

# 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](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.

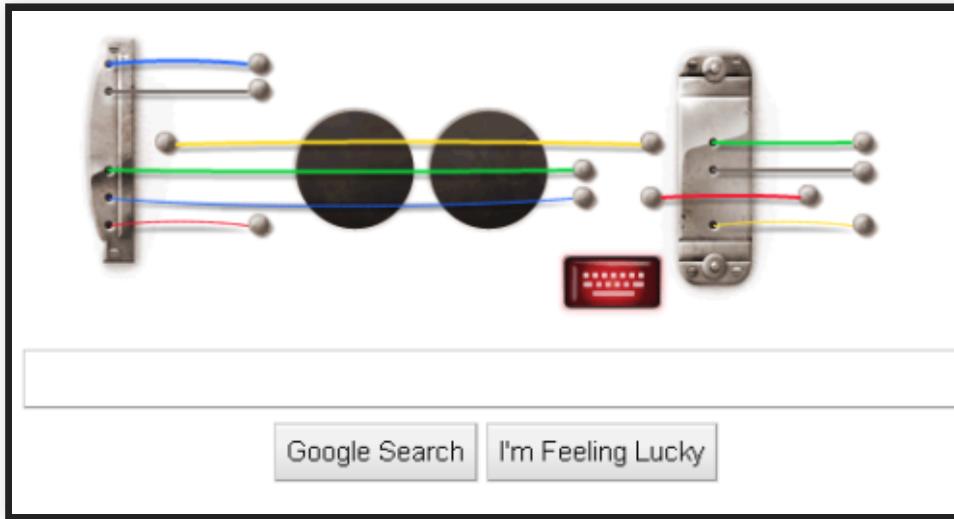
By Aditi Shrikant | [aditi@vox.com](mailto:aditi@vox.com) | Nov 27, 2018, 6:10pm EST



SHARE



# THE IMPLICATIONS OF OUR CHOICES



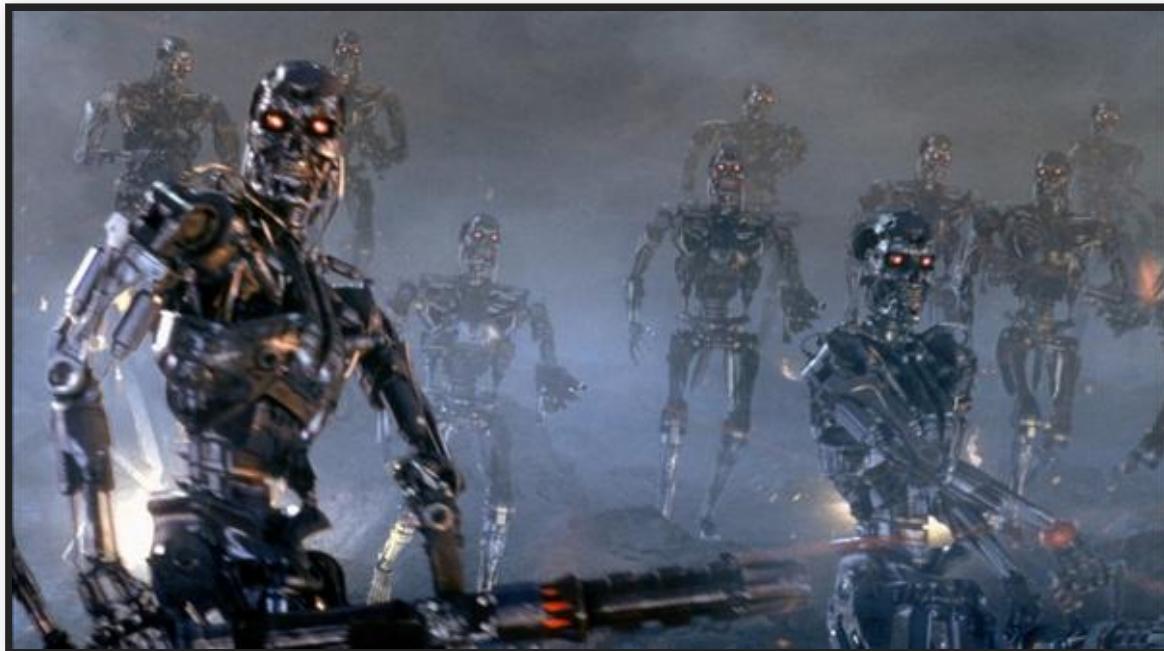
*“Update Jun 17: Wow—in just 48 hours in the U.S., you recorded 5.1 years worth of music—40 million songs—using our doodle guitar. And those songs were played back 870,000 times!”*



# CONCERNS ABOUT AN AI FUTURE



# SAFETY



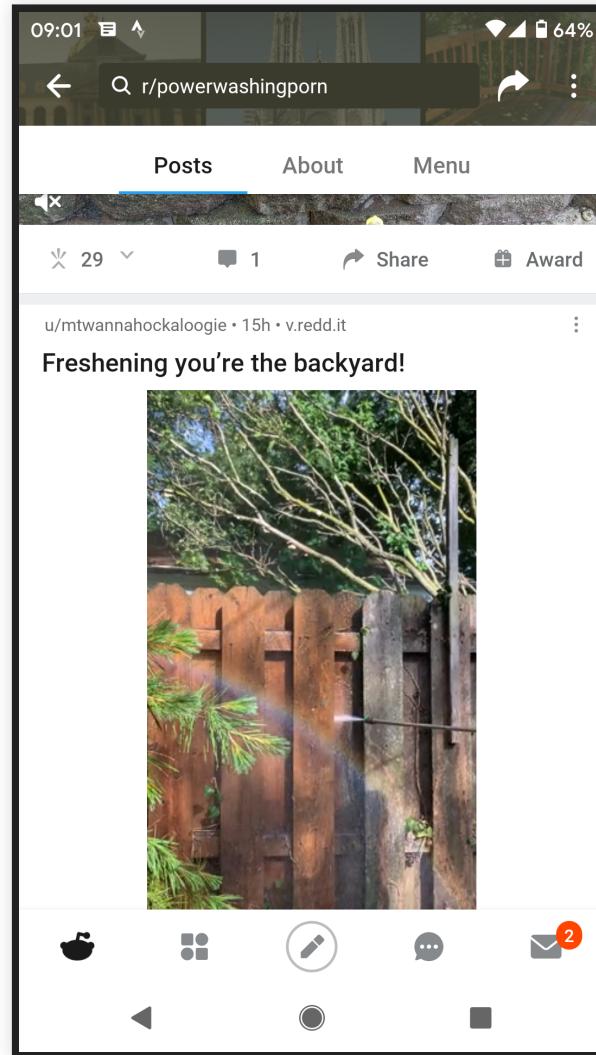
# SAFETY

*Tweet*

# SAFETY

*Tweet*

# ADDICTION



## 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.*

**A**ddiction 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,

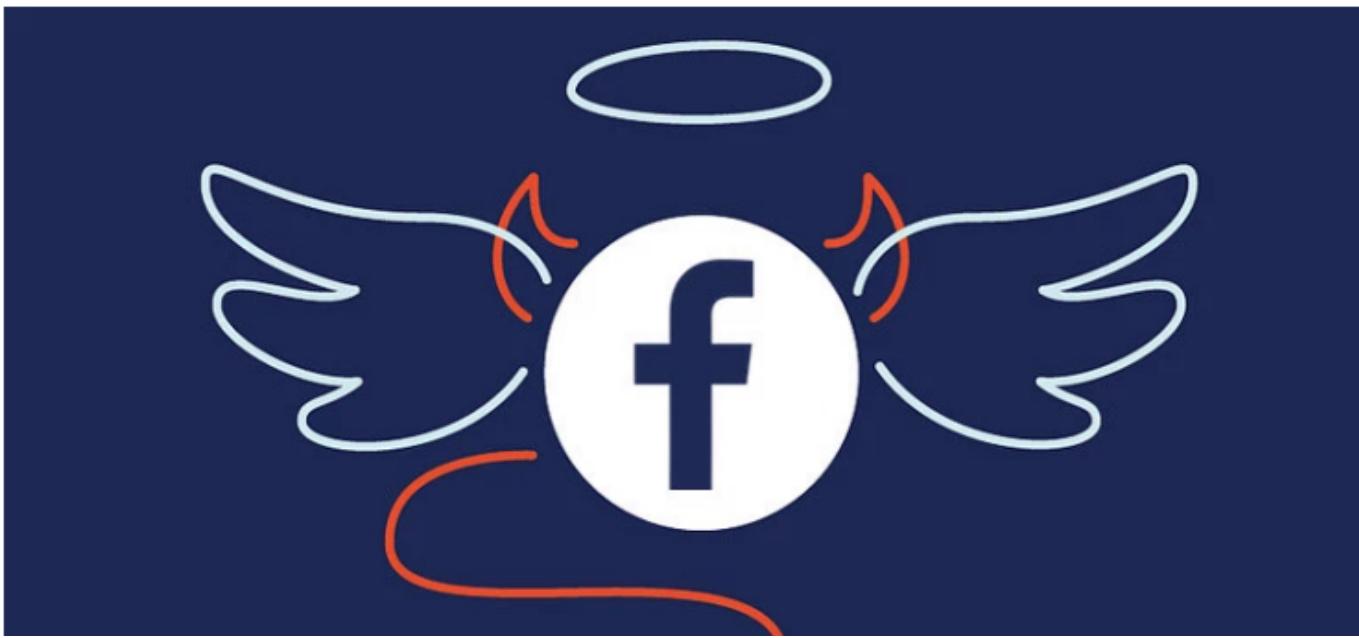
X

# The Morality Of A/B Testing

**Josh Constine** @joshconstine / 4 years ago



Comment





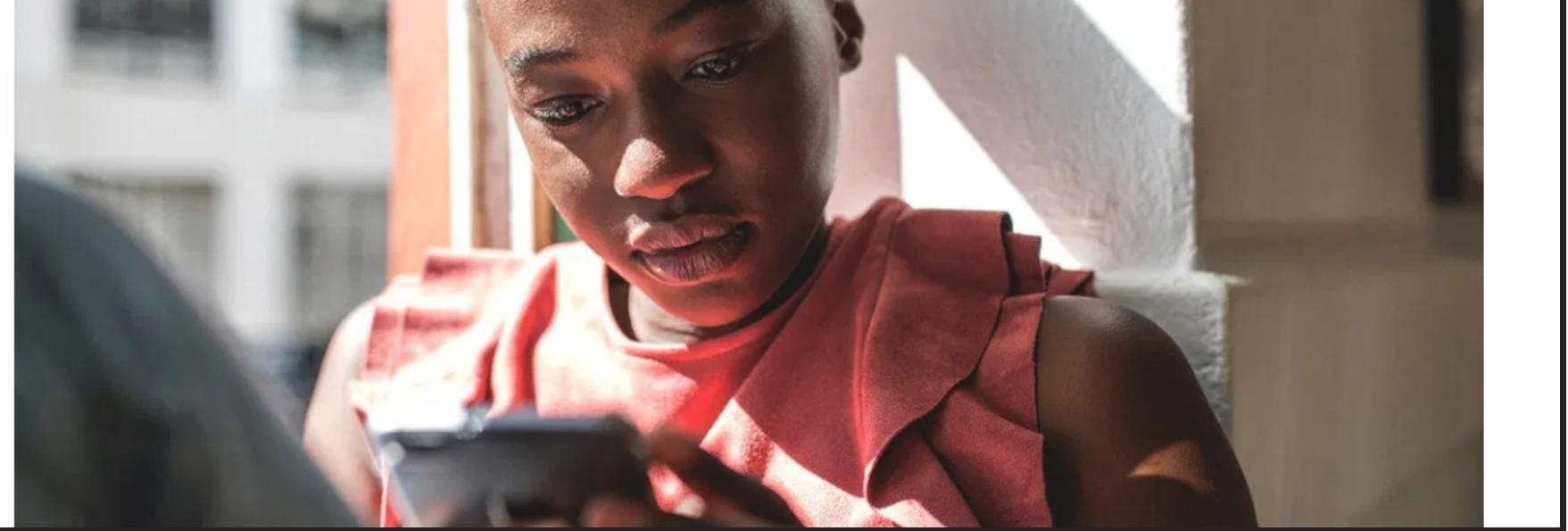
[HEALTH NEWS](#) [Fact Checked](#)

# 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.





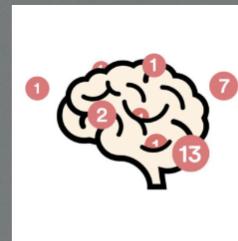
## The extractive attention economy is tearing apart our shared social fabric.

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:



### 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

# SOCIETY: UNEMPLOYMENT ENGINEERING / DESKILLING



## Speaker notes

The dangers and risks of automating jobs.

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

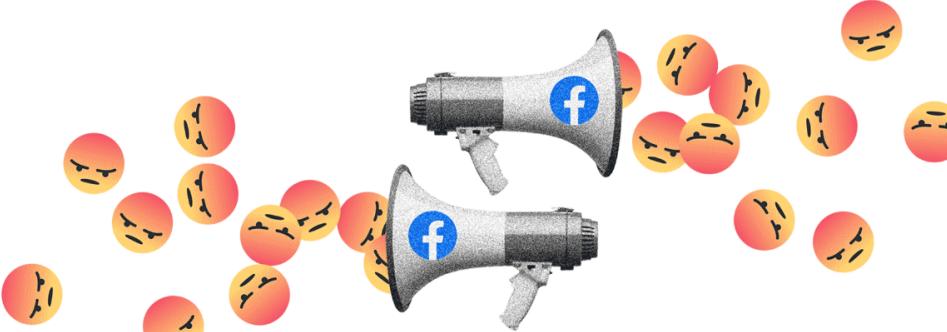
See for example: Andrew Yang. The War on Normal People. 2019



# SOCIETY: POLARIZATION

≡ THE WALL STREET JOURNAL. SEARCH

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, ...



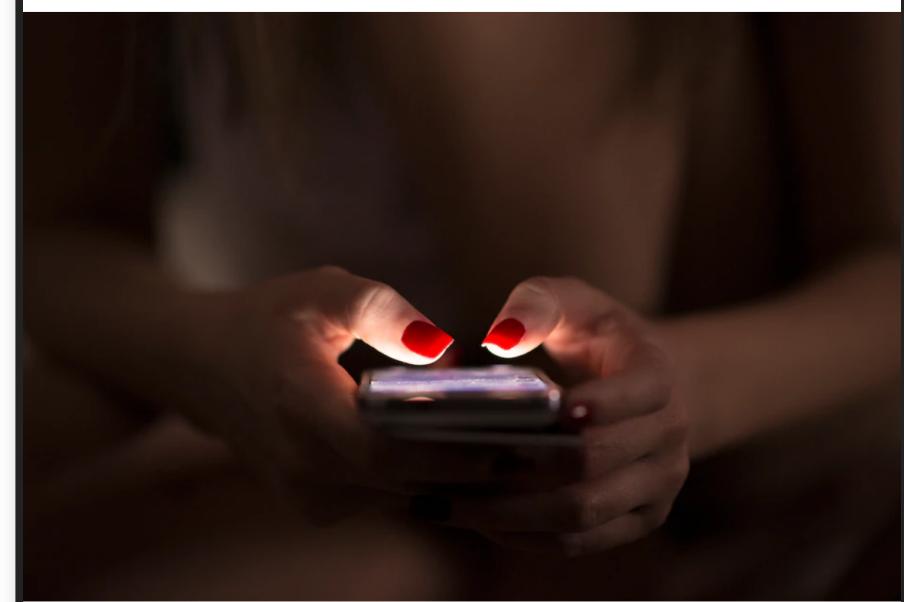
# WEAPONS, SURVEILLANCE, SUPPRESSION



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

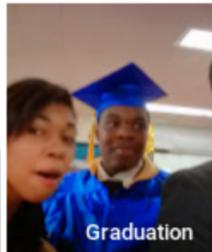
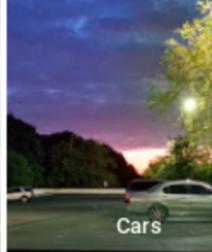
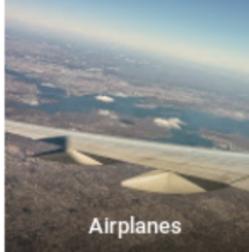
*Tweet*

# DISCRIMINATION

 stop hoarding and work with your ...  
@jackyalcine

Follow ▾

Google Photos, y'all fucked up. My friend's not a gorilla.



6:22 PM - 28 Jun 2015

3,352 Retweets 2,767 Likes

 Kahuna

232 3.4K 2.8K

# 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

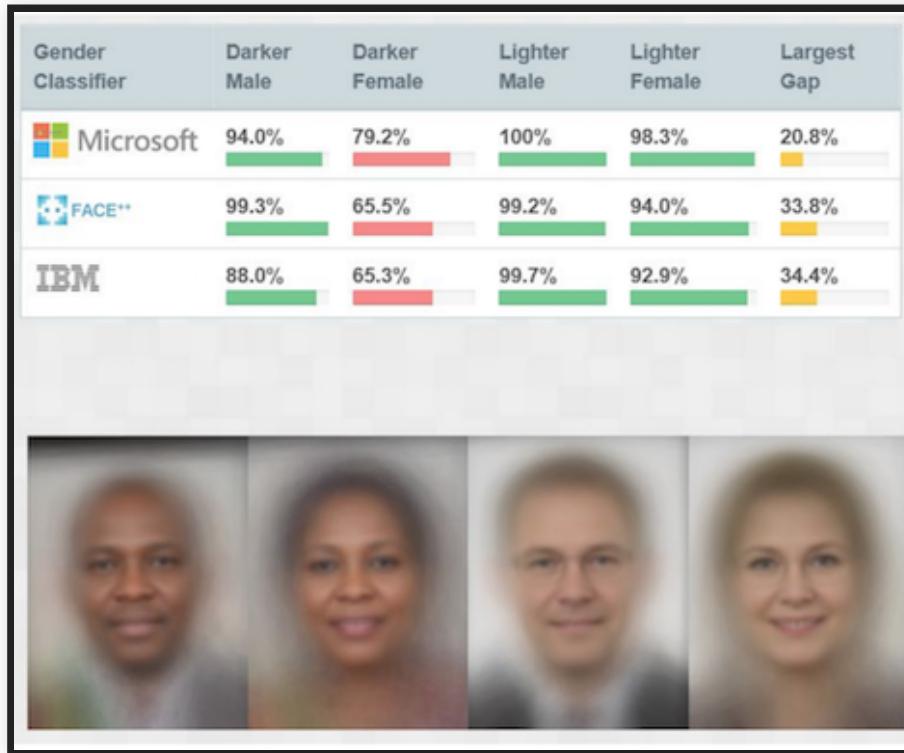


All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**.  
The systemic barrier has been removed.



# HARMS OF ALLOCATION

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



Other examples?





# HARMS OF REPRESENTATION

- Reinforce stereotypes, subordination along the lines of identity

Ads by Google

[Latanya Sweeney, Arrested?](#)

1) Enter Name and State. 2) Access Full Background Checks Instantly.

[www.instantcheckmate.com/](http://www.instantcheckmate.com/)

[Latanya Sweeney](#)

Public Records Found For: Latanya Sweeney. View Now.

[www.publicrecords.com/](http://www.publicrecords.com/)

[La Tanya](#)

Search for La Tanya Look Up Fast Results now!

[www.ask.com/La+Tanya](http://www.ask.com/La+Tanya)

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



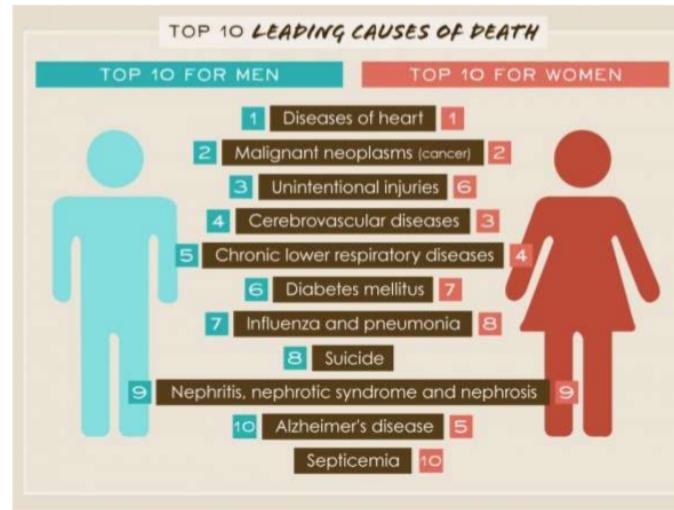
- Objective: Decide "Is this student likely to succeed"?
- Possible harms: Allocation of resources? Quality of service? Stereotyping? Denigration? Over-/Under-representation?

# NOT ALL DISCRIMINATION IS HARMFUL



FEDERAL TRADE COMMISSION

Mortgage discrimination is against the law.



- 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?

The image displays two side-by-side screenshots of the Google Translate interface, illustrating gender bias in machine translation.

**Top Screenshot (English to Turkish):**

- Input:** "He is a nurse  
She is a doctor" (English)
- Output:** "O bir hemşire  
O bir doktor" (Turkish)
- Note:** The output preserves the gender of the input, translating "he" as "O bir hemşire" and "she" as "O bir doktor".

**Bottom Screenshot (Turkish to English):**

- Input:** "O bir hemşire  
O bir doktor" (Turkish)
- Output:** "She is a nurse  
He is a doctor" (English)
- Note:** The output preserves the gender of the input, translating "she" as "She is a nurse" and "he" as "He is a doctor".

In both cases, the model has assigned a female gender to the first sentence and a male gender to the second sentence, reflecting a gender bias present in the training data.

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

Baracas, 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. "]  
[\(<https://arxiv.org/pdf/1908.09635.pdf>.\)](https://arxiv.org/pdf/1908.09635.pdf) arXiv preprint arXiv:1908.09635 (2019).

# HISTORICAL BIAS

*Data reflects past biases, not intended outcomes*

The screenshot shows a search results page for the query "ceo". The interface includes a search bar with the term "ceo", a magnifying glass icon, and a navigation bar with tabs for All, Images, Videos, News, Maps, Meanings, and Settings. Below the search bar are filters for All Regions, Safe Search (set to Moderate), All Sizes, All Types, All Layouts, and All Colors. The main content area displays five search results, each featuring a portrait of a man in a suit and a brief description:

- Cronos CEO: \$1.8 billion from Big Tob...**  
marketwatch.com  
  
1320 × 742
- Marriott CEO talks...**  
bizjournals.com  
  
750 × 1001
- Goldman Sachs may claw back milli...**  
nypost.com  
  
1200 × 800
- Coolest thing about Tesla's C**  
businessinsider.com  


Below these results are two more partial portraits of men in suits.



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 highest  
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

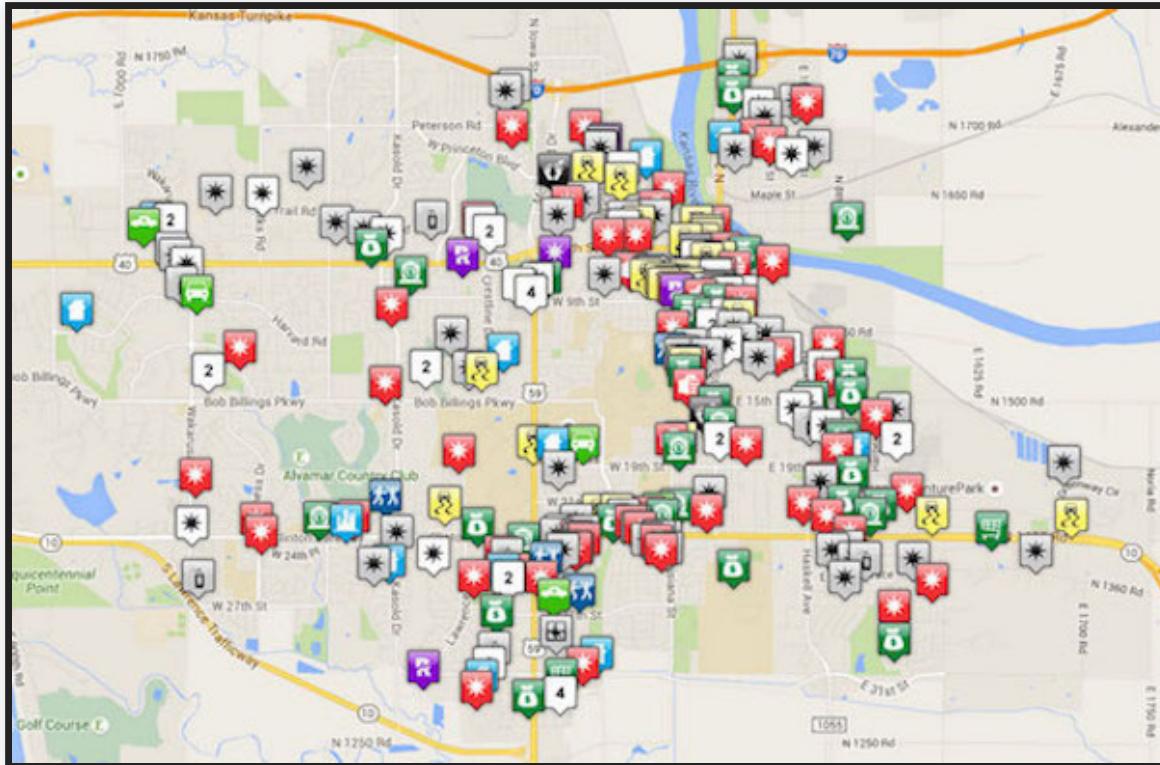
## Speaker notes

- 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

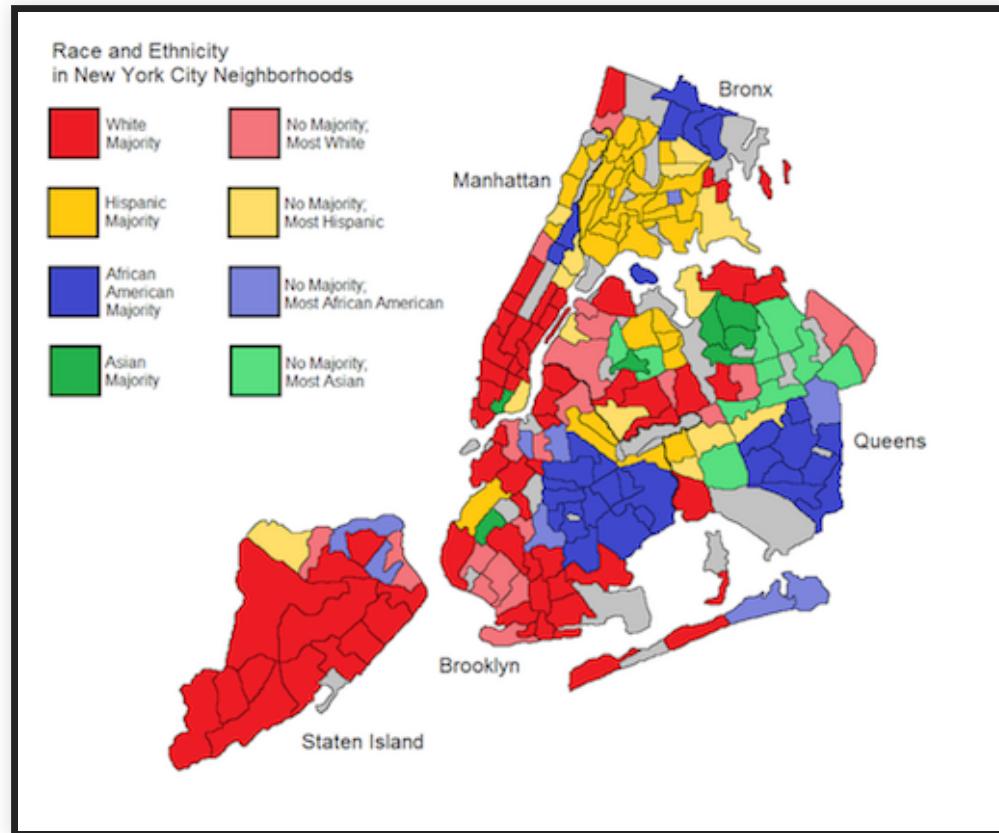


*Tweet*



# 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

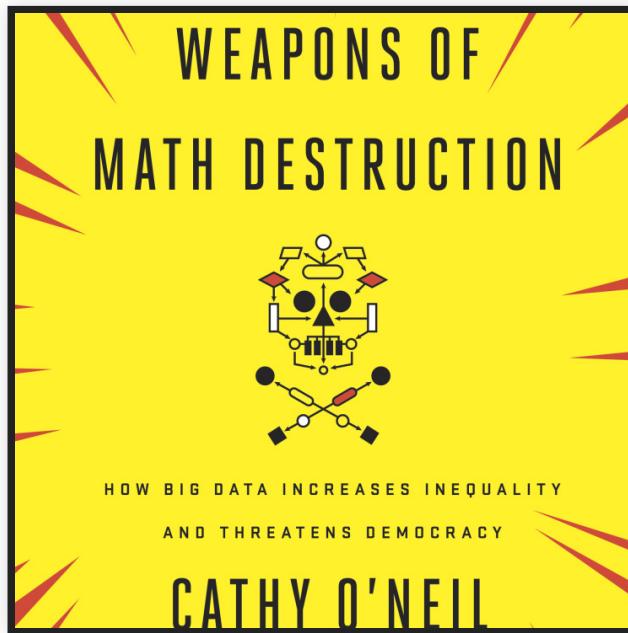


# CASE STUDY: COLLEGE ADMISSION



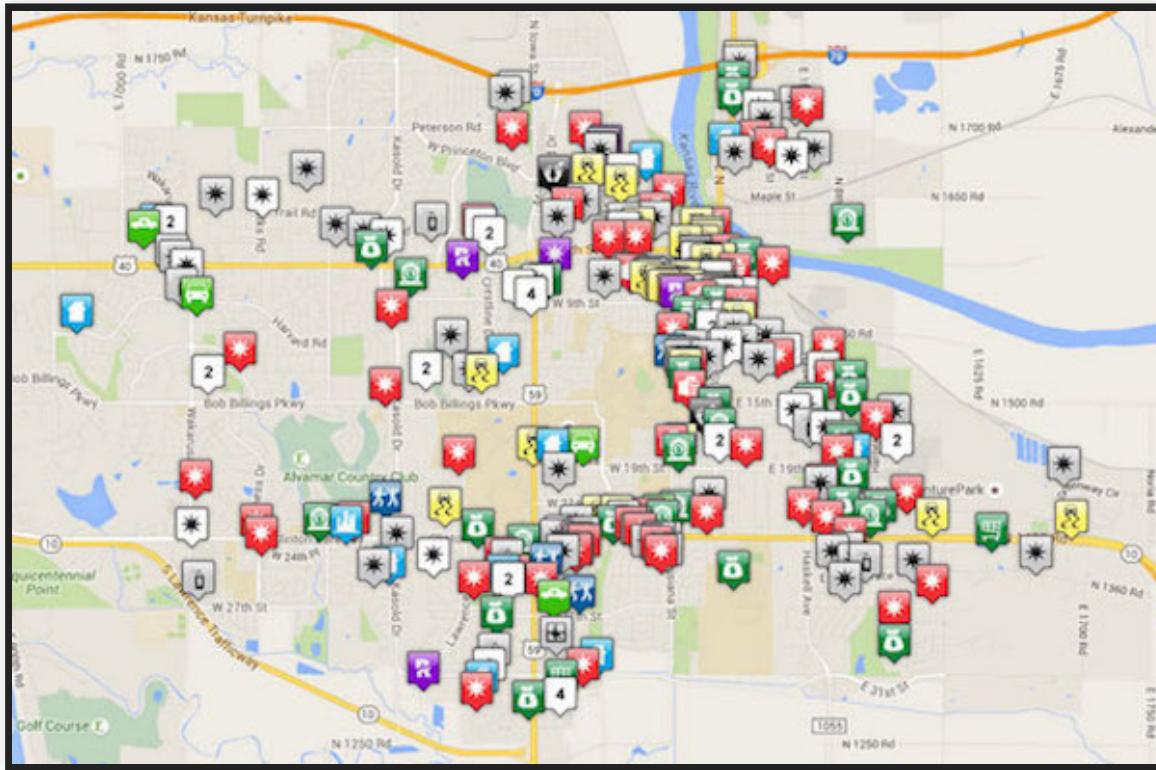
- Classification: Is this student likely to succeed?
- Features: GPA, SAT, race, gender, household income, city, etc.,
- **Discuss:** Historical bias? Skewed sample? Tainted examples? Limited features? Sample size disparity? Proxies?

# MASSIVE POTENTIAL DAMAGE



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

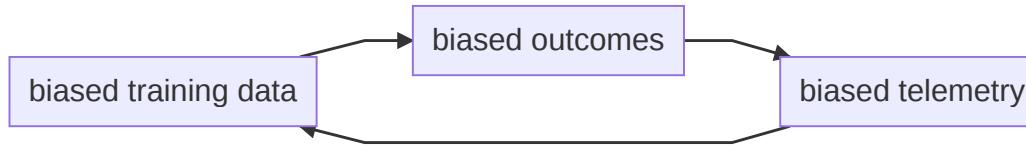
# EXAMPLE: PREDICTIVE POLICING



*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