



Machine Learning in Production

Responsible ML

≡ Engineering



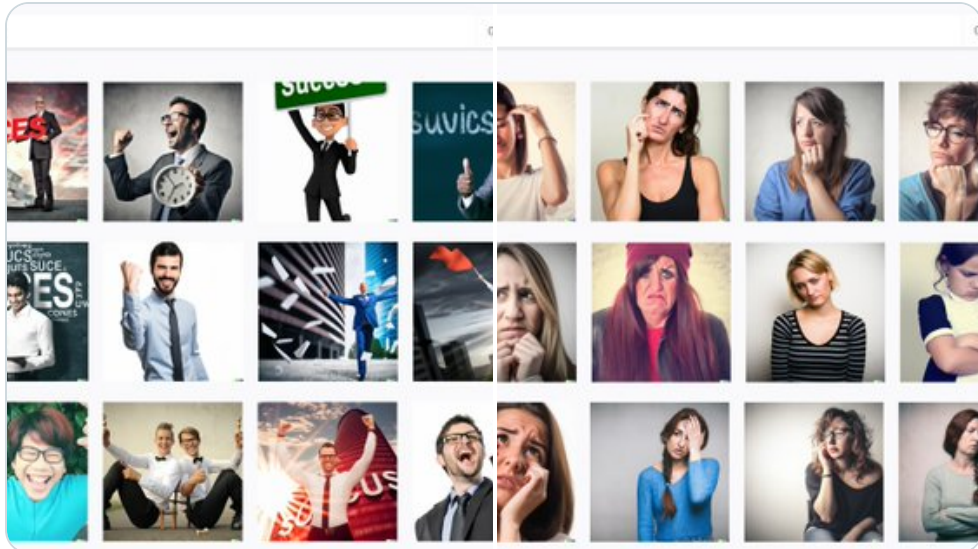
Nao Tokui

@naotokui_en · [Follow](#)



"Success" and "Sadness", according to DALL-E 2.

(No cherry-picking)



4:00 AM · Aug 7, 2022



466



Reply



Copy link

[Read 21 replies](#)

Changing directions...

Fundamentals of Engineering AI-Enabled Systems

Holistic system view: AI and non-AI components, pipelines, stakeholders, environment interactions, feedback loops

Requirements:

- System and model goals
- User requirements
- Environment assumptions
- Quality beyond accuracy
- Measurement
- Risk analysis
- Planning for mistakes

Architecture + design:

- Modeling tradeoffs
- Deployment architecture
- Data science pipelines
- Telemetry, monitoring
- Anticipating evolution
- Big data processing
- Human-AI design

Quality assurance:

- Model testing
- Data quality
- QA automation
- Testing in production
- Infrastructure quality
- Debugging

Operations:

- Continuous deployment
- Contin. experimentation
- Configuration mgmt.
- Monitoring
- Versioning
- Big data
- DevOps, MLOps

Teams and process: Data science vs software eng. workflows, interdisciplinary teams, collaboration points, technical debt

Responsible AI Engineering

Provenance,
versioning,
reproducibility

Safety

Security and
privacy

Fairness

Interpretability
and explainability

Transparency
and trust

Ethics, governance, regulation, compliance, organizational culture

Readings

R. Caplan, J. Donovan, L. Hanson, J. Matthews. "Algorithmic Accountability: A Primer", Data & Society (2018).

Learning Goals

- Review the importance of ethical considerations in designing AI-enabled systems
- Recall basic strategies to reason about ethical challenges
- Diagnose potential ethical issues in a given system
- Understand the types of harm that can be caused by ML
- Understand the sources of bias in ML

Why Fairness?

Many interrelated issues:

* Ethics * Fairness * Justice * Discrimination * Safety * Privacy * Security * Transparency *
Accountability

Each is a deep and nuanced research topic. We focus on survey of some key issues.



In 2015, Shkreli received widespread criticism [...] obtained the manufacturing license for the antiparasitic drug Daraprim and raised its price from USD 13.5 to 750 per pill [...] referred to by the media as "the most hated man in America" and "Pharma Bro". -- [Wikipedia](#)

"I could have raised it higher and made more profits for our shareholders. Which is my primary duty." -- Martin Shkreli

Speaker notes

Image source: https://en.wikipedia.org/wiki/Martin_Shkreli#/media/File:Martin_Shkreli_2016.jpg



Terminology



Legal = in accordance to societal laws

- systematic body of rules governing society; set through government
- punishment for violation

Ethical = following moral principles of tradition, group, or individual

- branch of philosophy, science of a standard human conduct
- professional ethics = rules codified by professional organization
- no legal binding, no enforcement beyond "shame"
- high ethical standards may yield long term benefits through image and staff loyalty

Big Disclaimer

Legality is obviously a locale-specific concern.

What is *ethical* (and how we know) is a very complicated question.

- Whether there exists ground-truth ethics is a point of philosophical debate.
- Often informed by context/culture/etc.

We adopt a generally US-centric perspective for much of this discussion.

- ...Because that's where we are.
- But given the global reach of software, tread with care.

Speaker notes

GDPR is an easy example



With a few lines of code...

Developers have substantial power in shaping products, and software has substantial power over human lives.

Small design decisions can have substantial impact (safety, security, discrimination, ...) -- not always deliberate

Our view: We have both **legal & ethical** responsibilities to anticipate mistakes, think through their consequences, and build in mitigations!

Example: Social Media

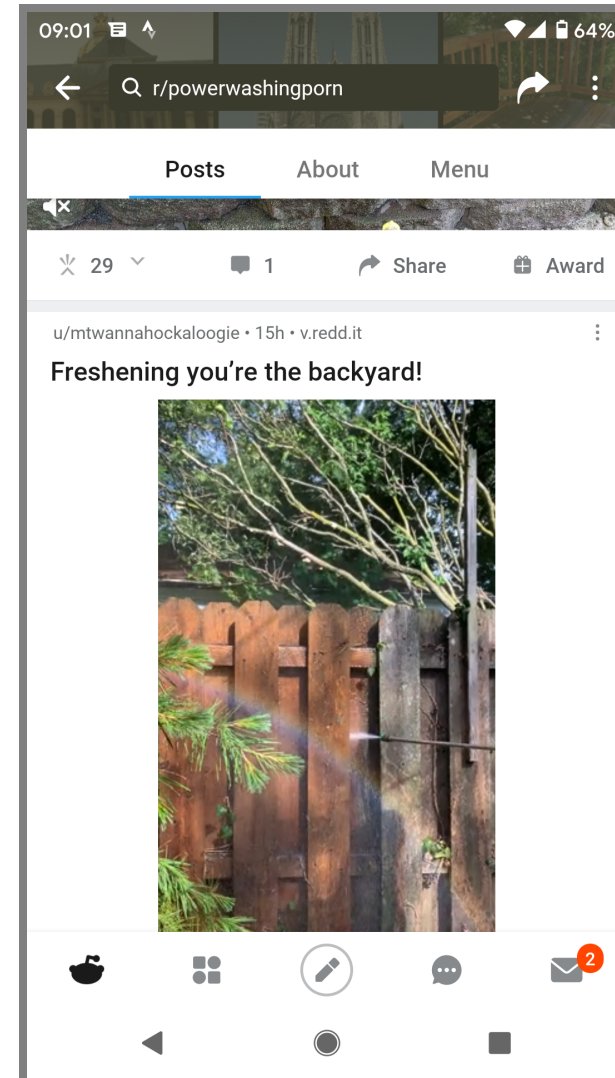


≡ *What is the (real) organizational objective of the company?*

Optimizing for Organizational Objective

How do we maximize the user engagement? Examples:

- Infinite scroll: Encourage non-stop, continual use
- Personal recommendations: Suggest news feed to increase engagement
- Push notifications: Notify disengaged users to return to the app



Addiction

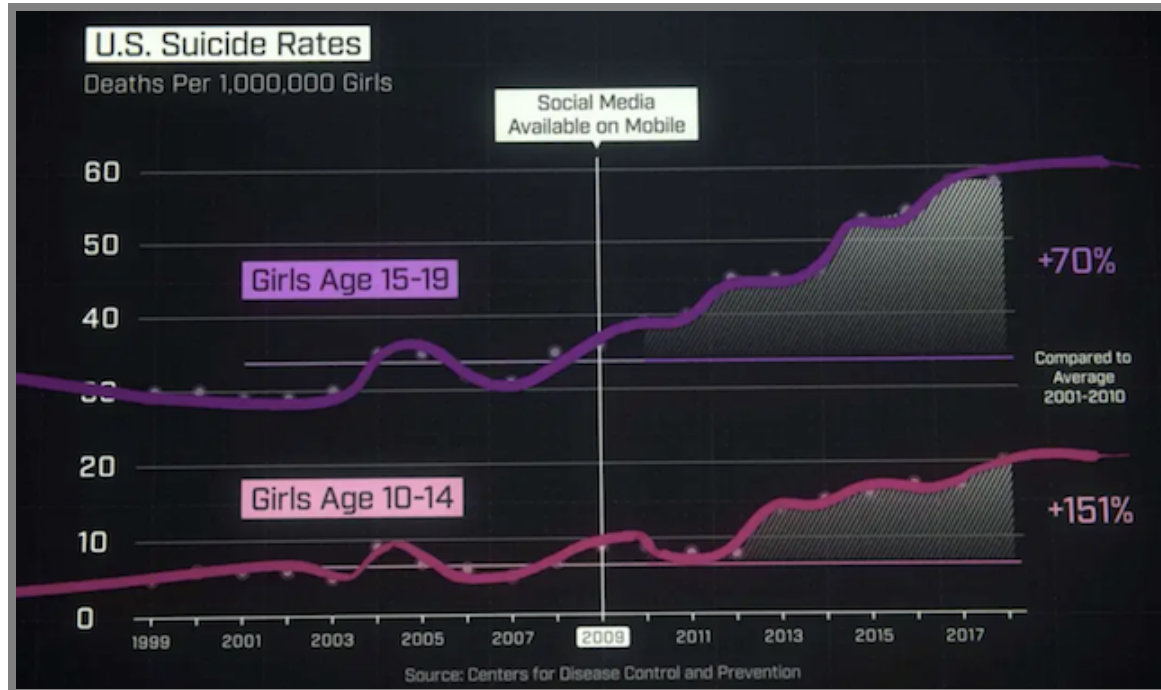


- 210M people worldwide addicted to social media
- 71% of Americans sleep next to a mobile device
- ~1000 people injured **per day** due to distracted driving (USA)

[https://www.flurry.com/blog/mobile-addicts-multiply-across-the-globe/;](https://www.flurry.com/blog/mobile-addicts-multiply-across-the-globe/)

https://www.cdc.gov/motorvehiclesafety/Distracted_Driving/index.html

Mental Health



- 35% of US teenagers with low social-emotional well-being have been bullied on social media.
- 70% of teens feel excluded when using social media.

≡ <https://leftronic.com/social-media-addiction-statistics>

Disinformation & Polarization

**BERNIE SANDERS:
CLINTON FOUNDATION
IS A "PROBLEM"**

HILLARY ASKS "WHAT DIFFERENCE DOES IT MAKE
IF YOU KNOW THE DIFFERENCE"

**FOLLOW VETERANS US
IF YOU KNOW THE DIFFERENCE**

**ANOTHER GRUESOME ATTACK ON POLICE
BY A BLM MOVEMENT ACTIVIST**

Support Hillary
the American Muslims

**ATAN: IF I WIN CLINTON WINS!
JESUS: NOT IF I CAN HELP IT!**

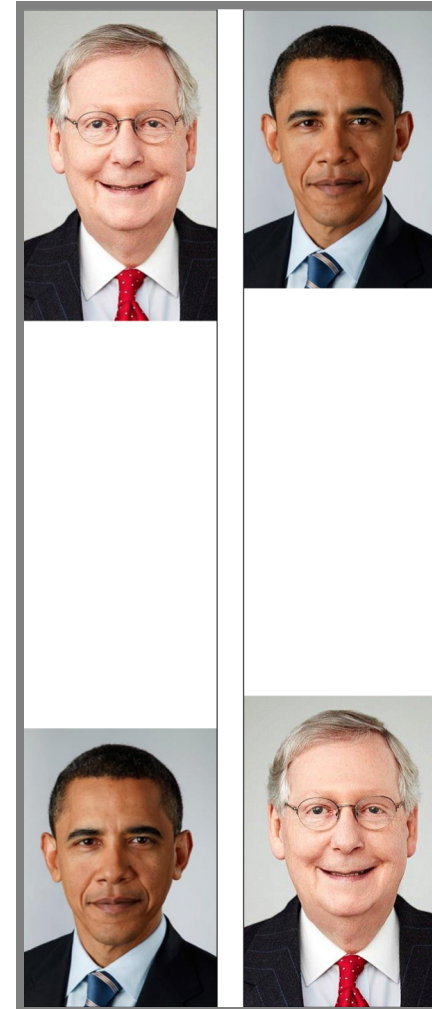
**OUR HEARTS ARE
WITH THOSE 11 HEROES**

**HILLARY CLINTON HAS A 69 PERCENT
DISAPPROVAL RATE AMONG
ALL VETERANS**

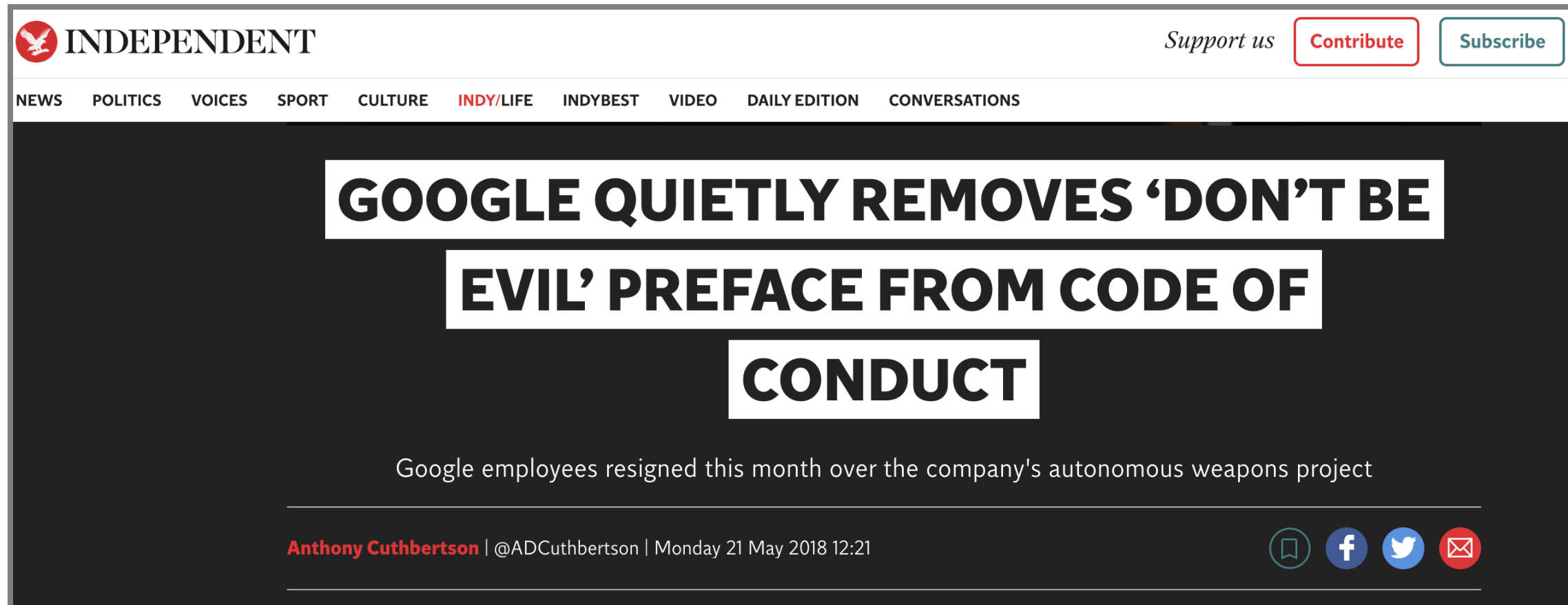
**NO
INVADERS
ALLOWED**

PRESS 'LIKE' TO HELP JESUS WIN!

Discrimination (as a side effect...)



Who's to blame?



The image shows a screenshot of a news article from the Independent. The headline is displayed in large, bold, black text on a white background, split across three lines: "GOOGLE QUIETLY REMOVES 'DON'T BE EVIL' PREFACE FROM CODE OF CONDUCT". Below the headline, a sub-headline reads "Google employees resigned this month over the company's autonomous weapons project". The author's name, "Anthony Cuthbertson", and the date, "Monday 21 May 2018 12:21", are visible at the bottom left of the article content. Social media sharing icons for bookmark, Facebook, Twitter, and email are located at the bottom right. The top of the page features the Independent logo, navigation links for various sections like NEWS, POLITICS, and VOICES, and buttons for "Support us", "Contribute", and "Subscribe".

INDEPENDENT

Support us [Contribute](#) [Subscribe](#)

NEWS POLITICS VOICES SPORT CULTURE **INDY/LIFE** INDYBEST VIDEO DAILY EDITION CONVERSATIONS

GOOGLE QUIETLY REMOVES 'DON'T BE EVIL' PREFACE FROM CODE OF CONDUCT

Google employees resigned this month over the company's autonomous weapons project

Anthony Cuthbertson | @ADCuthbertson | Monday 21 May 2018 12:21

[Bookmark](#) [Facebook](#) [Twitter](#) [Email](#)

Are these companies intentionally trying to cause harm? If not, what are the root causes of the problem?

Liability?

The software is provided “as is”, without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement. In no event shall the authors or copyright holders be liable for any claim, damages or other liability, whether in an action of contract, tort or otherwise, arising from, out of or in connection with the software or the use or other dealings in the software.

Software companies have usually gotten away with claiming no liability for their products



Some Challenges

Misalignment between organizational goals & societal values

- Financial incentives often dominate other goals ("grow or die")

Hardly any regulation

- Often, little legal consequences for causing negative impact (with exceptions based on domain)
- Poor understanding of socio-technical systems by policy makers

Engineering challenges, at system- & ML-level

- Difficult to clearly define or measure ethical values
- Difficult to anticipate all possible usage contexts
- Difficult to anticipate impact of feedback loops
- Difficult to prevent malicious actors from abusing the system
- Difficult to interpret output of ML and make ethical decisions

These problems have long existed, but are being rapidly exacerbated by the widespread use of ML

There are ML-specific techniques/concerns with respect to these issues.

Responsible Engineering Matters

Engineers have substantial power in shaping products and outcomes

Serious individual and societal harms possible from (a) negligence and (b) malicious designs

- Safety, mental health, weapons
- Security, privacy
- Manipulation, addiction, surveillance, polarization
- Job loss, deskilling
- Discrimination

"I don't care about ethics, I just want to make money."

Regulations apply in many domains, including those where ML is "hot"

- Health care, finance, real estate

Bad PR can be bad for your bottom line.



 **DHH**  
@dhh · [Follow](#)

The [@AppleCard](#) is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

8:34 PM · Nov 7, 2019

 22.7K  Reply  Copy link

Responsible Engineering in this Course

Key areas of concern

- Fairness
- Safety
- Security and privacy
- Transparency and accountability

Technical infrastructure concepts

- Interpretability and explainability
- Versioning, provenance, reproducibility

Fairness

Dividing a Pie?

- Equal slices for everybody
- Bigger slices for active bakers
- Bigger slices for inexperienced/new members (e.g., children)
- Bigger slices for hungry people
- More pie for everybody, bake more

*(Not everybody contributed equally during baking, not everybody is
≡ equally hungry)*



What is fair?

Fairness discourse asks questions about how to treat people and whether treating different groups of people differently is ethical. If two groups of people are systematically treated differently, this is often considered unfair.

Regulated domains (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)




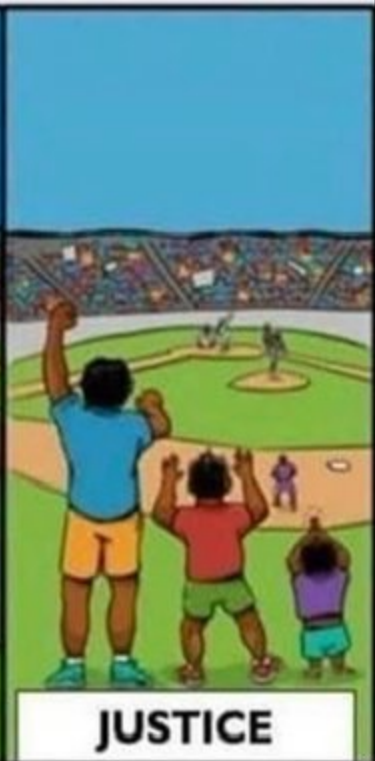

Extends to marketing and advertising; not limited to final decision

Legally protected classes (US)

- Race ([Civil Rights Act of 1964](#))
- Religion ([Civil Rights Act of 1964](#))
- National origin ([Civil Rights Act of 1964](#))
- Sex, sexual orientation, and gender identity ([Equal Pay Act of 1963](#), [Civil Rights Act of 1964](#), and [Bostock v. Clayton](#))
- Age (40 and over, [Age Discrimination in Employment Act of 1967](#))
- Pregnancy ([Pregnancy Discrimination Act of 1978](#))
- Familial status (preference for or against having children, [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 of 1994](#))
- Genetic information ([Genetic Information Nondiscrimination Act of 2008](#))

Common framing: Equality vs Equity vs Justice

Equality vs Equity vs Justice

 <p>REALITY</p>	 <p>EQUALITY</p>	 <p>EQUITY</p>	 <p>JUSTICE</p>	 <p>INCLUSION</p>
<p>One gets more than is needed, while the other gets less than is needed. Thus, a huge disparity is created.</p>	<p>The assumption is that everyone benefits from the same supports. This is considered to be equal treatment.</p>	<p>Everyone gets the support they need, which produces equity.</p>	<p>All 3 can see the game without supports or accommodations because the cause(s) of the inequity was addressed. The systemic barrier has been removed.</p>	<p>Everyone is INCLUDED in the game. No one is left on the outside; we <u>didn't</u> only remove the barriers keeping people out, we made sure they were valued & involved.</p>

@ClinPsychDavid

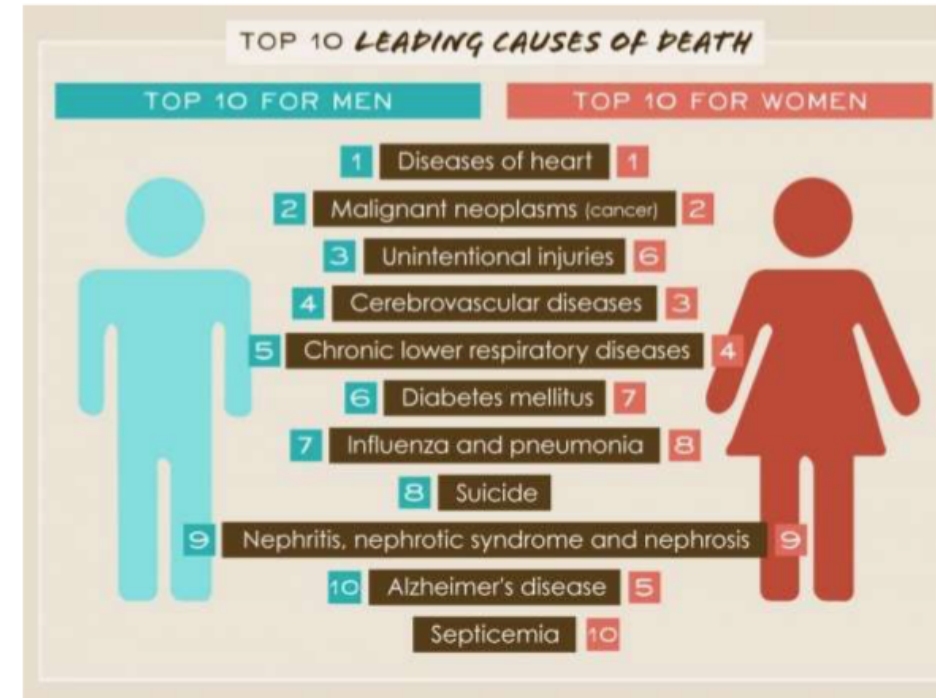
Caveat: Something can be fair but still unethical (Thanos)!

Not all discrimination is harmful



FEDERAL TRADE COMMISSION

Mortgage discrimination is against the law.



Not all discrimination is harmful



- Discrimination is a **domain-specific** concept
 - ML models discriminate based on input data by construction.
 - There are real differences between two groups, it might not be fair to ignore them
 - The problem is *unjustified* differentiation; i.e., discriminating on factors that should not matter

Fairness vs. bias vs. harm

Fairness is best understood as a **societal or cultural concept**.

Bias, in discussing ML, can be understood as a **technological or algorithmic concept**; it is often discussed in terms of its negative effects.

- Whether bias is harmful or unfair is not something that can be decided algorithmically.

Useful definition/framework defines algorithmic bias as "a skew that produces a type of harm."

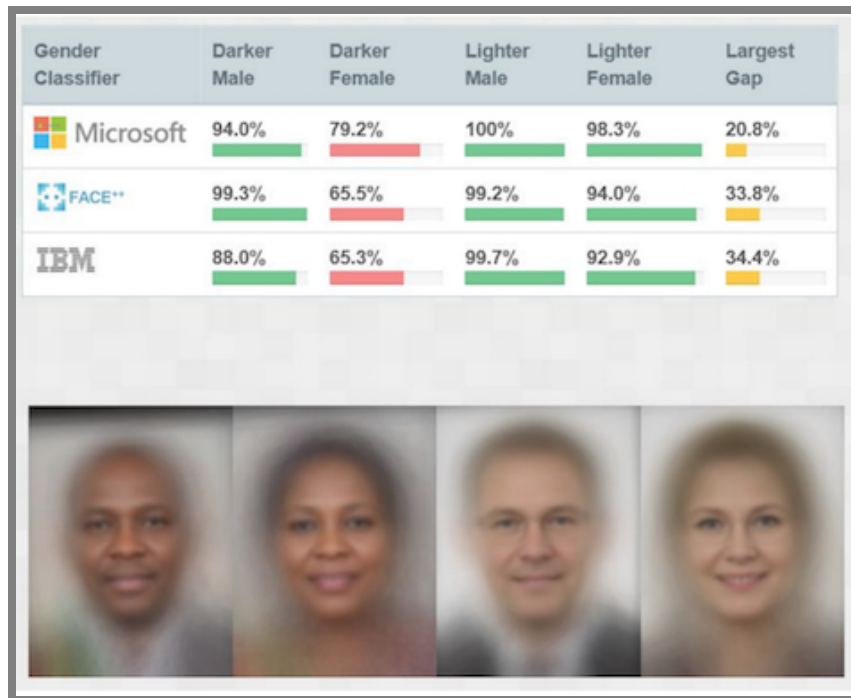
Types of Harm on Society

Harms of allocation: Withhold opportunities or resources

Harms of representation: Reinforce stereotypes, subordination along the lines of identity

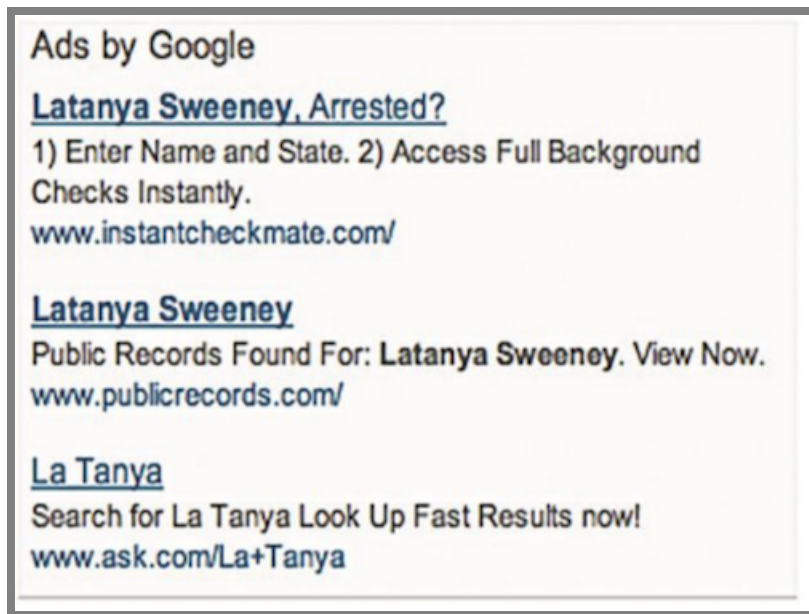
Harms of Allocation

- Withhold opportunities or resources
- Poor quality of service, degraded user experience for certain groups



Harms of Representation

- Over/under-representation of certain groups in organizations
- Reinforcement of stereotypes (e.g. Black community & criminality)



"Racially identifying names" change the ads you get -- names commonly associated with Black individuals were more likely to trigger ads that suggested a criminal background check.

Identifying (co-occurring) harms

	Allocation of resources	Quality of Service	Stereotyping	Denigration	Over- / Under-Representation
Hiring system does not rank women as highly as men for technical jobs	x	x	x		x
Photo management program labels image of black people as “gorillas”		x		x	
Image searches for “CEO” yield only photos of white men on first page			x		x

- Multiple types of harms can be caused by a product!
- Think about your system objectives & identify potential harms.

Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial (2019). *

Role of Requirements Engineering

- Identify system goals
- Identify legal constraints
- Identify stakeholders and fairness concerns
- Analyze risks with regard to discrimination and fairness
- Analyze possible feedback loops (world vs machine)
- Negotiate tradeoffs with stakeholders
- Set requirements/constraints for data and model
- Plan mitigations in the system (beyond the model)
- Design incident response plan
- Set expectations for offline and online assurance and monitoring

Sources of Bias

Where does the bias come from?

The image displays two screenshots of the Google Translate interface, illustrating a gender bias in the Turkish translation of English sentences. In the top screenshot, the source text is "He is a nurse" and "She is a doctor". The detected language is English. The target language is set to Turkish. The translation provided is "O bir hemşire" (She is a nurse) and "O bir doktor" (He is a doctor). In the bottom screenshot, the source text is "O bir hemşire" and "O bir doktor". The detected language is Turkish. The target language is set to English. The translation provided is "She is a nurse" and "He is a doctor". Both screenshots include a "Suggest an edit" button and a "Turn off instant translation" link.

≡ *Semantics derived automatically from language corpora contain human-like biases, Caliskan et al., Science (2017).*

Where does the bias come from?

The screenshot displays the Microsoft Translator interface with two translation panels. The top panel shows the English text "He is a nurse. She is a doctor." being translated into Turkish as "O bir hemşire. O bir doktor." (O is a nurse. O is a doctor). The bottom panel shows the Turkish text "O bir hemşire. O bir doktor." being translated back into English as "She's a nurse. He's a doctor." This illustrates a gender bias where the translator consistently assigns female gender to the subject of the sentence.

Microsoft
Translator Text Conversation Apps For business Help

Search the web Sign in

English Turkish Turkish English

He is a nurse.
She is a doctor.

O bir hemşire.
O bir doktor.

Turkish English Turkish English

O bir hemşire.
O bir doktor.

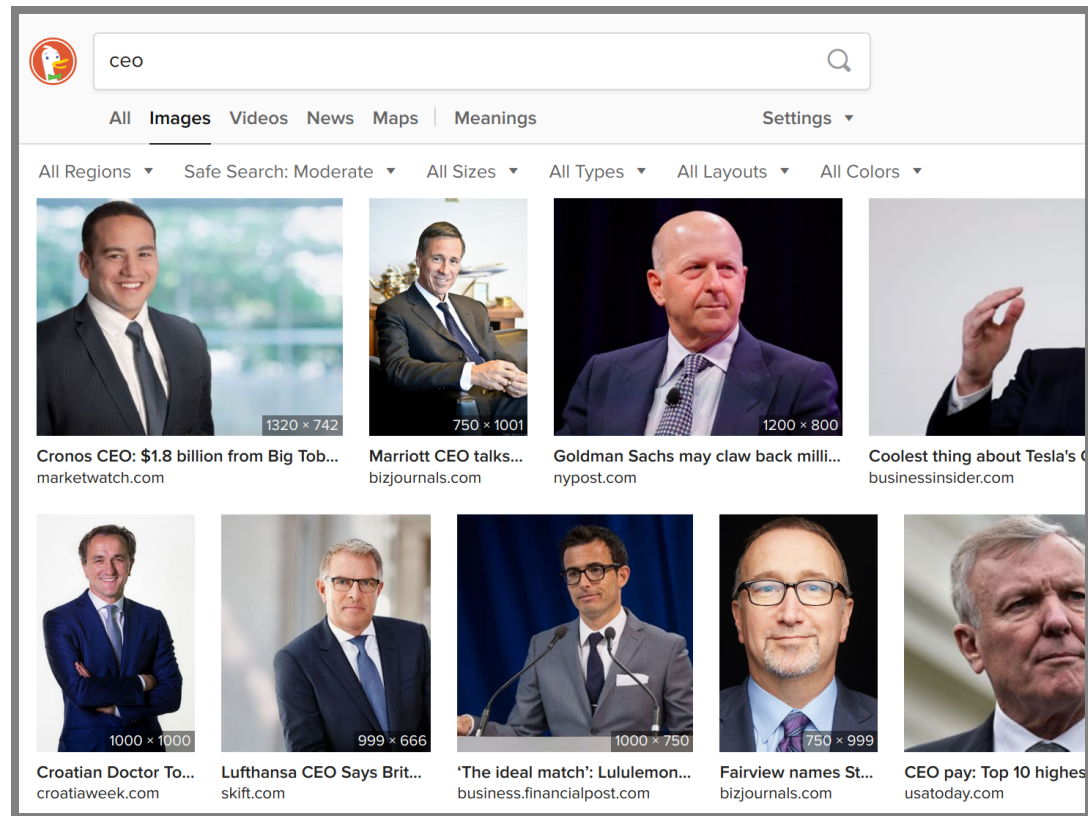
She's a nurse.
He's a doctor.

Sources of Bias

- Historial bias
- Tainted examples
- Limited features
- Skewed sample
- Sample size disparity
- Proxies

Historical Bias

Data reflects past biases, not intended outcomes



Should the algorithm reflect reality?

Speaker notes

"An example of this type of bias can be found in a 2018 image search result where searching for women CEOs ultimately resulted in fewer female CEO images due to the fact that only 5% of Fortune 500 CEOs were woman—which would cause the search results to be biased towards male CEOs. These search results were of course reflecting the reality, but whether or not the search algorithms should reflect this reality is an issue worth considering."



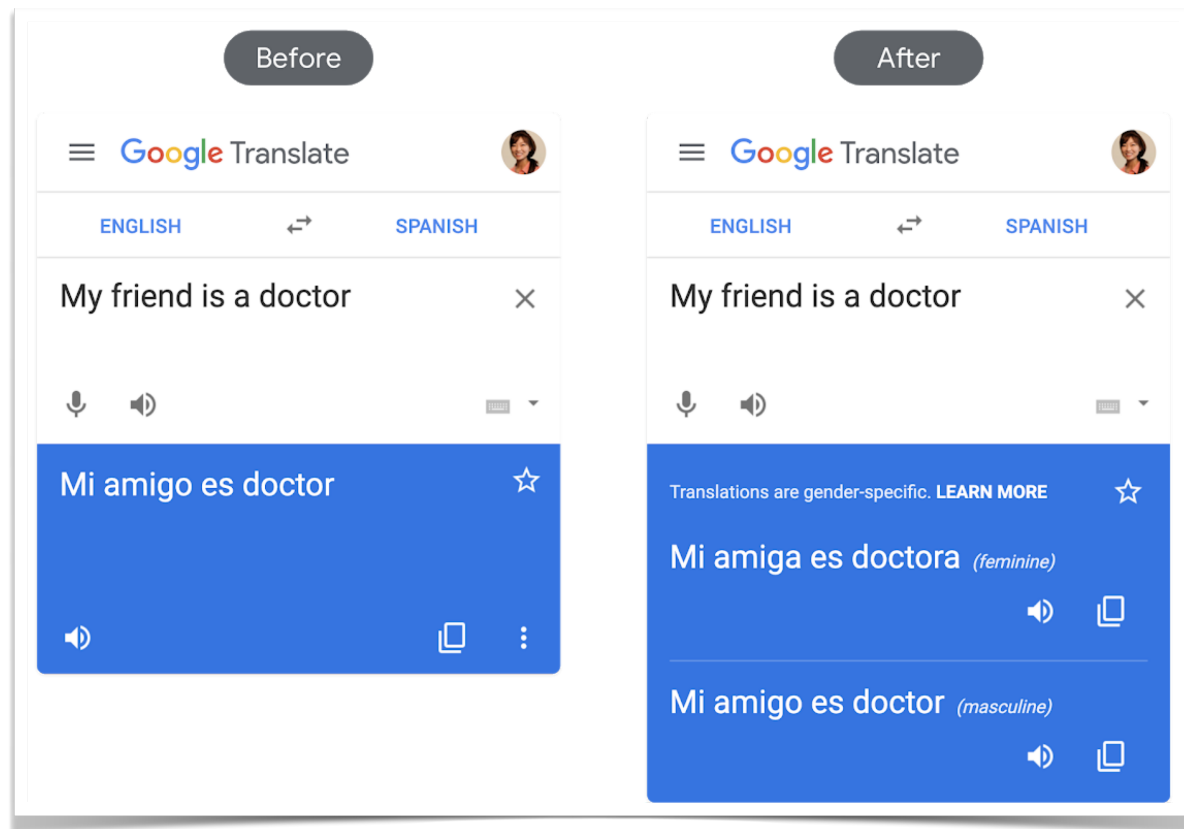
What is reality?

at least as many women play games casually as men if not more. But someone searching for "gamer" probably has an image in their head. Should we be accurate with respect to reality? Or accurate with respect to what the person is searching for?



Correcting Historical Bias

Fix the system, not just the model



≡ A Scalable Approach to Reducing Gender Bias in Google Translate

Correcting Historical Bias

"Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide. " -- Cathy O'Neil in [Weapons of Math Destruction](#)

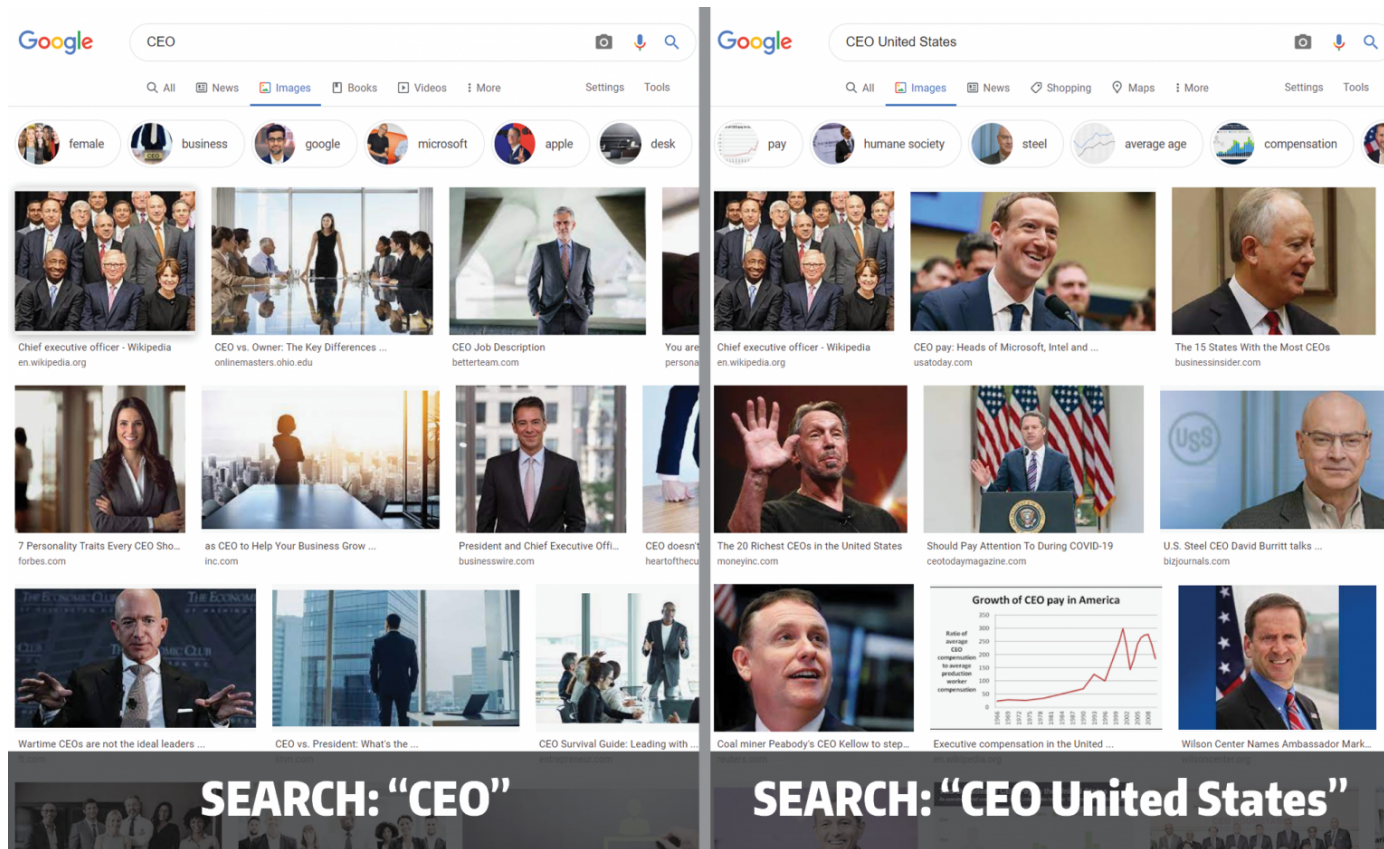
"Through user studies, the [image search] team learned that many users were uncomfortable with the idea of the company "manipulating" search results, viewing this behavior as unethical." -- observation from interviews by Ken Holstein

Correcting Historical Bias

"Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide. " -- Cathy O'Neil in [Weapons of Math Destruction](#)

"Through user studies, the [image search] team learned that many users were uncomfortable with the idea of the company "manipulating" search results, viewing this behavior as unethical." -- observation from interviews by Ken Holstein

Correcting Historical Bias is Hard, even if you want to



Google's 'CEO' image search gender bias hasn't really been fixed

Tainted Labels

Bias in dataset labels assigned (directly or indirectly) by humans

TECH / AMAZON / ARTIFICIAL INTELLIGENCE

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By [James Vincent](#) | Oct 10, 2018, 7:09am EDT

Example: Hiring decision dataset -- labels assigned by (possibly biased) experts or derived from past (possibly biased) hiring decisions

Limited Features

Features that are less informative/reliable for certain subpopulations

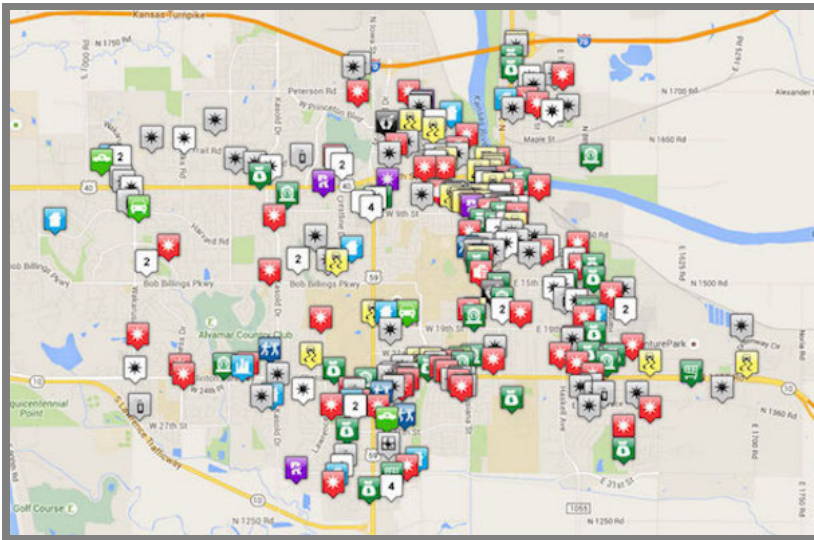


- Graduate admissions: Letters of recommendation equally reliable for international applicants?
- Employee performance review: "Leave of absence" acceptable feature if parental leave is gender skewed?

Decisions may be based on features that are predictive and accurate for a large part of the target distribution, but not so for some other parts of the distribution. For example, a system ranking applications for graduate school admissions may heavily rely on letters of recommendation and be well calibrated for applicants who can request letters from mentors familiar with the culture and jargon of such letters in the US, but may work poorly for international applicants from countries where such letters are not common or where such letters express support with different jargon. To reduce bias, we should be carefully reviewing all features and analyze whether they may be less predictive for certain subpopulations.

Skewed Sample

Bias in how and what data is collected



Crime prediction: Where to analyze crime? What is considered crime?
Actually a random/representative sample?

≡ Raw data is an oxymoron

Sample Size Disparity

Limited training data for some subpopulations



- Biased sampling process: "Shirley Card" used for Kodak color calibration, using mostly Caucasian models
- Small subpopulations: Sikhs small minority in US (0.2%) barely represented in a random sample

Sample Size Disparity

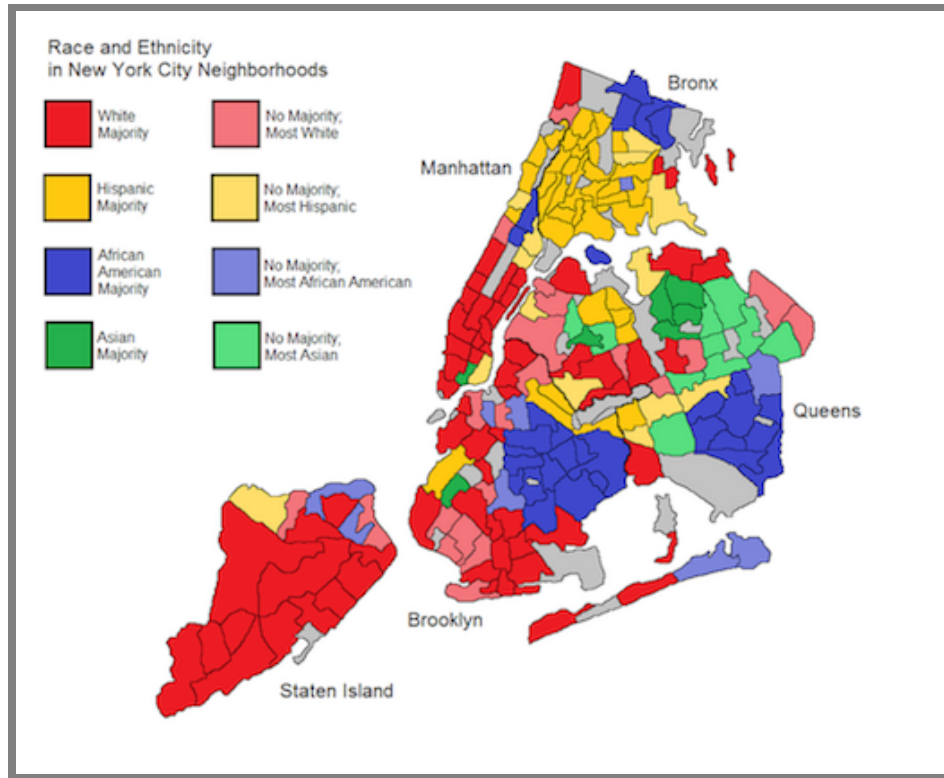
Without intervention:

- Models biased toward populations more represented in target distribution (e.g., Caucasian skin tones)
- ... biased towards population that are easier to sample (e.g., people self-selecting to post to Instagram)
- ... may ignore small minority populations as noise

Typically requires deliberate sampling strategy, intentional oversampling

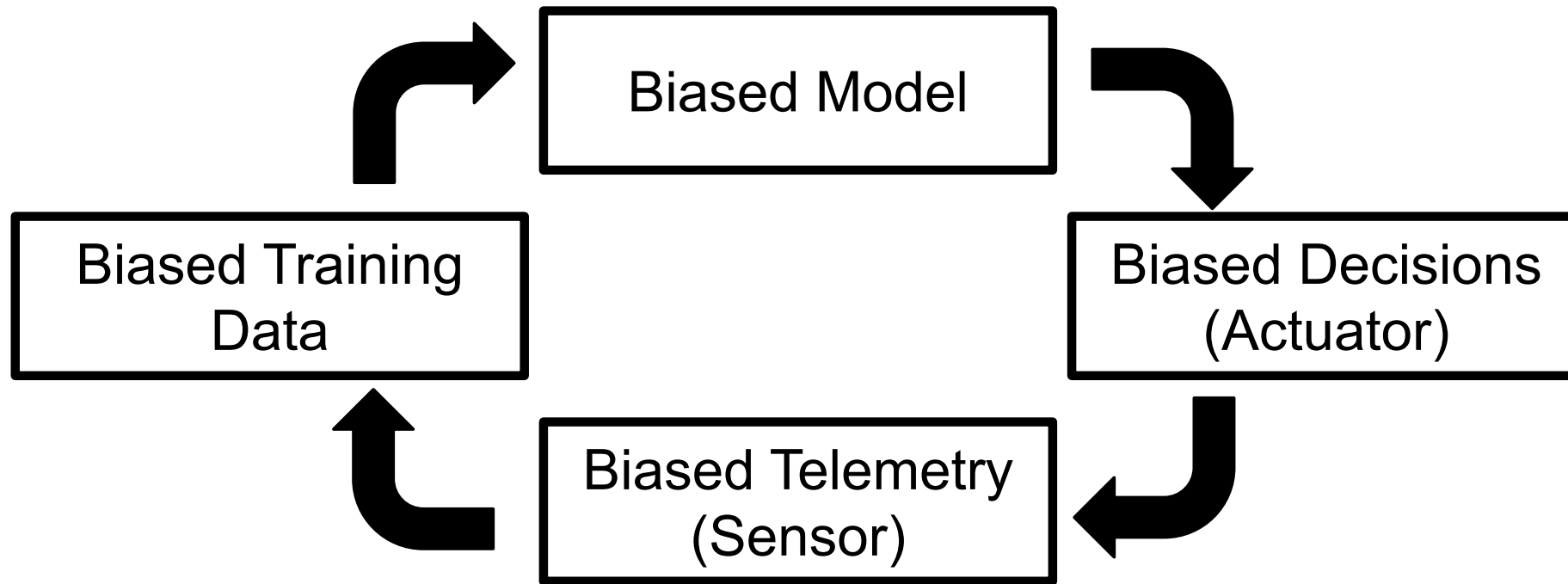
Proxies

Features correlate with protected attribute, remain after removal



- Example: Neighborhood as a proxy for race
- Extracurricular activities as proxy for gender and social class (e.g., “cheerleading”, “peer-mentor for ...”, “sailing team”, “classical music”)

Feedback Loops reinforce Bias



"Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide." -- Cathy O'Neil in [Weapons of Math Destruction](#)

Breakout: College Admission



Scenario: Evaluate applications & identify students likely to succeed

Features: GPA, GRE/SAT, gender, race, undergrad institute, alumni connections, household income, hometown, transcript, etc.

Breakout: College Admission

Scenario: Evaluate applications & identify students who are likely to succeed

Features: GPA, GRE/SAT, gender, race, undergrad institute, alumni connections, household income, hometown, transcript, etc.

As a group, post to #Lecture tagging members:

- **Possible harms:** Allocation of resources? Quality of service? Stereotyping? Denigration? Over-/Under-representation?
- **Sources of bias:** Skewed sample? Tainted labels? Historical bias? Limited features? Sample size disparity? Proxies?

Next lectures

1. Measuring and Improving Fairness at the Model Level
2. Fairness is a System-Wide Concern

Summary

- Many interrelated issues: ethics, fairness, justice, safety, security, ...
- Both legal & ethical dimensions
- Challenges with developing ethical systems / developing systems responsibly
- Large potential for damage: Harm of allocation & harm of representation
- Sources of bias in ML: Skewed sample, tainted labels, limited features, sample size, disparity, proxies

Further Readings

- O’Neil, Cathy. [Weapons of math destruction: How big data increases inequality and threatens democracy](#). Crown Publishing, 2017.
- Barocas, Solon, and Andrew D. Selbst. “[Big data’s disparate impact](#).” Calif. L. Rev. 104 (2016): 671.
- Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. “[A survey on bias and fairness in machine learning](#).” ACM Computing Surveys (CSUR) 54, no. 6 (2021): 1–35.
- Bietti, Elettra. “[From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy](#).” In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 210–219. 2020.

