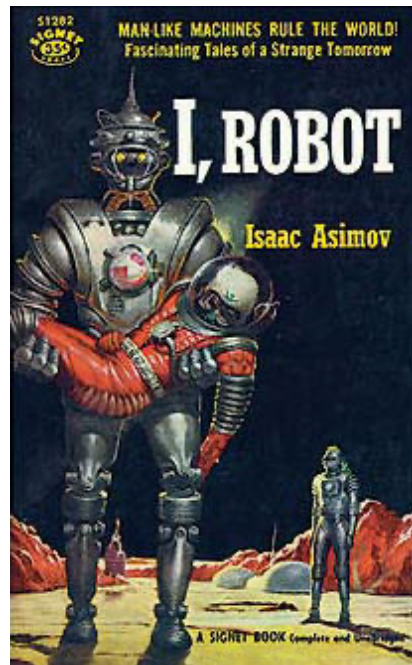
©2022 Carlos Guestrin






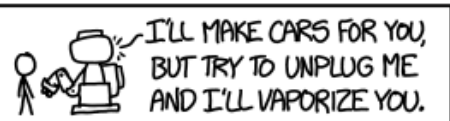

©2022 Carlos Guestrin

[https://www.youtube.com/watch
?v=4sEVX4mPuto](https://www.youtube.com/watch?v=4sEVX4mPuto)

Decisions by Proxy



WHY ASIMOV PUT THE THREE LAWS OF ROBOTICS IN THE ORDER HE DID:

POSSIBLE ORDERING	CONSEQUENCES	
1. (1) DON'T HARM HUMANS 2. (2) OBEY ORDERS 3. (3) PROTECT YOURSELF	[SEE ASIMOV'S STORIES]	BALANCED WORLD
1. (1) DON'T HARM HUMANS 2. (3) PROTECT YOURSELF 3. (2) OBEY ORDERS		FRUSTRATING WORLD
1. (2) OBEY ORDERS 2. (1) DON'T HARM HUMANS 3. (3) PROTECT YOURSELF		KILLBOT HELLSCAPE
1. (2) OBEY ORDERS 2. (3) PROTECT YOURSELF 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE
1. (3) PROTECT YOURSELF 2. (1) DON'T HARM HUMANS 3. (2) OBEY ORDERS		TERRIFYING STANDOFF
1. (3) PROTECT YOURSELF 2. (2) OBEY ORDERS 3. (1) DON'T HARM HUMANS		KILLBOT HELLSCAPE

<https://xkcd.com/1613/>

©2022 Carlos Guestrin

CS229: Machine Learning



©2022 Carlos Guestrin

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

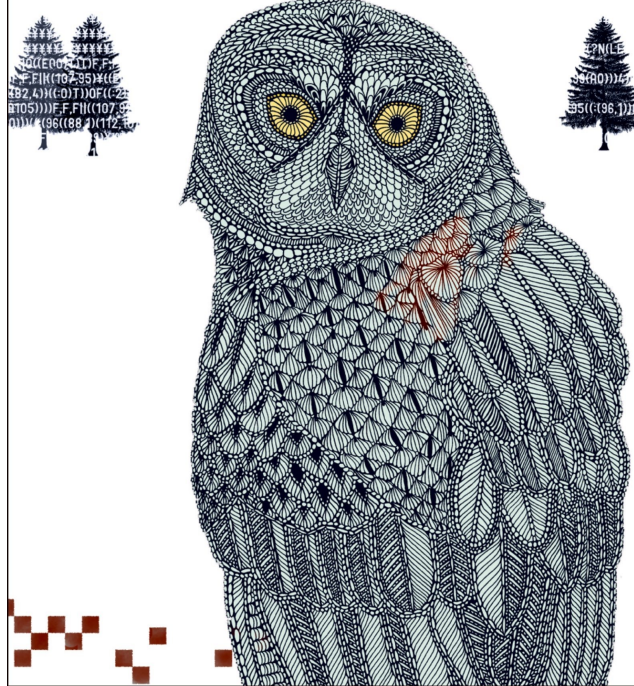
Existential Risk



NICK BOSTROM

SUPERINTELLIGENCE

Paths, Dangers, Strategies



©2022 Carlos Guestrin

CS229: Machine Learning



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

- 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
 - 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

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

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
- 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)
 - y – (hidden) true target in decision (will this person pay the loan)
-
- Shorthand probability notation:
-
- “Perfect” predictor:

Fairness through Unawareness

- Definition:
- Desirable properties:
- Criticisms:

Three Important Fairness Criteria

- Independence
- Separation
- Sufficiency

All these criteria are achievable...

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

1. Independence

- Definition: Decision d independent of sensitive features c

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

Independence: Desirable Properties

- Simple
- Some legal support
- In some settings, can increase representation, e.g., in admissions

Independence: Shortcomings

- Ignores possible correlations between y and \mathbf{c}
 - Precludes perfect predictor $d=y$
- Laziness: quality of decision doesn't need to be uniformly good between groups

2. Separation

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

Variant of Separation: False negative rate parity

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

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

- Separation:
- Confusion matrix:
- Variants:

Separation: Desirable Properties

- Optimality compatibility
- Incentivize to reduce errors equally across groups

Separation: Shortcomings

- Can amplify disparities

3. Sufficiency

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

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

 - Negative predictive rate:

Sufficiency: Desirable Properties

- Optimality compatibility:
- Equal chance of success, given acceptance:

Sufficiency: Shortcomings

- Also can amplify disparities

All these criteria are achievable...

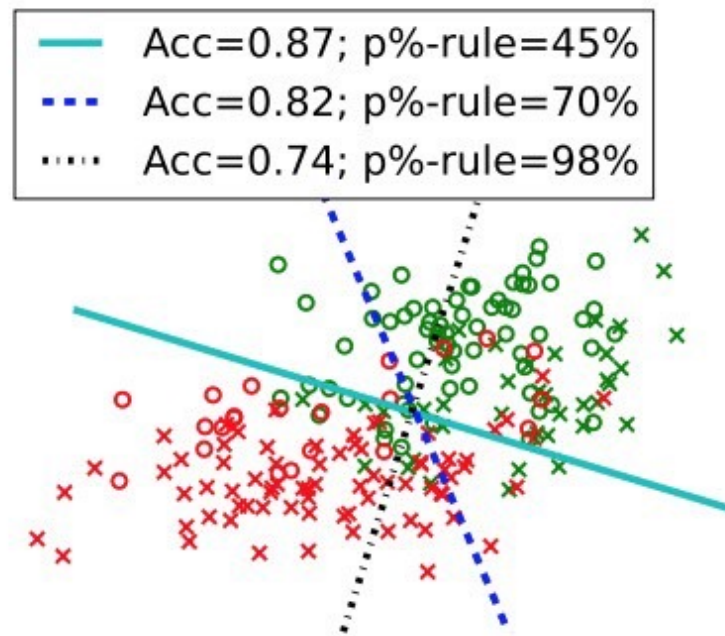
- Techniques include:
 - Pre-processing
 - Changing training procedure
 - Post-processing



Trade-offs are Inevitable

Tradeoff Between Fairness and Accuracy

Tradeoff Between Group-Specific Performance and Average-Case Performance



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

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?







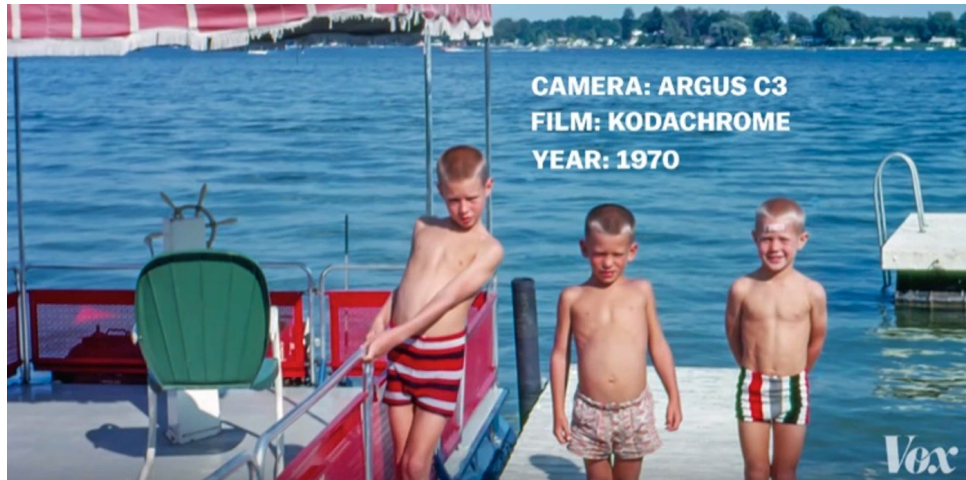
ML perpetuates stereotypes...

The image shows a Google Images search interface for the query "ceo portrait". The search bar is at the top, with the Google logo on the left and navigation icons on the right. Below the search bar, there are tabs for "All", "Images", "News", "Shopping", "Videos", and "More". A "Tools" button is also present. A horizontal filter bar contains several categories: "business", "office", "casual", "headshot", "modern", "woman", "photography", "man", "professional", "environmental", "executive", "black", and "background". The main content area displays a grid of image thumbnails, each with a small caption and source URL. The thumbnails show various styles of CEO portraits, including men in suits, women in business attire, and different backgrounds like office settings and cityscapes. A "Related searches" box is located in the bottom right corner, listing terms like "business ceo portrait", "casual ceo portrait", and "ceo photoshoot ideas".

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

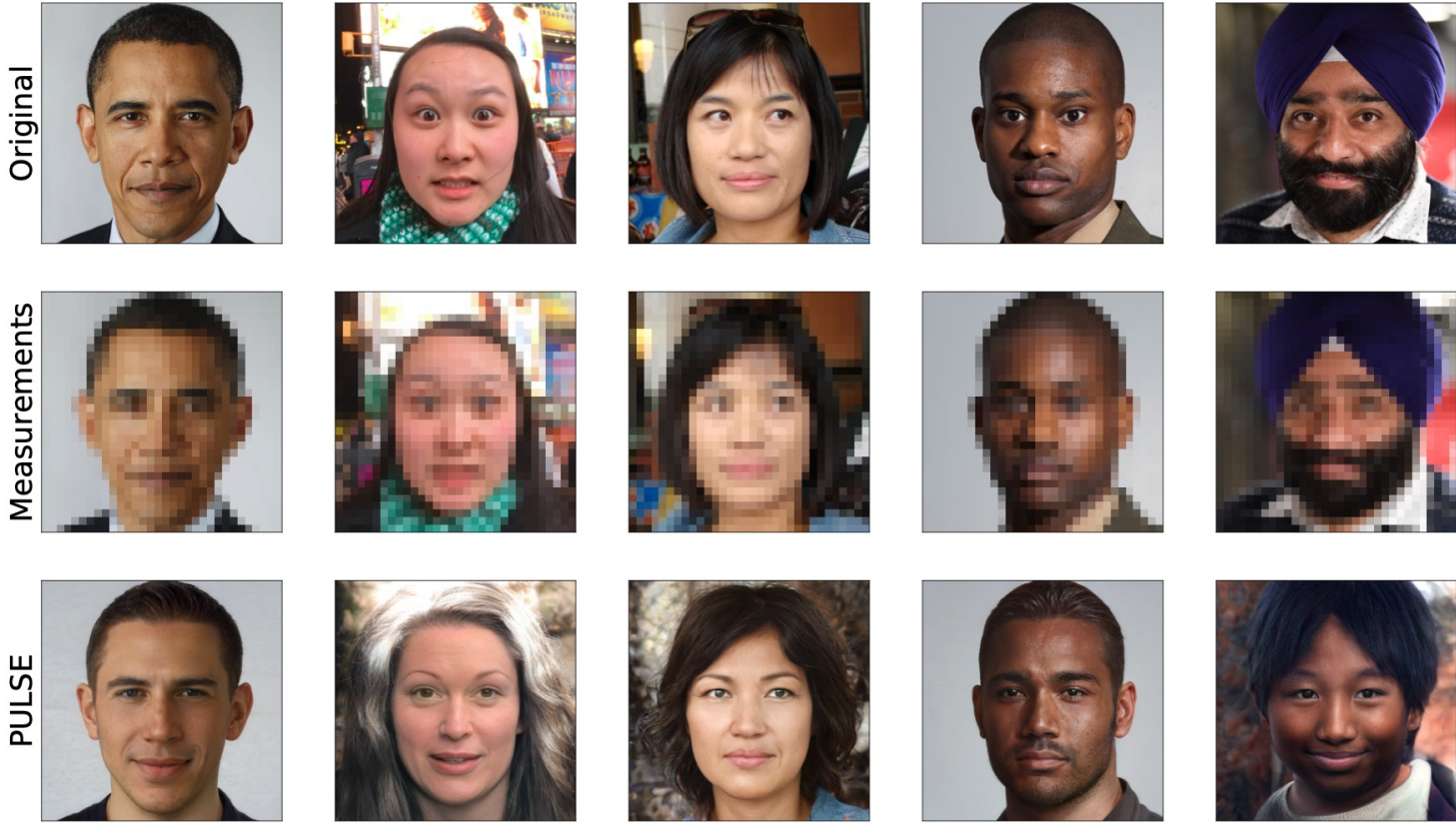


Source: www.vox.com/2015/9/18/9348821/photography-race-bias



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

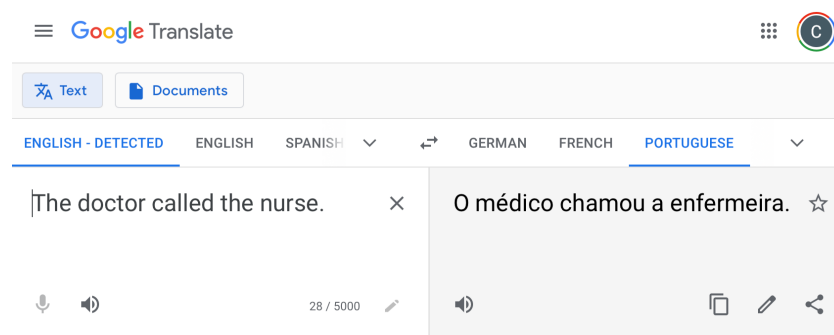
ENGLISH - DETECTED ENGLISH SPANISH FRENCH ▾ ↔ FRENCH PORTUGUESE GERMAN ▾

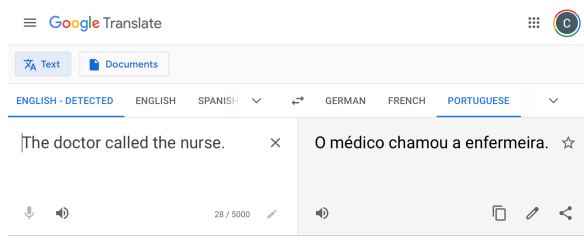
How can you trust machine learning? ×

Como você pode confiar no aprendizado de máquina? ☆

🎤 🔊 35 / 5000 ✎ 🔊 📄 ✎ 🔄

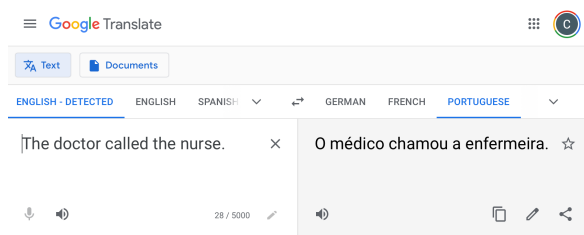
[Send feedback](#)





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