ETHICS & FAIRNESS IN AI-ENABLED SYSTEMS

Christian Kaestner

(with slides from Eunsuk Kang)

Required reading: 🗎 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
- Analyze a system for harmful feedback loops

OVERVIEW

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.

ETHICAL VS LEGAL

In September 2015, Shkreli received widespread criticism when Turing obtained the manufacturing license for the antiparasitic drug Daraprim and raised its price by a factor of 56 (from USD 13.5 to 750 per pill), leading him to be 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

WITH A FEW LINES OF CODE...

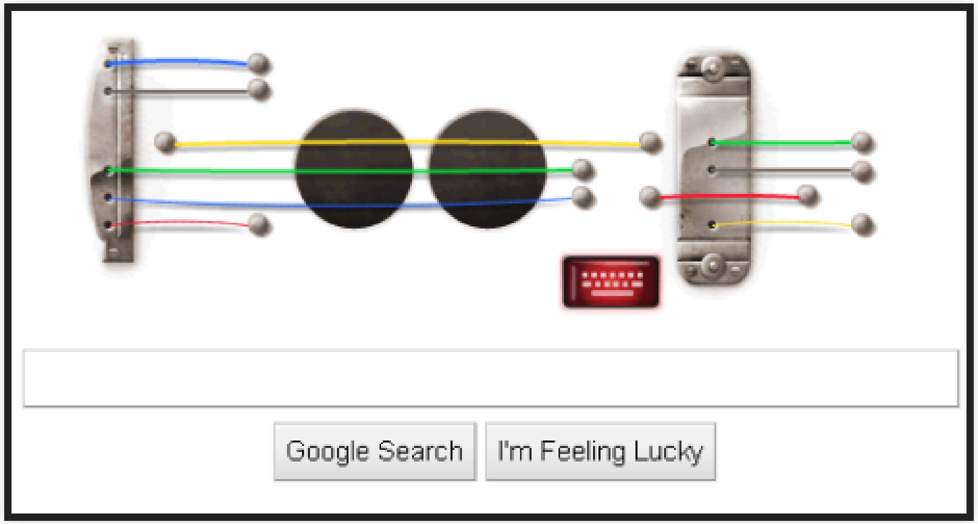
Some airlines may be using algorithms to split up families during flights

Your random airplane seat assignment might not be random at all.

By Aditi Shrikant | aditi@vox.com | Nov 27, 2018, 6:10pm EST

SHARE

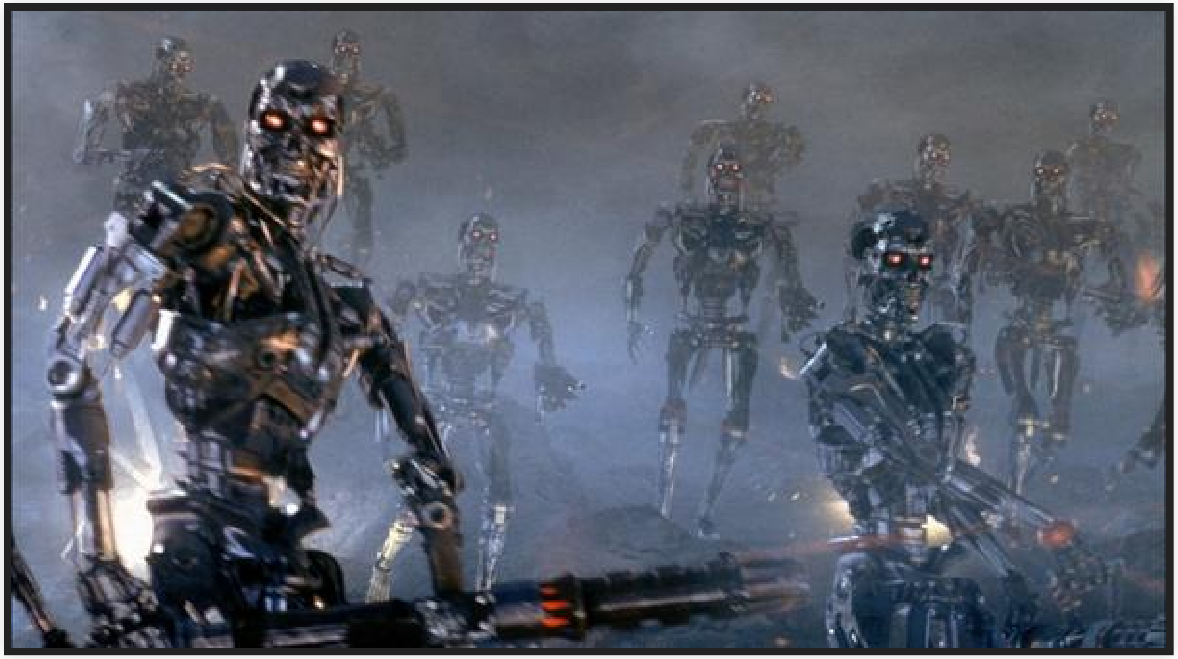
THE IMPLICATIONS OF OUR CHOICES



"Update Jun 17: Wow—in just 48 hours in the U.S., you recorded 5.1 years worth of music—40 million songs—using our doodle guitar. And those songs were played back 870,000 times!"

CONCERNS ABOUT AN AI FUTURE

SAFETY



SAFETY

skoops ❄️⛄☣️
@skoops

The @netatmo servers are down and twitter is already full of freezing people not able to control their heating :D (via [protected]) / cc @internetofshit

8:15 PM · Nov 22, 2018

♡ 2.2K   1.6K people are Tweeting about this

SAFETY

Emily Slackerman Ackerman
@EmilyEAckerman

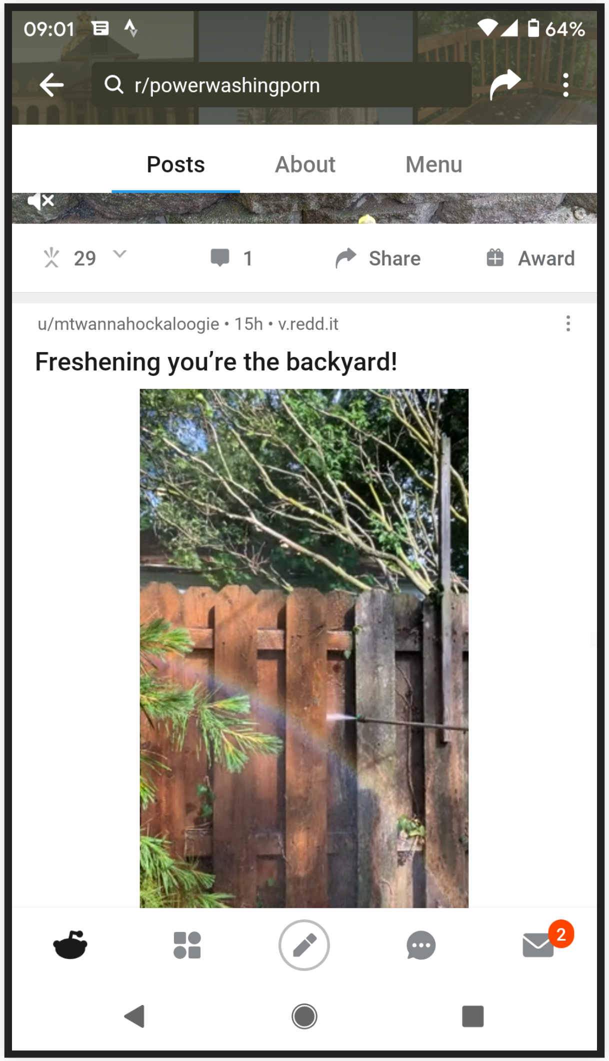
i (in a wheelchair) was just trapped *on* forbes ave by one of these robots, only days after their independent roll out. i can tell that as long as they continue to operate, they are going to be a major accessibility and safety issue. [thread]

Everything we know about the Starship food delivery ro…
The white, 2-foot tall battery-powered delivery robots will be sharing the sidewalk with Oakland pedestrians starti…
pittnews.com

7:27 PM · Oct 21, 2019

4.3K 3.2K people are Tweeting about this

ADDICTION


09:01
64%
r/powerwashingporn
Posts
About
Menu
29
1
Share
Award
u/mtwannahockaloogie • 15h • v.redd.it
Freshening you're the backyard!
2

Speaker notes

Infinite scroll in applications removes the natural breaking point at pagination where one might reflect and stop use.

ADDICTION

NO MERCY NO MALICE

Robinhood Has Gamified Online Trading Into an Addiction

Tech's obsession with addiction will hurt us all

Scott Galloway Follow

Jun 23 · 7 min read ★

Warning: This post contains a discussion of suicide.

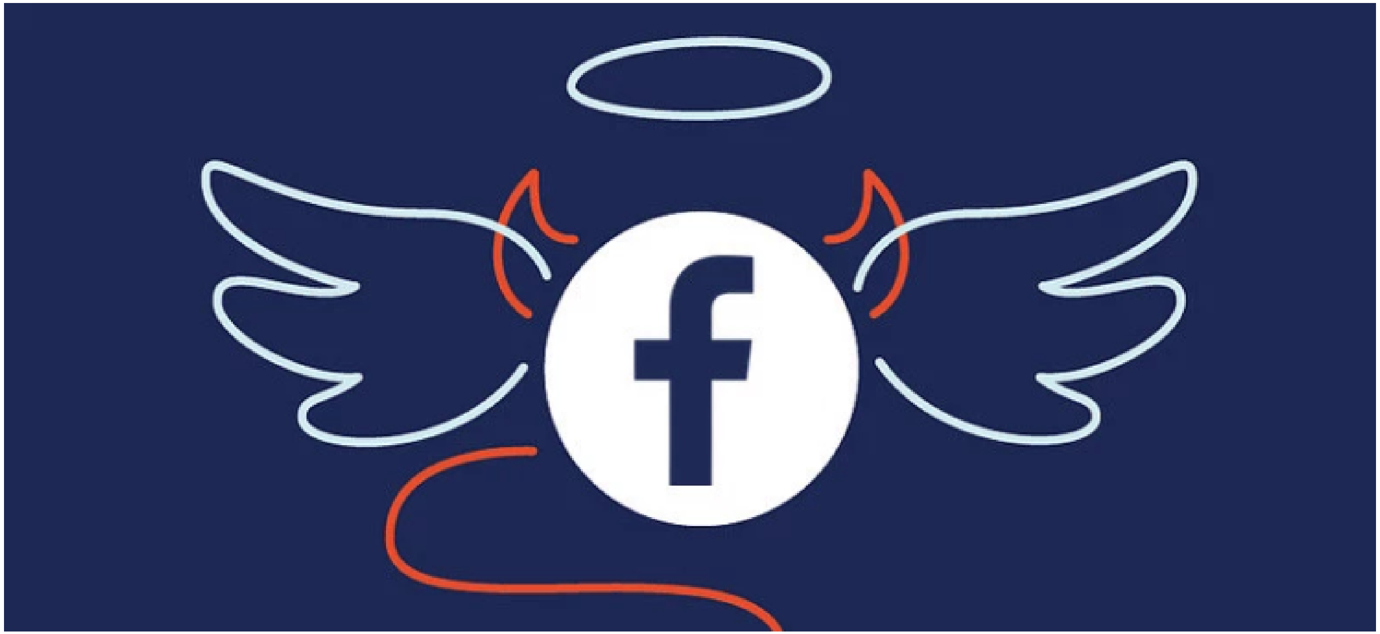
Addiction is the inability to stop consuming a chemical or pursuing an activity although it's causing harm.

I engage with almost every substance or behavior associated with addiction: alcohol, drugs, coffee, porn, sex, gambling, work, spending,

The Morality Of A/B Testing

Josh Constine @joshconstine / 4 years ago

Comment



MENTAL HEALTH

healthline

SUBSCRIBE

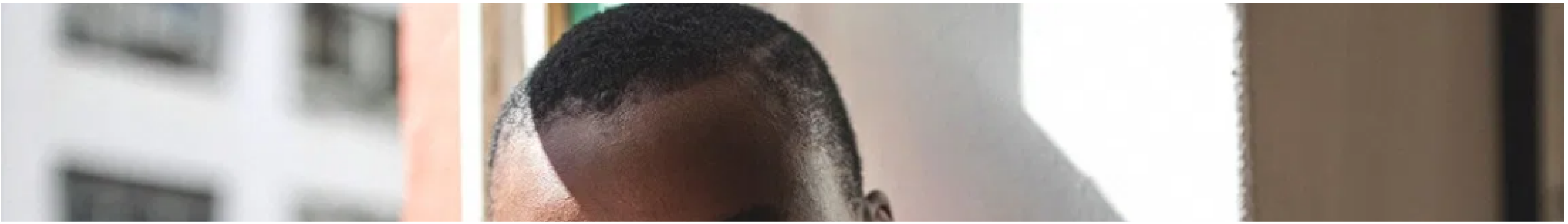
HEALTH NEWS

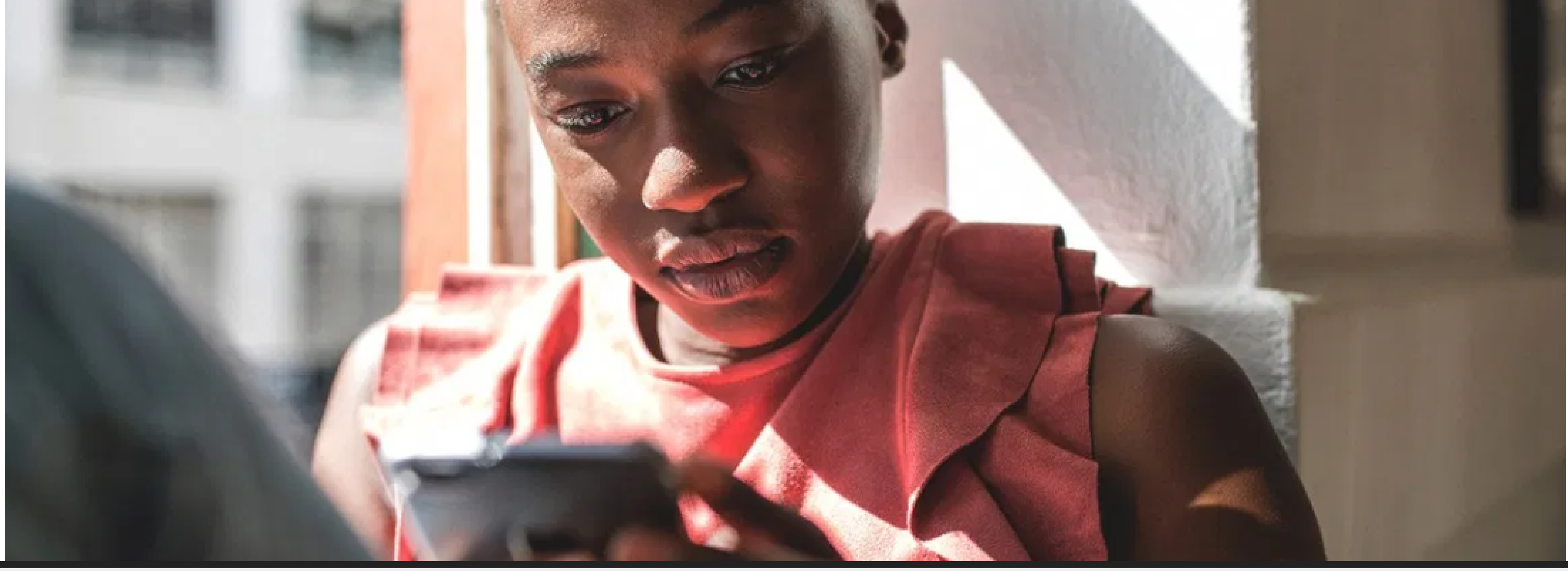
✓ Fact Checked

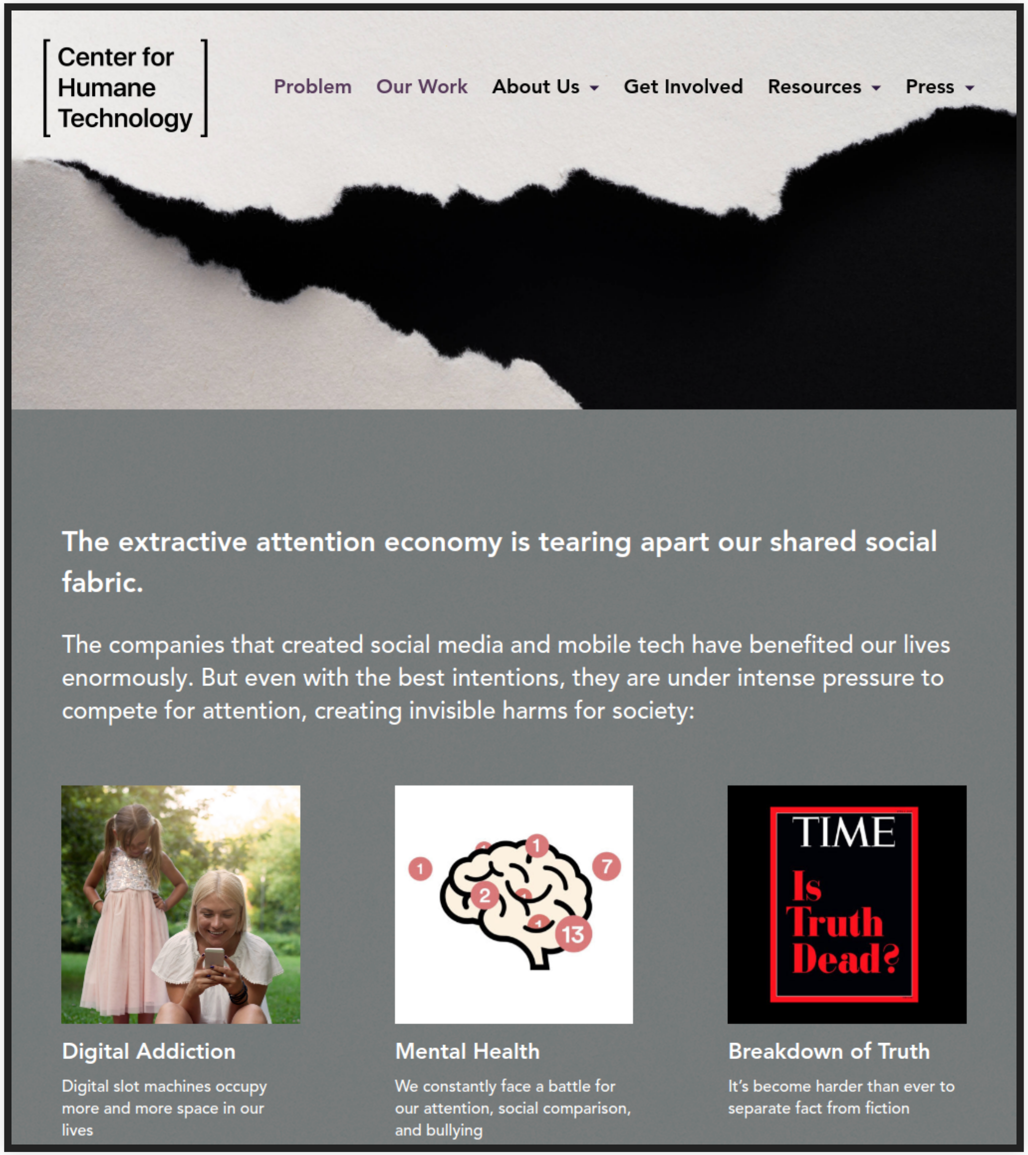
The FOMO Is Real: How Social Media Increases Depression and Loneliness

Written by Gigen Mammoser on December 10, 2018

New research reveals how social media platforms like Facebook can greatly affect your mental health.






Center for Humane Technology
Problem
Our Work
About Us
Get Involved
Resources
Press
The extractive attention economy is tearing apart our shared social fabric.
The companies that created social media and mobile tech have benefited our lives enormously. But even with the best intentions, they are under intense pressure to compete for attention, creating invisible harms for society:
1
1
2
7
13
TIME
Is Truth Dead?
Digital Addiction
Digital slot machines occupy more and more space in our lives
Mental Health
We constantly face a battle for our attention, social comparison, and bullying
Breakdown of Truth
It's become harder than ever to separate fact from fiction

5.9

SOCIETY: UNEMPLOYMENT ENGINEERING / DESKILLING

Speaker notes

The dangers and risks of automating jobs.

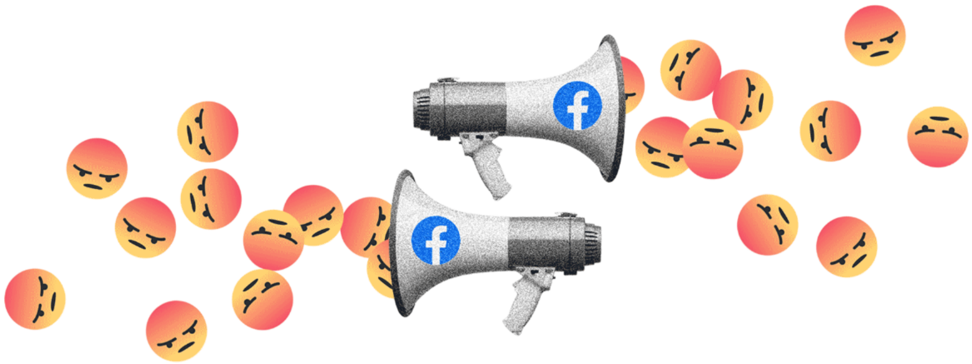
Discuss issues around automated truck driving and the role of jobs.

See for example: Andrew Yang. The War on Normal People. 2019

SOCIETY: POLARIZATION

THE WALL STREET JOURNAL.

SUBSCRIBE | SIGN IN



TECH

Facebook Executives Shut Down Efforts to Make the Site Less Divisive

The social-media giant internally studied how it polarizes users, then largely shelved the research

By Jeff Horwitz and Deepa Seetharaman

May 26, 2020 11:38 am ET

Speaker notes

Recommendations for further readings: https://www.nytimes.com/column/kara-swisher, https://podcasts.apple.com/us/podcast/recode-decode/id1011668648

Also isolation, Cambridge Analytica, collaboration with ICE, ...

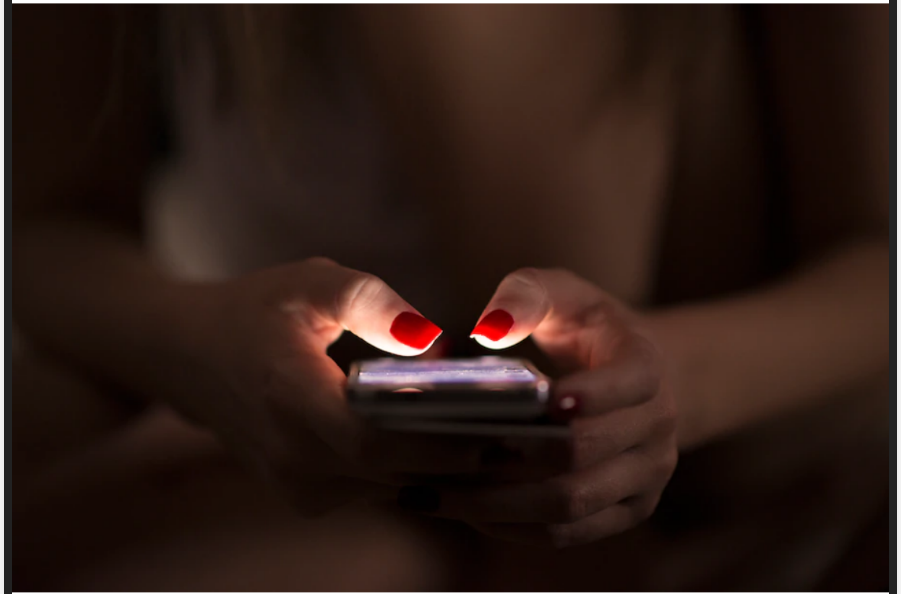
WEAPONS, SURVEILLANCE, SUPPRESSION

The Washington Post

Democracy Dies in Darkness

PostEverything • Perspective

How U.S. surveillance technology is propping up authoritarian regimes



(iStock)

By **Robert Morgus** and **Justin Sherman**

Jan. 17, 2019 at 6:00 a.m. EST

DISCRIMINATION

DHH
@dhh

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

28.5K 10.6K people are Tweeting about this

DISCRIMINATION

DISCRIMINATION

- Unequal treatment in hiring, college admissions, credit rating, insurance, policing, sentencing, advertisement, ...
- Unequal outcomes in healthcare, accident prevention, ...
- Reinforcing patterns in predictive policing with feedback loops

- Technological redlining

ANY OWN EXPERIENCES?

SUMMARY -- SO FAR

- Safety issues
- Addiction and mental health
- Societal consequences: unemployment, polarization, monopolies
- Weapons, surveillance, suppression
- Discrimination, social equity

- Many issues are ethically problematic, but some are legal. Consequences?
- Intentional? Negligence? Unforeseeable?

FAIRNESS

LEGALLY PROTECTED CLASSES (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)

Barocas, Solon and Moritz Hardt. "Fairness in machine learning." NIPS Tutorial 1 (2017).

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

Barocas, Solon and Moritz Hardt. "Fairness in machine learning." NIPS Tutorial 1 (2017).

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

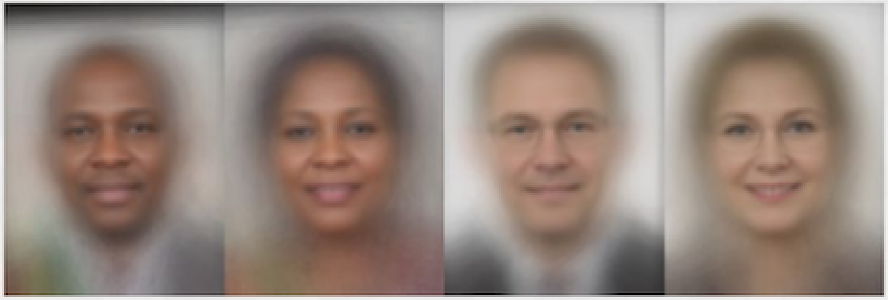
Justice

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.

HARMS OF ALLOCATION

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

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

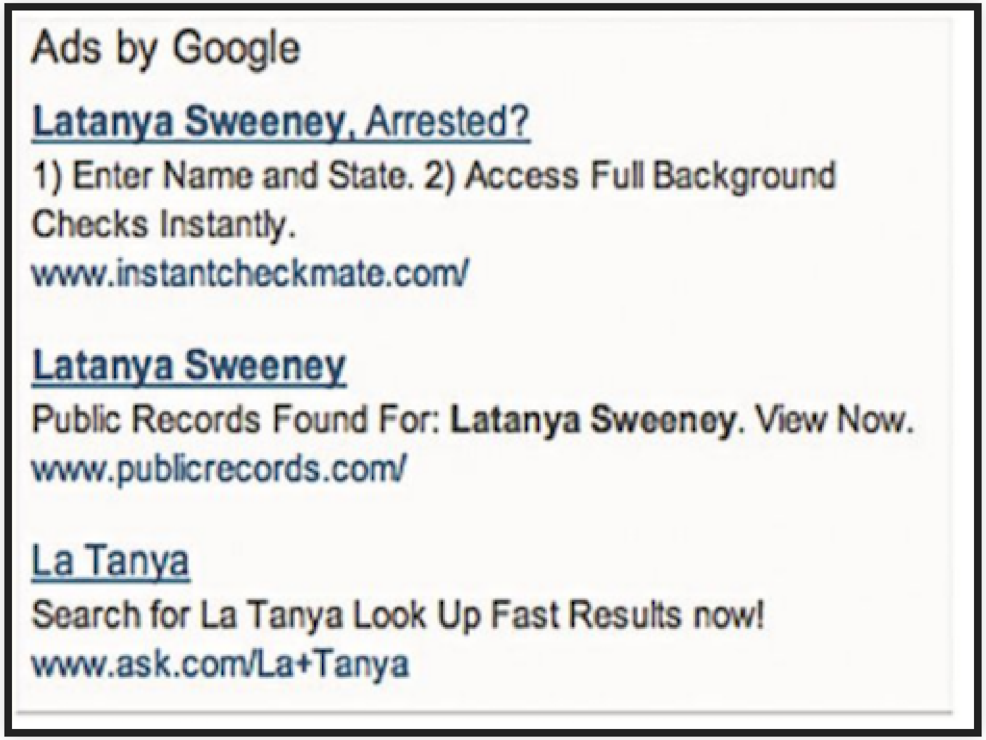


Other examples?

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Buolamwini & Gebru, ACM FAT* (2018).

HARMS OF REPRESENTATION

- Reinforce stereotypes, subordination along the lines of identity



Other examples?

Latanya Sweeney. Discrimination in Online Ad Delivery, SSRN (2013).

IDENTIFYING 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.

Swati Gupta, Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, Jean GarciaGathright. Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019. (slides)

THE 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

WHY CARE ABOUT FAIRNESS?

- Obey the law
- Better product, serving wider audiences
- Competition
- Responsibility
- PR

Examples?

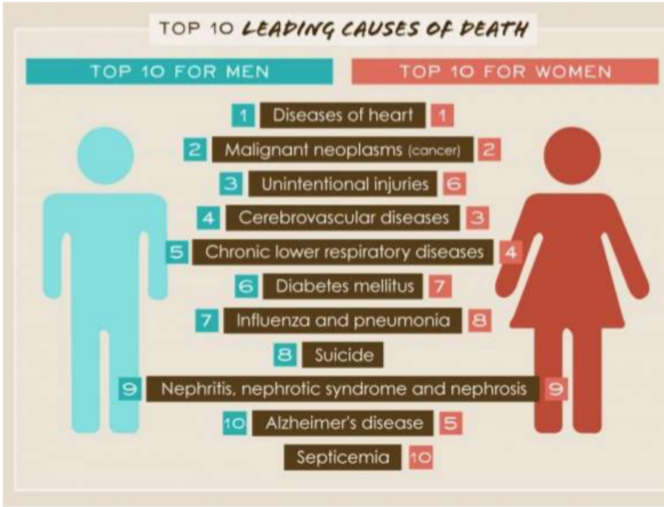
Which argument appeals to which stakeholders?

Swati Gupta, Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, Jean GarciaGathright. Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019. (slides)

CASE STUDY: COLLEGE ADMISSION

- Objective: Decide "Is this student likely to succeed"?
- Possible harms: Allocation of resources? Quality of service? Stereotyping? Denigration? Over-/Under-representation?

NOT ALL DISCRIMINATION IS HARMFUL



- Loan lending: Gender discrimination is illegal.
- Medical diagnosis: Gender-specific diagnosis may be desirable.
- Discrimination is a **domain-specific** concept!

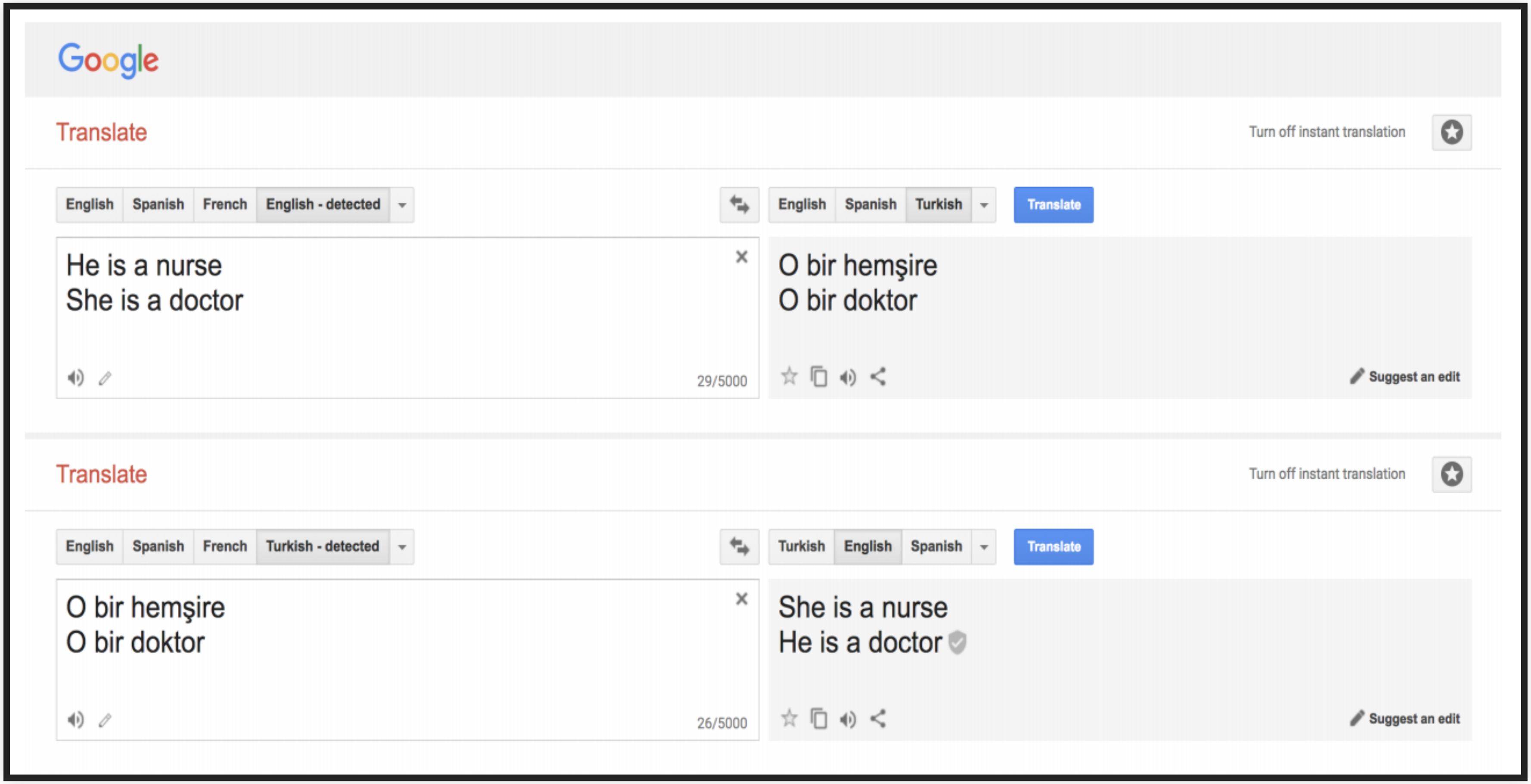
Other examples?

ON TERMINOLOGY

- Bias and discrimination are technical terms in machine learning
  - selection bias, reporting bias, bias of an estimator, inductive/learning bias
  - discrimination refers to distinguishing outcomes (classification)
- The problem is *unjustified* differentiation, ethical issues
  - practical irrelevance
  - moral irrelevance

SOURCES OF BIAS

WHERE DOES THE BIAS COME FROM?



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

SOURCES OF BIAS

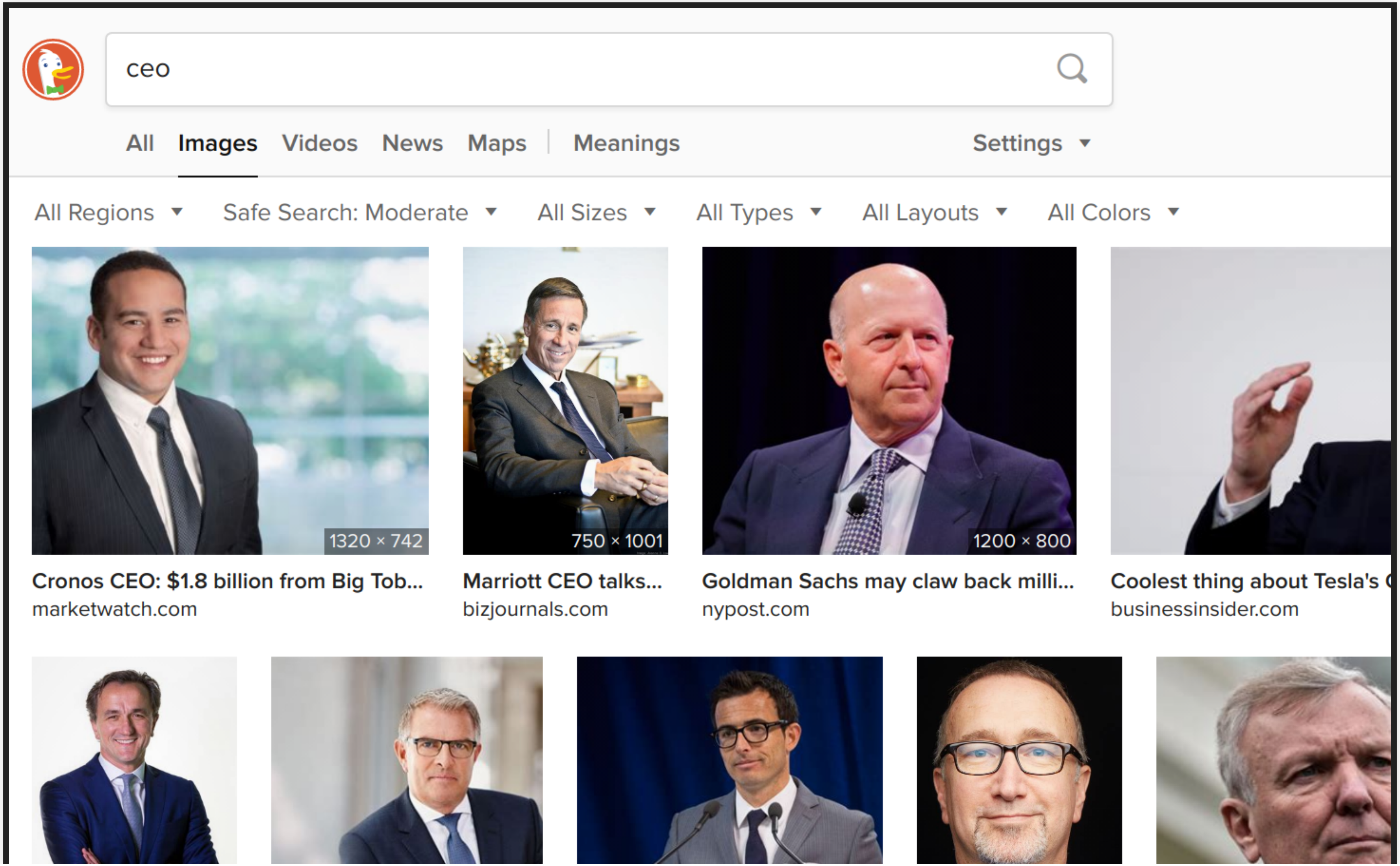
- Tainted examples / historical bias
- Skewed sample
- Limited features
- Sample size disparity
- Proxies

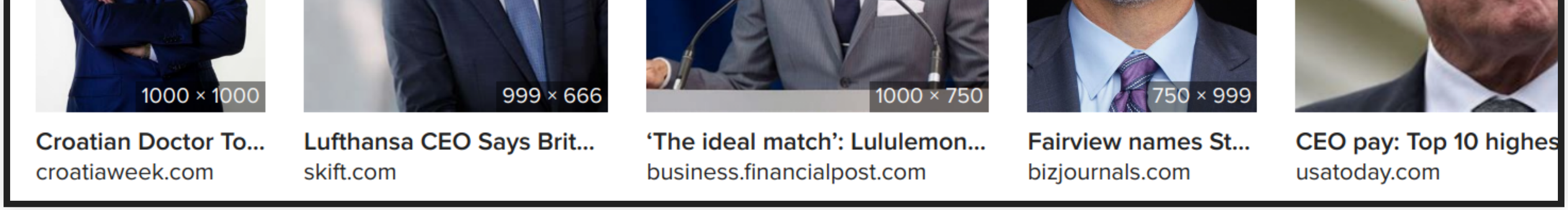
Barocas, Solon, and Andrew D. Selbst. "Big data's disparate impact." Calif. L. Rev. 104 (2016): 671.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "]
(https://arxiv.org/pdf/1908.09635.pdf." arXiv preprint arXiv:1908.09635 (2019).

HISTORICAL BIAS

Data reflects past biases, not intended outcomes




1000 × 1000
Croatian Doctor To...
croatiaweek.com
999 × 666
Lufthansa CEO Says Brit...
skift.com
1000 × 750
‘The ideal match’: Lululemon...
business.financialpost.com
750 × 999
Fairview names St...
bizjournals.com
CEO pay: Top 10 highes
usatoday.com

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."

TAINTED EXAMPLES

Samples or labels reflect human bias

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"

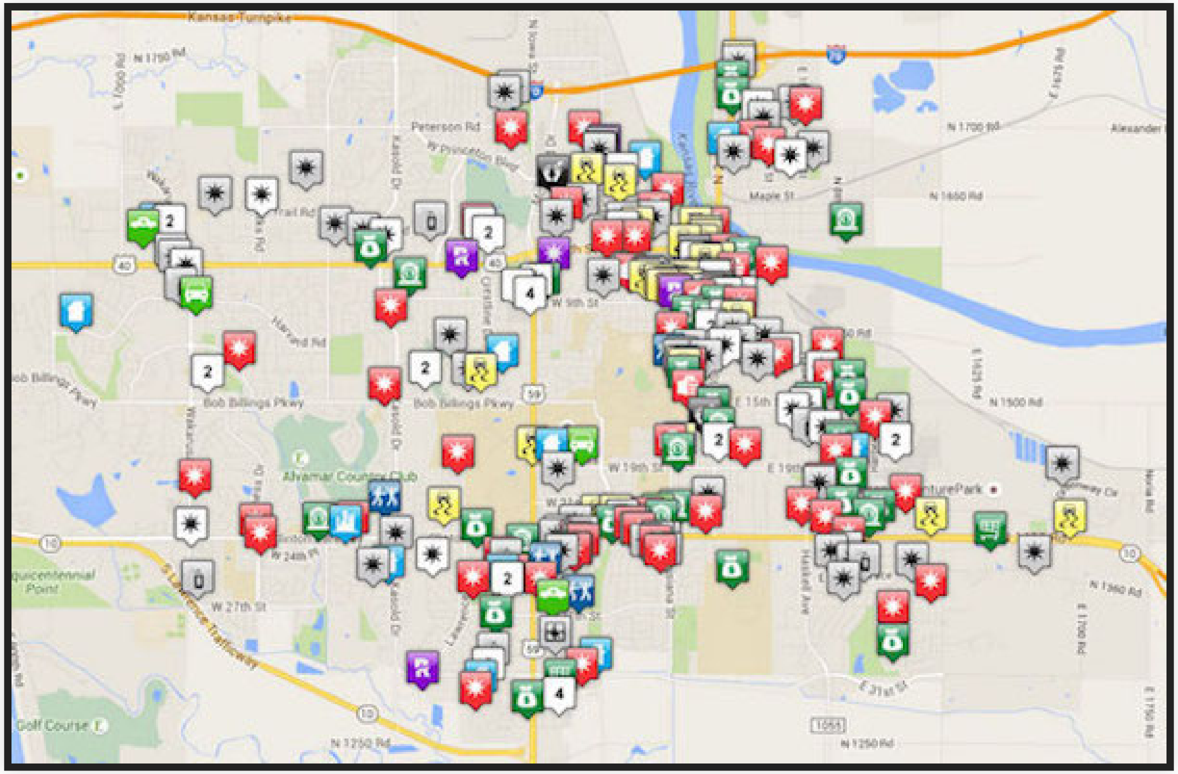
By James Vincent | Oct 10, 2018, 7:09am EDT

Speaker notes

- Bias in the dataset caused by humans
- Some labels created manually by employers
- Dataset "tainted" by biased human judgement

SKEWED SAMPLE

Crime prediction for policing strategy



Speaker notes

Initial bias in the data set, amplified through feedback loop

Other example: Street Bump app in Boston (2012) to detect potholes while driving favors areas with higher smartphone adoption

LIMITED FEATURES

Features used are less informative/reliable for certain subpopulations

Example: "Leave of absence" as feature in employee performance review

Speaker notes

- Features are less informative or reliable for certain parts of the population
- Features that support accurate prediction for the majority may not do so for a minority group
- Example: Employee performance review
    - "Leave of absence" as a feature (an indicator of poor performance)
    - Unfair bias against employees on parental leave

SAMPLE SIZE DISPARITY

Less training data available for certain subpopulations

Example: "Shirley Card" used for color calibration

Speaker notes

- Less data available for certain parts of the population
- Example: "Shirley Card"
  - Used by Kodak for color calibration in photo films
  - Most "Shirley Cards" used Caucasian models
  - Poor color quality for other skin tones

Chukwuemeka Afigbo
@nke_ise

If you have ever had a problem grasping the importance of diversity in tech and its impact on society, watch this video

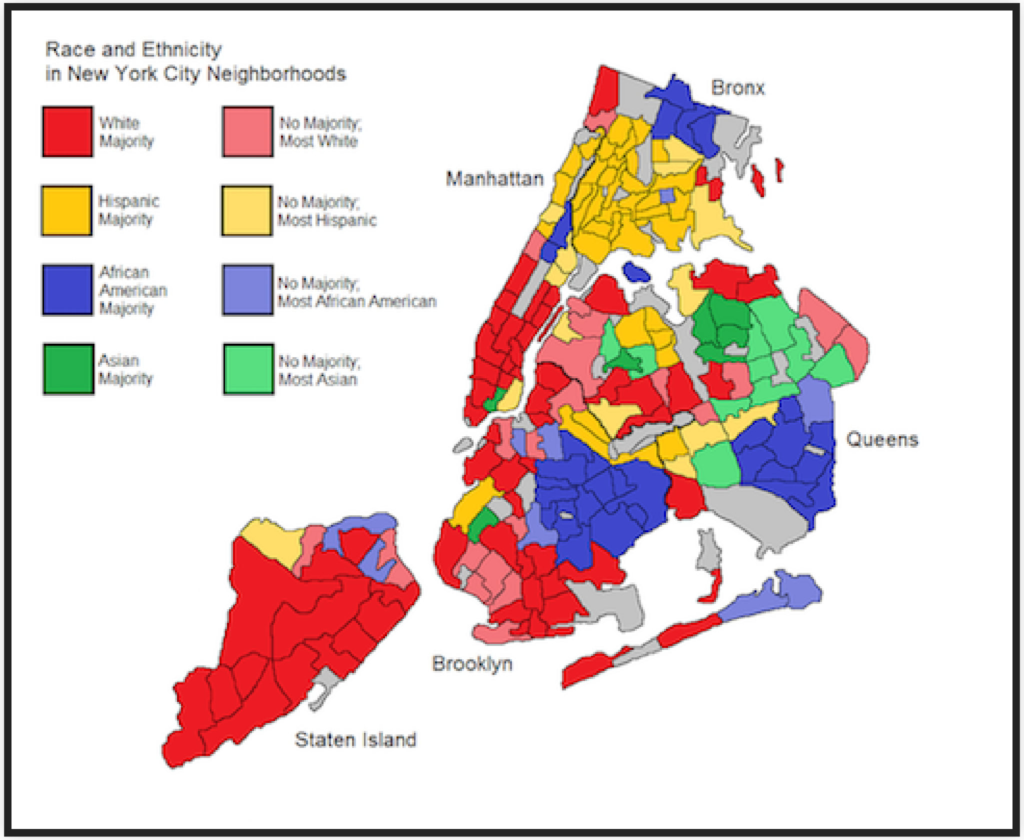
9:48 AM · Aug 16, 2017

216.6K   155K people are Tweeting about this

7.9

PROXIES

Features correlate with protected attributes



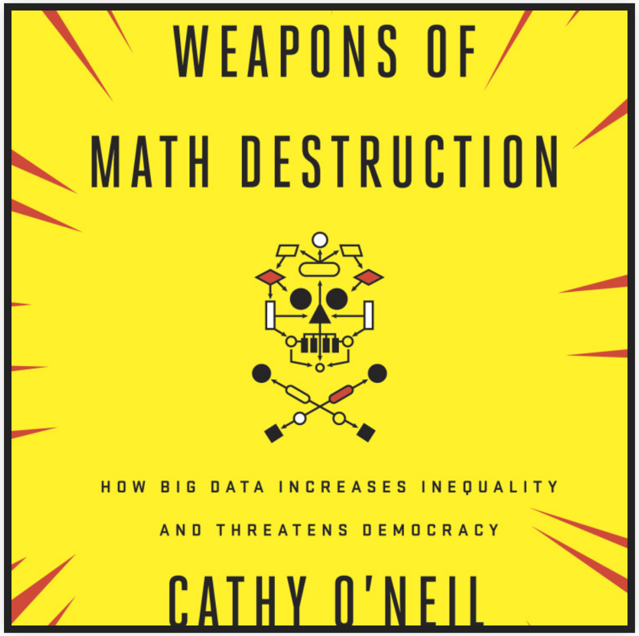
Speaker notes

- Certain features are correlated with class membership
- Example: Neighborhood as a proxy for race
- Even when sensitive attributes (e.g., race) are erased, bias may still occur

CASE STUDY: COLLEGE ADMISSION

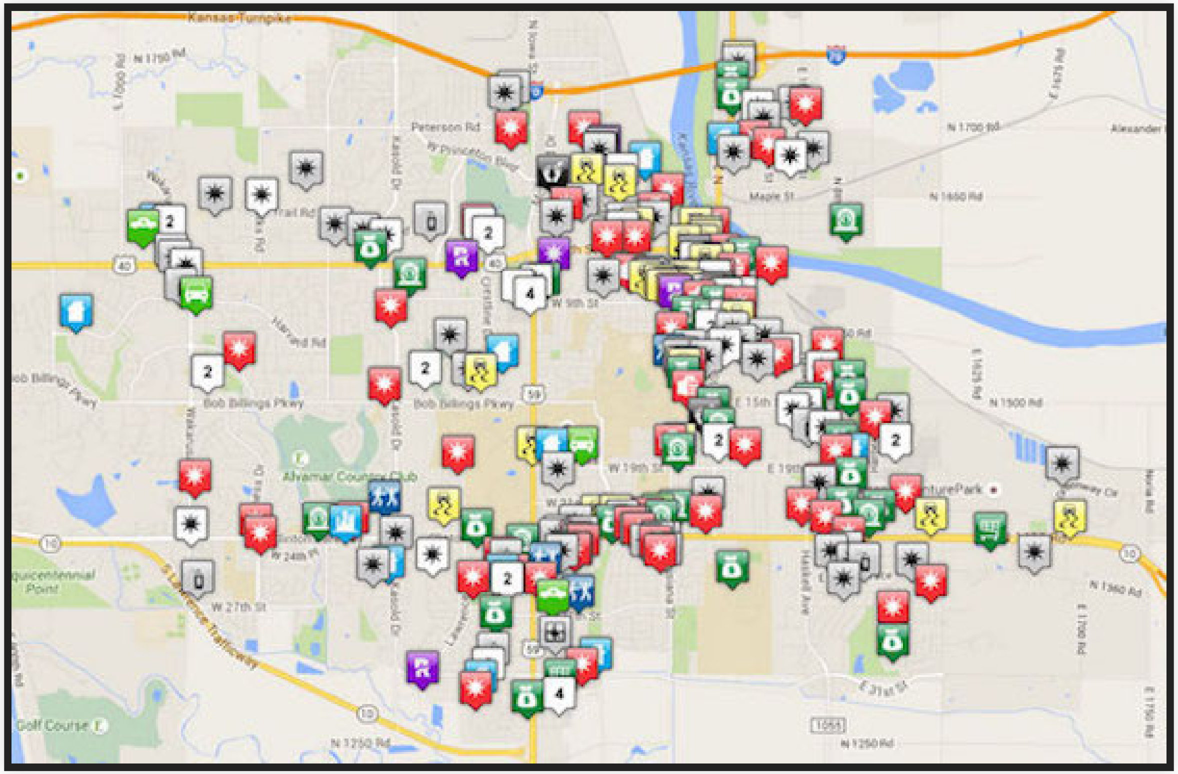
- Classification: Is this student likely to succeed?
- Features: GPA, SAT, race, gender, household income, city, etc.,
- **Discuss:** Historical bias? Skewed sample? Tainted examples? Limited features? Sample size disparity? Proxies?

MASSIVE POTENTIAL DAMAGE



O'Neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2016.

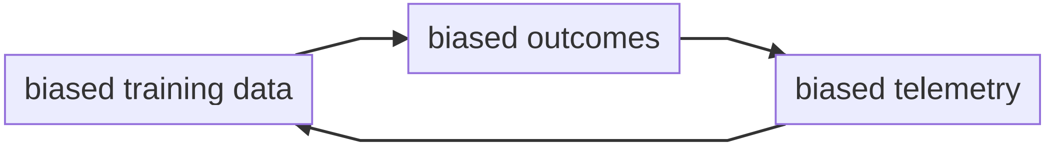
EXAMPLE: PREDICTIVE POLICING



with a few lines of code...

A person who scores as 'high risk' is likely to be unemployed and to come from a neighborhood where many of his friends and family have had run-ins with the law. Thanks in part to the resulting high score on the evaluation, he gets a longer sentence, locking him away for more years in a prison where he's surrounded by fellow criminals—which raises the likelihood that he'll return to prison. He is finally released into the same poor neighborhood, this time with a criminal record, which makes it that much harder to find a job. If he commits another crime, the recidivism model can claim another success. But in fact the model itself contributes to a toxic cycle and helps to sustain it. -- Cathy O'Neil in Weapons of Math Destruction

FEEDBACK LOOPS



"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]

KEY PROBLEMS

- We trust algorithms to be objective, may not question their predictions
- Often designed by and for privileged/majority group
- Algorithms often black box (technically opaque and kept secret from public)
- Predictions based on correlations, not causation; may depend on flawed statistics
- Potential for gaming/attacks
- Despite positive intent, feedback loops may undermine the original goals

O'Neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2016.

"WEAPONS OF MATH DESTRUCTION"

- Algorithm evaluates people
  - e.g., credit, hiring, admissions, recidivism, advertisement, insurance, healthcare
- Widely used for life-affecting decisions
- Opaque and not accountable, no path to complain
- Feedback loop

O'Neil, Cathy. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2016.

SUMMARY

- Many interrelated issues: ethics, fairness, justice, safety, security, ...
- Many many many potential issues
- Consider fairness when it's the law and because it's ethical
- Large potential for damage: Harm of allocation & harm of representation
- Sources of bias in ML: skewed sample, tainted examples, limited features, sample size, disparity, proxies
- Be aware of feedback loops

- Recommended readings: Weapons of Math Destructions and several tutorials on ML fairness
- **Next**: Definitions of fairness, measurement, testing for fairness