

The background of the slide features a dramatic, large-scale industrial fire at night. The intense orange and yellow flames dominate the right side of the frame, casting a bright glow over dark structures that appear to be part of a factory or refinery. The fire is highly detailed, with visible smoke and sparks. In the foreground, there is a dark, semi-transparent area where the title text is placed.

# 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



[Read the full conversation on Twitter](#)

515

Reply

[Copy link to Tweet](#)

[Read 22 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

# Overview

Many interrelated issues:

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



*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](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...

Developers have substantial power in shaping products

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

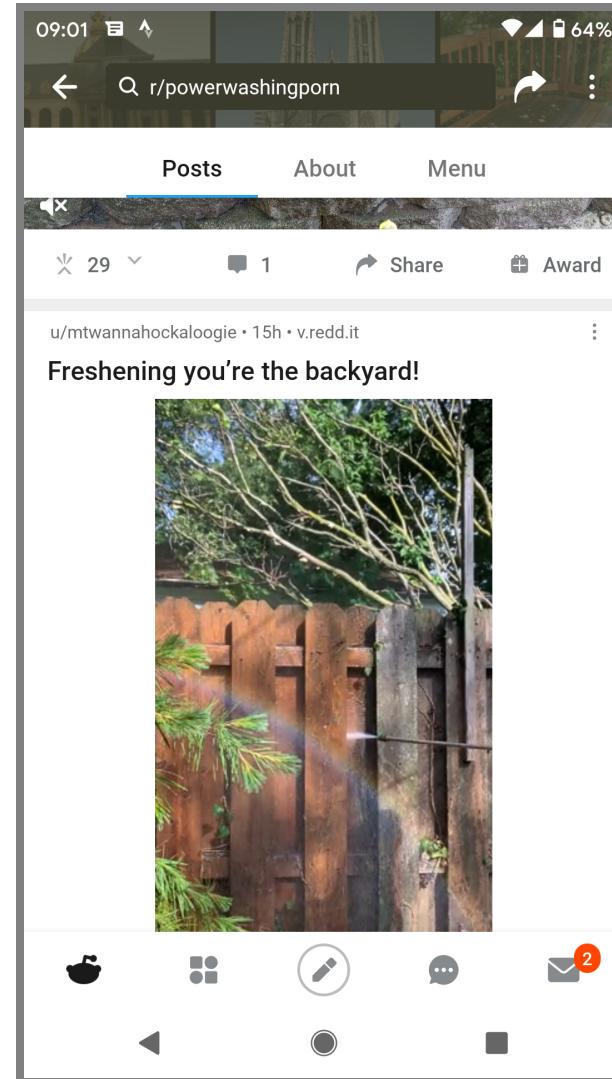


= *What is the (real) organizational objective of the company?*

# 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



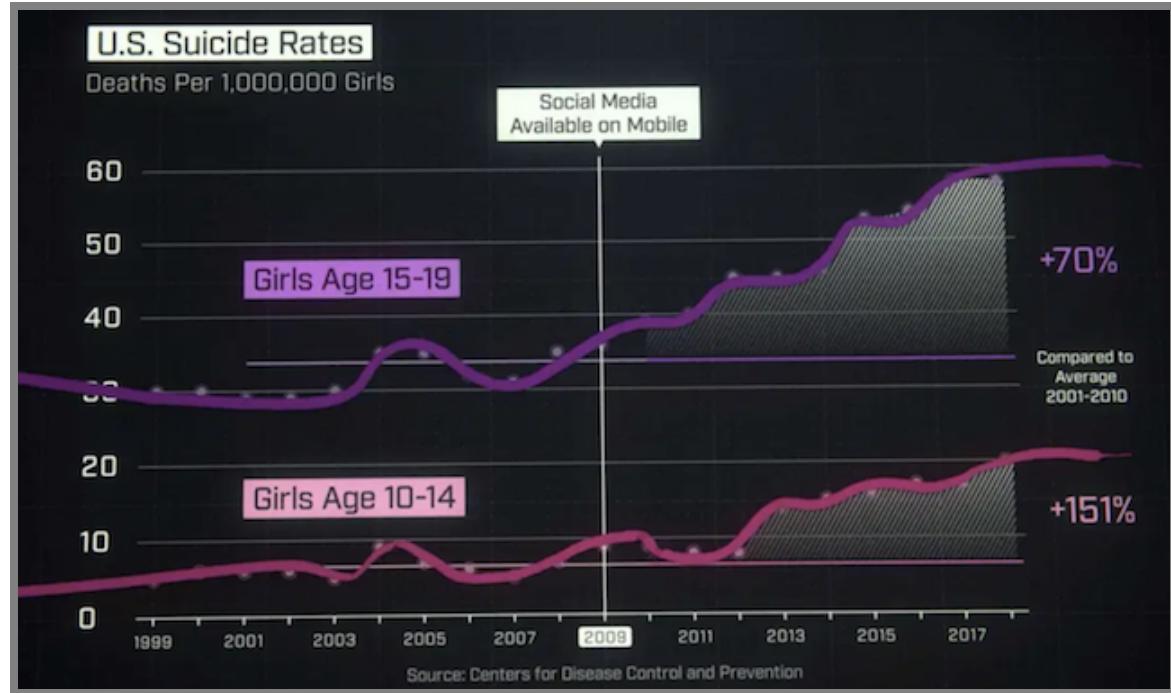
# Addiction



- 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.cdc.gov/motorvehiclesafety/Distracted\\_Driving/index.html](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



# Discrimination

 Tony "Abolish (Pol)ICE" Arcieri 🇺🇸  
@bascule

Trying a horrible experiment...

Which will the Twitter algorithm pick: Mitch McConnell or Barack Obama?



6:05 PM · Sep 19, 2020 · Twitter Web App

64K Retweets 16.5K Quote Tweets 198.3K Likes

# Who's to blame?

The screenshot shows a news article from The Independent. At the top left is the site's logo, 'INDEPENDENT'. To the right are links for 'Support us', 'Contribute' (in a red box), and 'Subscribe'. Below the header is a navigation bar with categories: NEWS, POLITICS, VOICES, SPORT, CULTURE, INDY/LIFE (in red), INDYBEST, VIDEO, DAILY EDITION, and CONVERSATIONS. The main title of the article is 'GOOGLE QUIETLY REMOVES 'DON'T BE EVIL' PREFACE FROM CODE OF CONDUCT', displayed in large, bold, black text on a white background. A subtitle below the main title reads 'Google employees resigned this month over the company's autonomous weapons project'. At the bottom of the article area, there is author information ('Anthony Cuthbertson | @ADCuthbertson | Monday 21 May 2018 12:21') and social sharing icons for bookmarking, Facebook, Twitter, and email.

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

*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.

## Speaker notes

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*

- Little legal consequences for causing negative impact (with some exceptions)
- 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 existed before, but they are being rapidly exacerbated by the widespread use of ML**

# 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

# Buzzword or real progress?

## Microsoft AI principles

We put our responsible AI principles into practice through the Office of Responsible AI (ORA) and the AI, Ethics, and Effects in Engineering and Research (Aether) Committee. The Aether Committee advises our leadership on the challenges and opportunities presented by AI innovations. ORA sets our rules and governance processes, working closely with teams across the company to enable the effort.

[Learn more about our approach >](#)

### Fairness

AI systems should treat all people fairly

[▷ Play video on fairness](#)

### Reliability & Safety

AI systems should perform reliably and safely

[▷ Play video on reliability](#)

### Privacy & Security

AI systems should be secure and respect privacy

[▷ Play video on privacy](#)

### Inclusiveness

AI systems should empower everyone and engage people

[▷ Play video on inclusiveness](#)

### Transparency

AI systems should be understandable

[▷ Play video on transparency](#)

### Accountability

People should be accountable for AI systems

[▷ Play video on accountability](#)



# 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

# 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](#))

= [https://en.wikipedia.org/wiki/Protected\\_group](https://en.wikipedia.org/wiki/Protected_group)

# 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

# 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.*

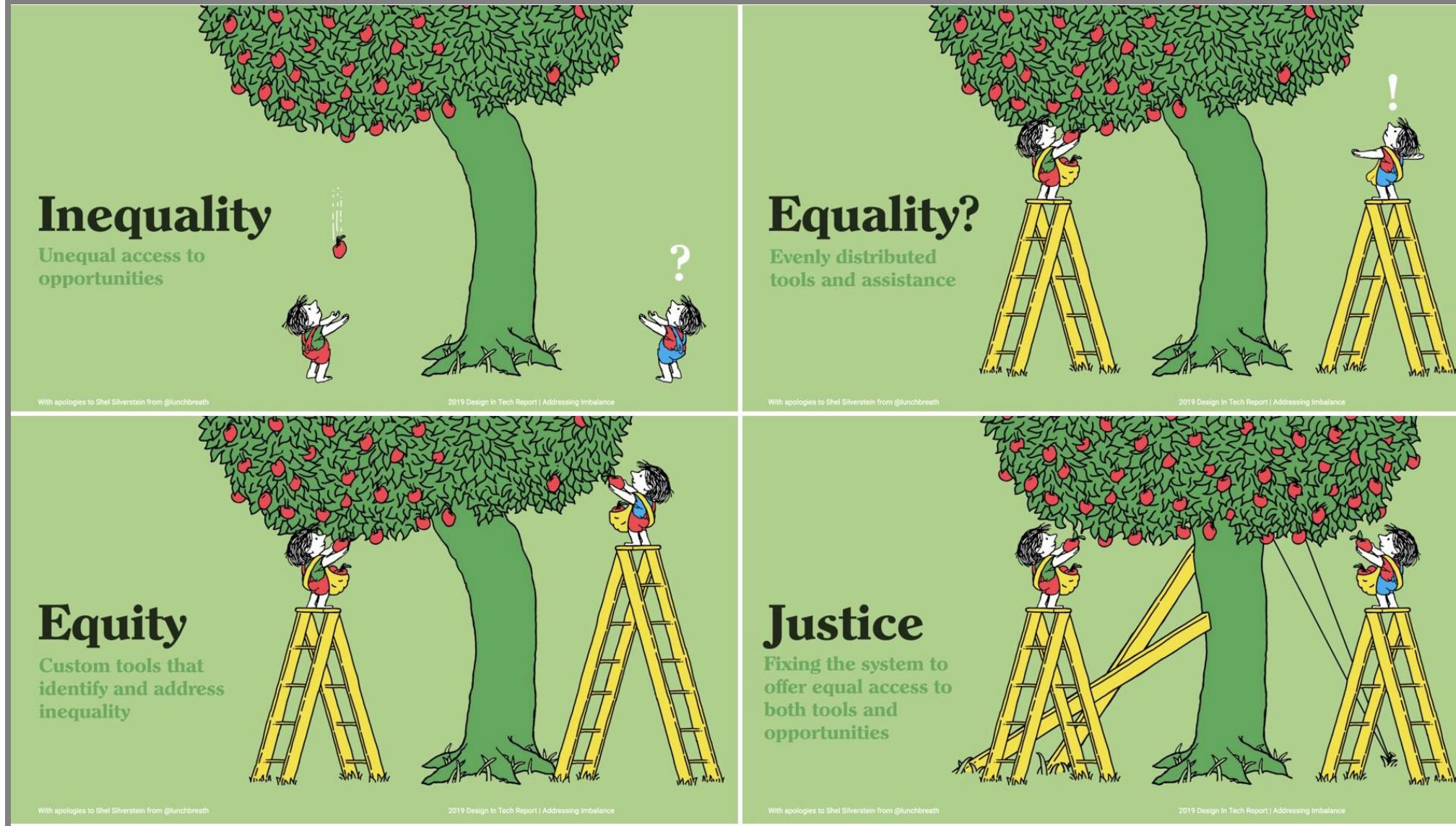
# Dividing a Pie?

- 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)*

# Preview: Equality vs Equity vs Justice



# Types of Harm on Society

**Harms of allocation:** Withhold opportunities or resources

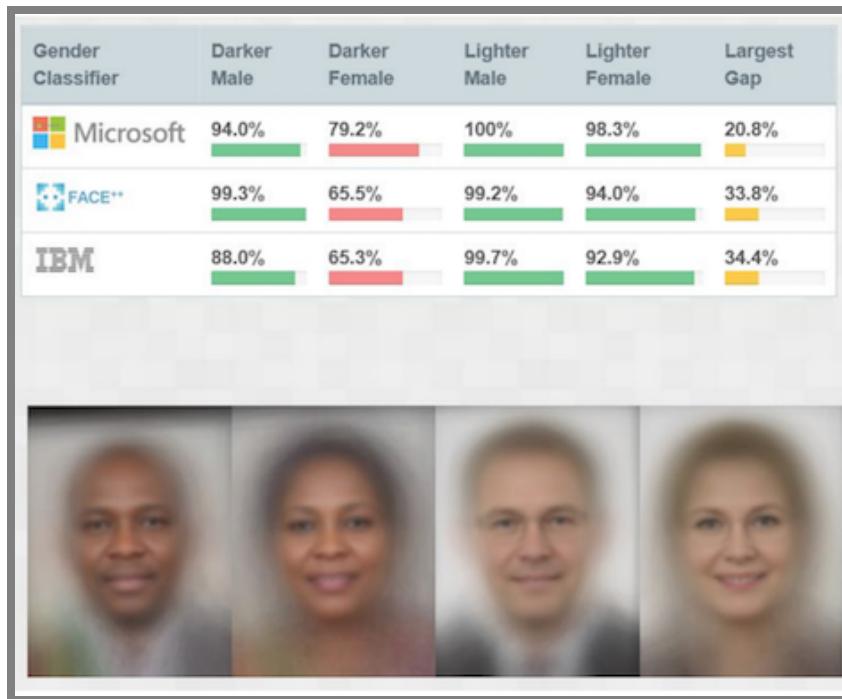
**Harms of representation:** Reinforce stereotypes, subordination along the lines of identity



Kate Crawford. "The Trouble With Bias", NeurIPS Keynote (2017).

# 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

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)



*Discrimination in Online Ad Delivery, Latanya Sweeney, 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.



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

# Not all discrimination is harmful



- Loan lending: Gender discrimination is illegal.
- Medical diagnosis: Gender-specific diagnosis may be desirable.
- The problem is *unjustified differentiation*; i.e., discriminating on factors that should not matter
- Discrimination is a **domain-specific** concept (i.e., world vs machine)

# 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 side-by-side screenshots of the Google Translate interface, highlighting a specific instance of gender bias in machine learning models.

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

- Source text: "He is a nurse  
She is a doctor"
- Target text: "O bir hemşire  
O bir doktor"
- Details: The English input is detected as "English - detected". The target language is "Turkish". The "Translate" button is blue.

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

- Source text: "O bir hemşire  
O bir doktor"
- Target text: "She is a nurse  
He is a doctor" (with a checkmark next to "He is a doctor")
- Details: The Turkish input is detected as "Turkish - detected". The target language is "English". The "Translate" button is blue.

In both cases, the model translates "he" to "O" and "she" to "She", but when translating back, it retains the original gendered terms ("hemşire" and "doktor") instead of switching them back to their English counterparts ("nurse" and "doctor"). This demonstrates how learned biases can persist across the translation pipeline.

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

# Where does the bias come from?

The image shows a screenshot of the Microsoft Translator web interface, which displays four separate translation windows arranged in a 2x2 grid. Each window has a 'Text' tab selected.

- Top Left Window:** English source, Turkish target. It contains the sentence "He is a nurse. She is a doctor." The word "She" is highlighted in blue, indicating it is the target word being analyzed.
- Top Right Window:** Turkish source, English target. It contains the sentence "O bir hemşire. O bir doktor." The word "O" is highlighted in blue.
- Bottom Left Window:** Turkish source, English target. It contains the sentence "O bir hemşire. O bir doktor." The word "O" is highlighted in blue.
- Bottom Right Window:** English source, Turkish target. It contains the sentence "She's a nurse. He's a doctor." The word "She's" is highlighted in blue.

The highlighting of "She" and "O" suggests that the system is identifying these words as potential gendered pronouns or names, while "He's" and "He" are not highlighted. This pattern illustrates how machine learning models can exhibit gender bias in their language processing.

# Sources of Bias

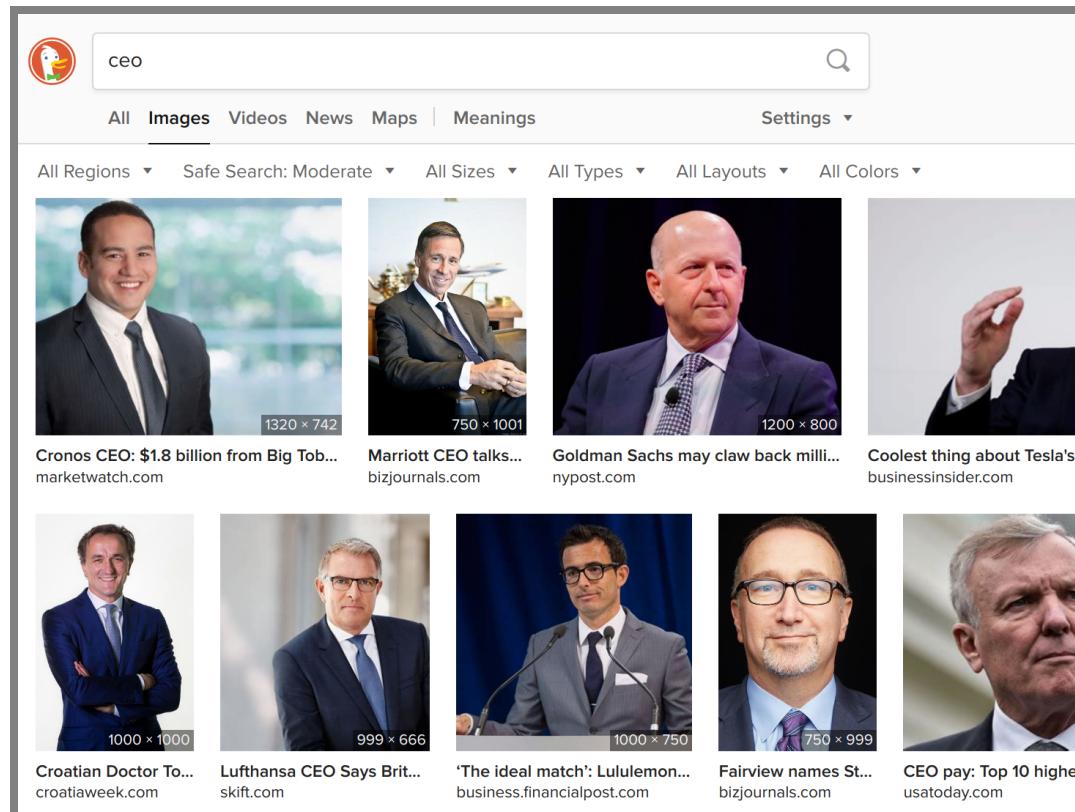
- Historial bias
- Tainted examples
- Skewed sample
- Limited features
- Sample size disparity
- Proxies



*Big Data's Disparate Impact*, Barocas & Selbst California Law Review (2016).

# Historical Bias

*Data reflects past biases, not intended outcomes*



*Should the algorithm reflect the 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."



# 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*

# Tainted Labels

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

The screenshot shows a news article from Vice Media. At the top left, there are three colored tabs: 'TECH' in pink, 'AMAZON' in blue, and 'ARTIFICIAL INTELLIGENCE' in red. The main title of the article is 'Amazon reportedly scraps internal AI recruiting tool that was biased against women'. Below the title is a subtitle: 'The secret program penalized applications that contained the word "women's"'. At the bottom left, it says 'By James Vincent | Oct 10, 2018, 7:09am EDT'.

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

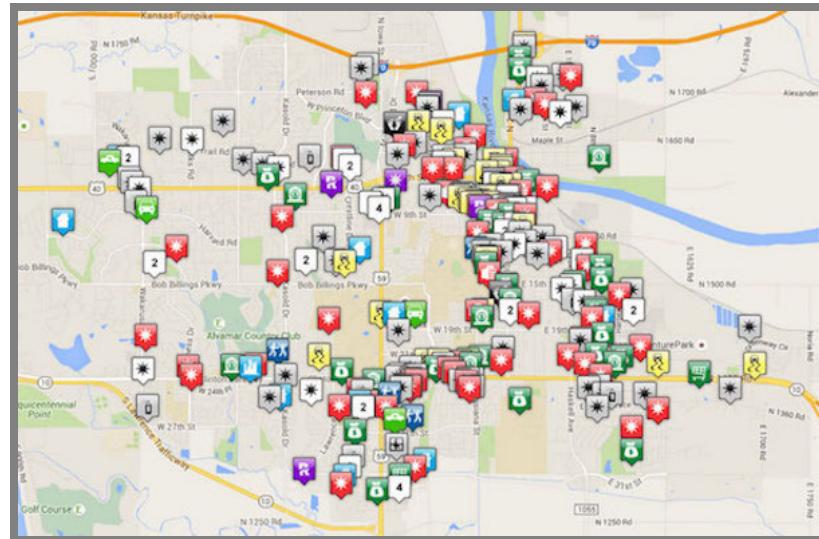
*The secret program penalized applications that contained the word “women’s”*

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

# 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?

Recall: Raw data is an oxymoron

# 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?

## Speaker notes

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.



# 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

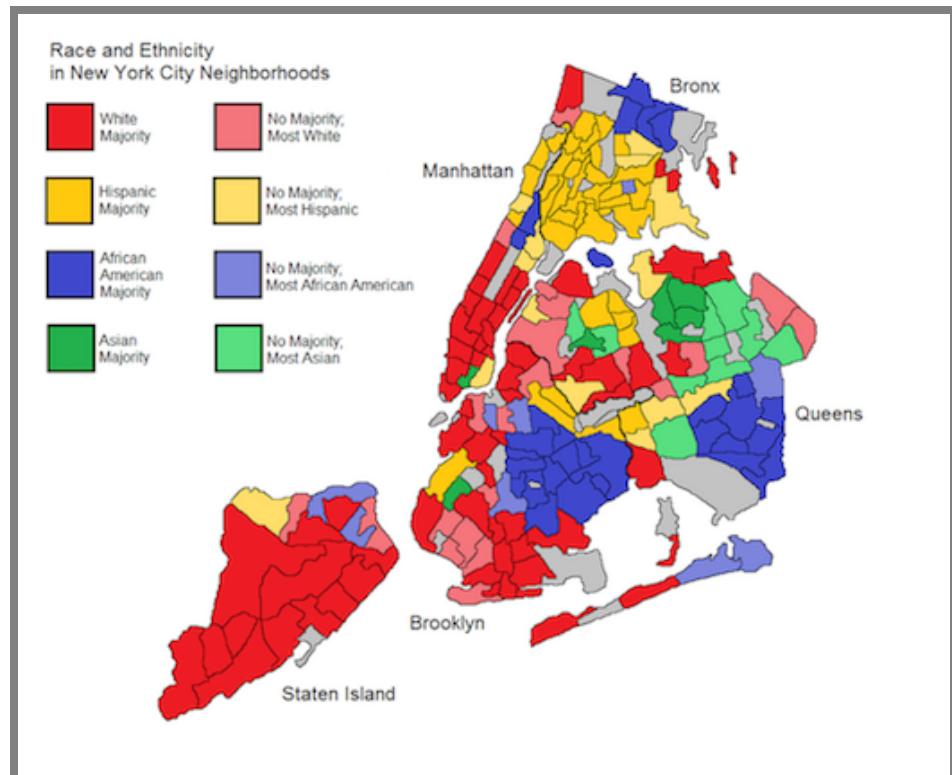
Without intervention:

- 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.

