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...

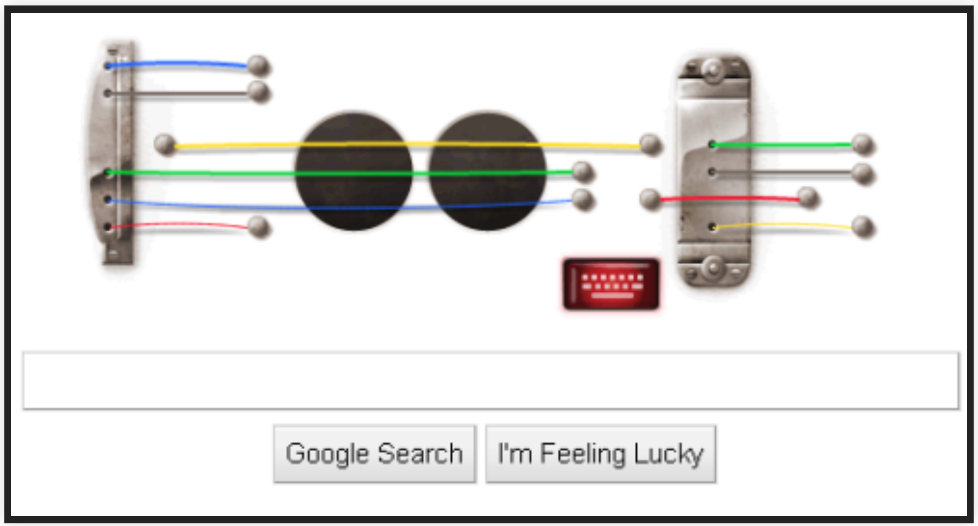
Some airlines may be using algorithms to split up families during flights

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

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

SHARE

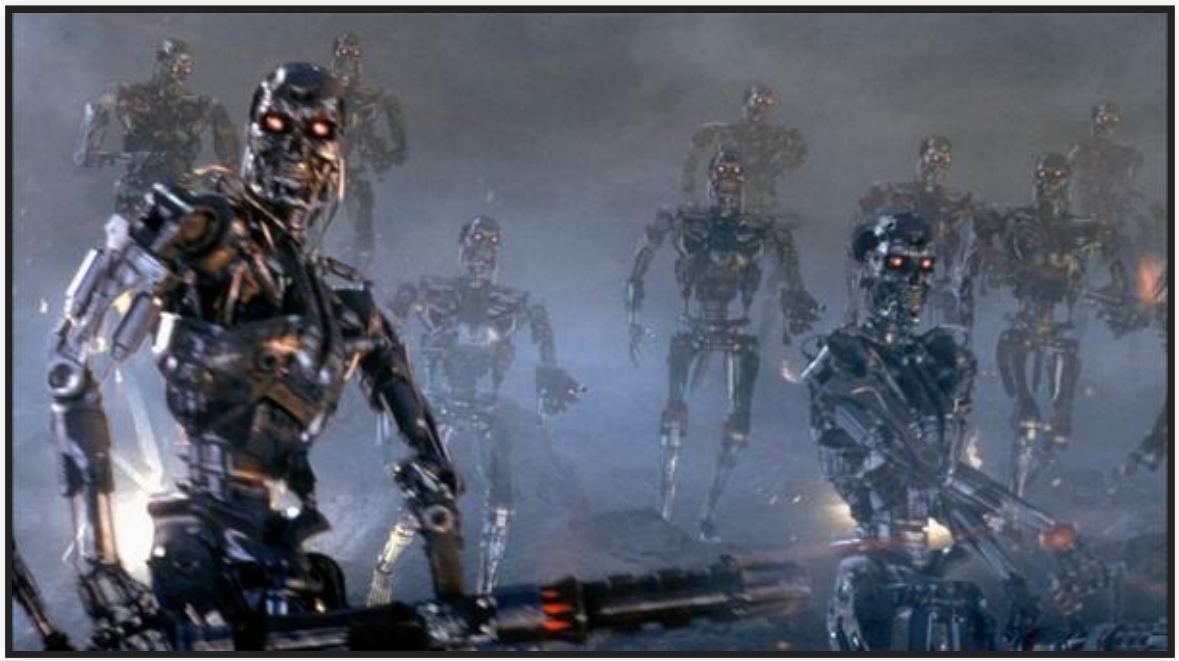
THE IMPLICATIONS OF OUR CHOICES



“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



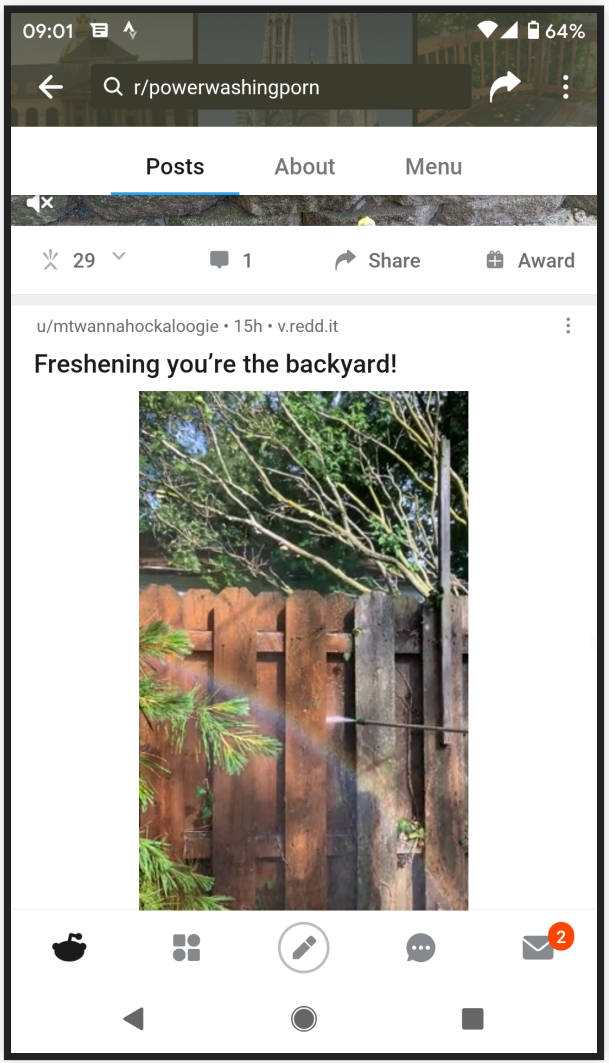
SAFETY


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
Kieran
@DivemasterK
netatmo Are your servers down ? I can't nnect to my app to turn on heating !!
11.18, 21:02 from Wicklow, Ireland
@netatmo hi my manual ovrride on my thermostat is not working and when i try using the app it comes up with an error with servers down. Can i overide at boiler end?
22.11.18, 20:58
Andy Mc
@ITakeSugar
Replying to @levisleedaniel and @netatmo
Is there a way to control the boiler even if the servers are down, it's freezing at the moment
22.11.18, 20:38
James Brown
@netatmo
same issue. Can't control heat
cannot login to netatmo.com
to control from there. What is
@netatmo ?
8:15 PM · Nov 22, 2018
2.2K
1.6K people ...

SAFETY

Tweet

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

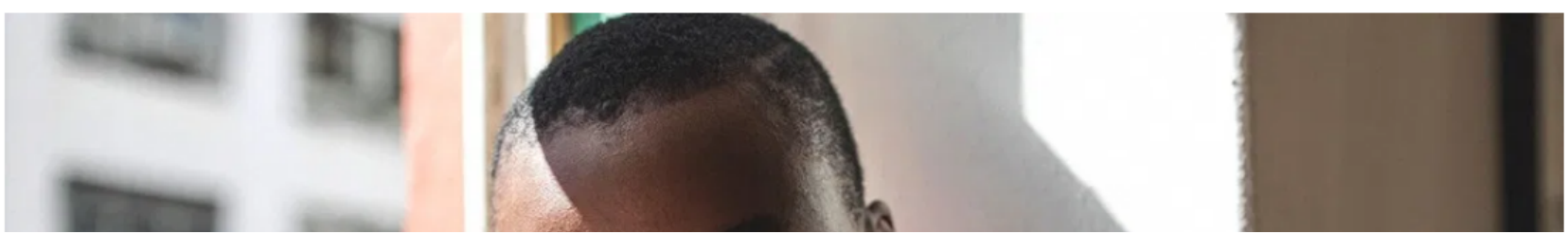
MENTAL HEALTH

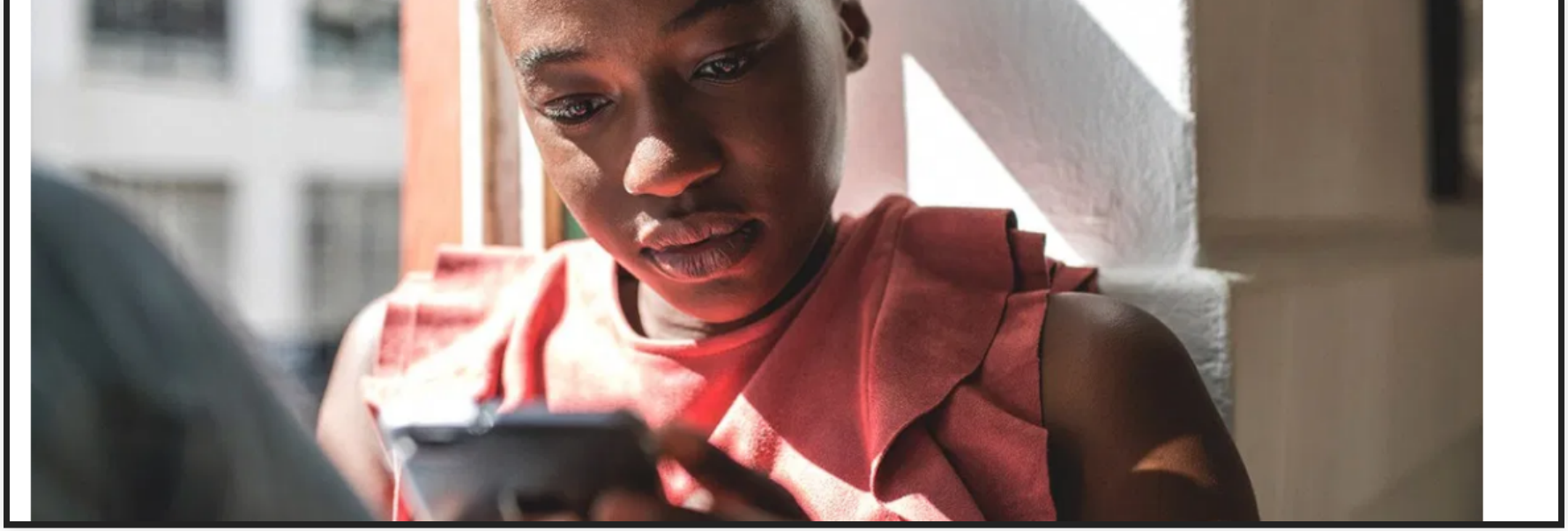
HEALTH NEWS ✓ Fact Checked

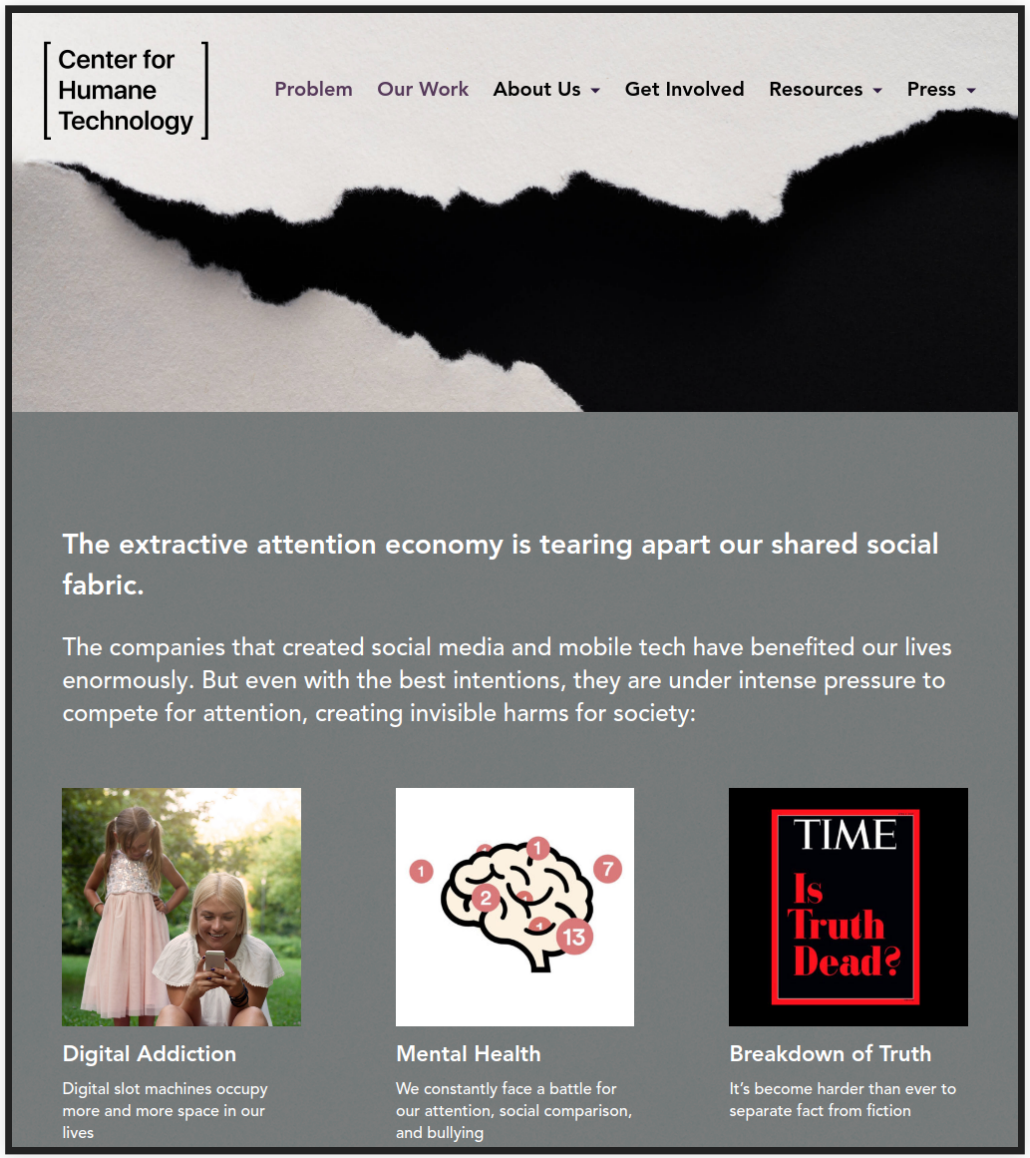
The FOMO Is Real: How Social Media Increases Depression and Loneliness

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:
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
TIME
Is Truth Dead?
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.

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

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

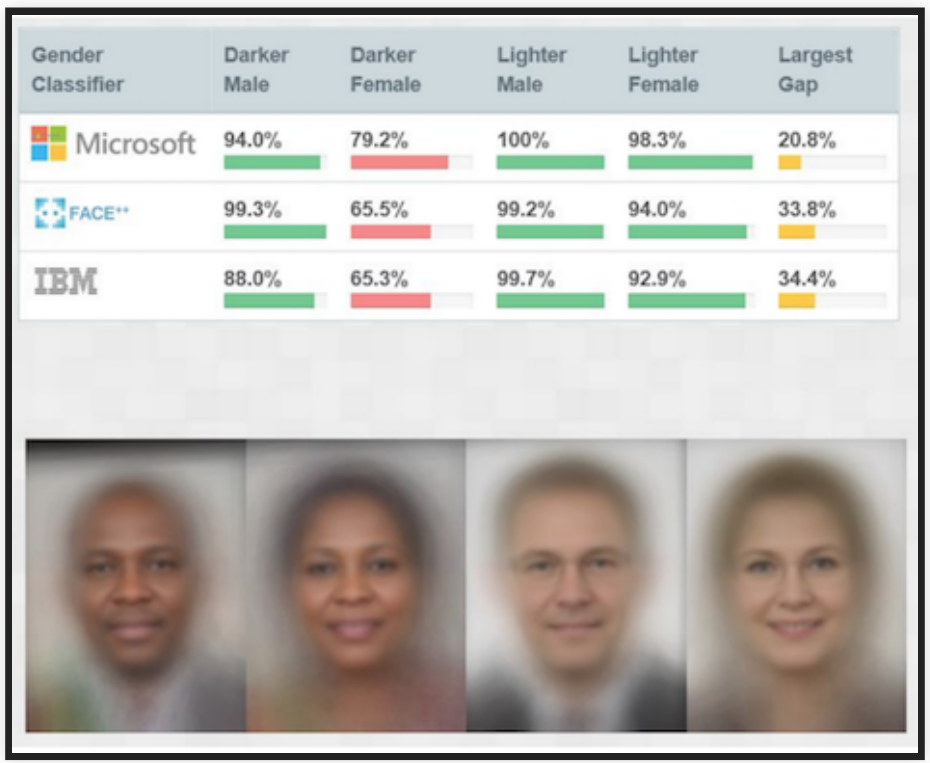
Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

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

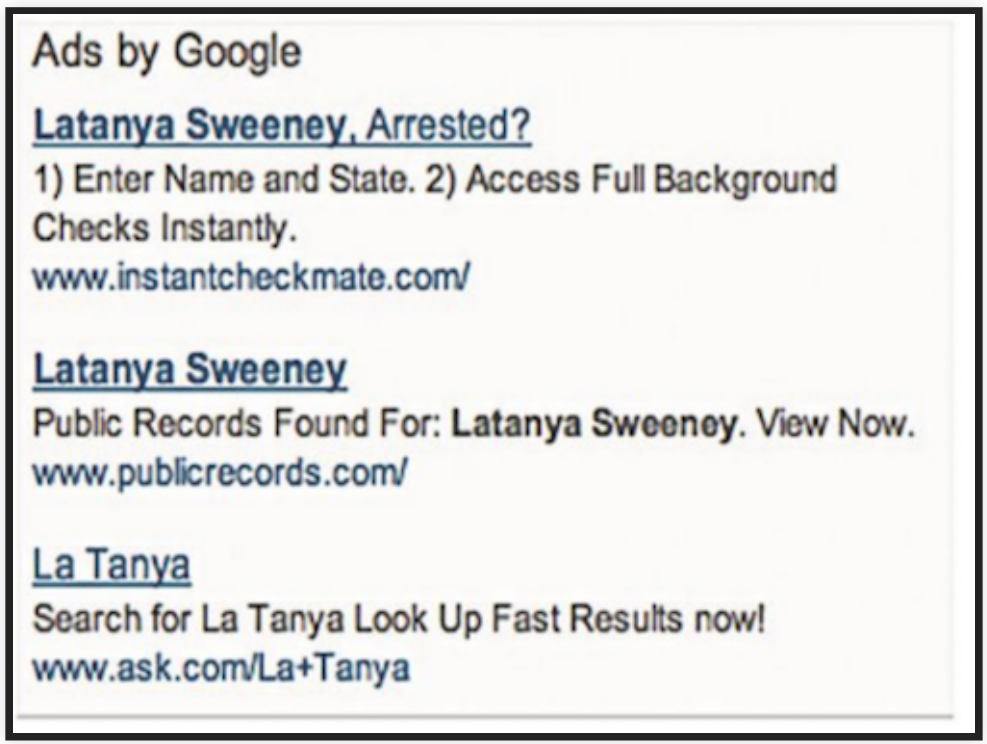


Other examples?

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

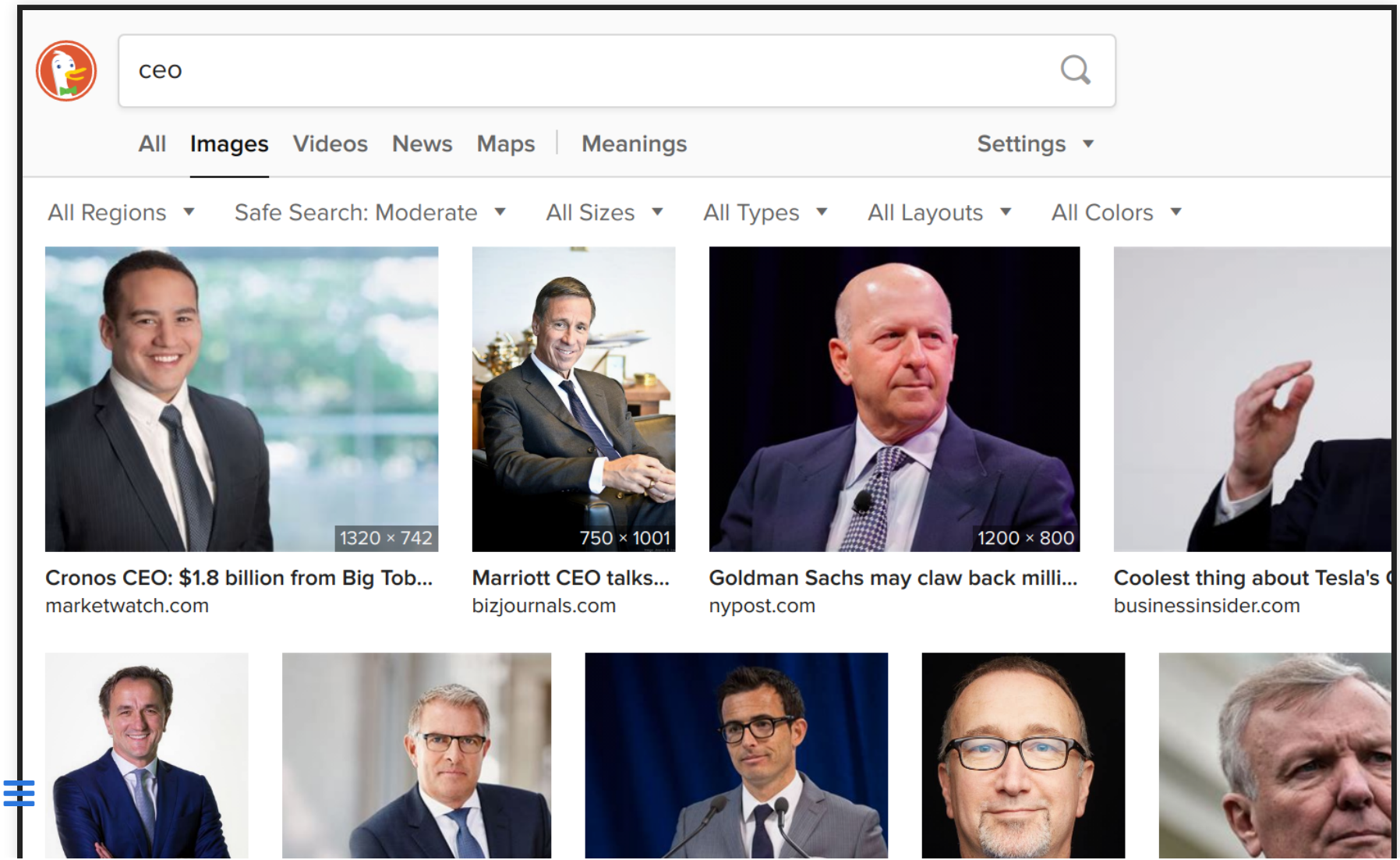
- Tainted examples / historical bias
- Skewed sample
- Limited features
- Sample size disparity
- Proxies

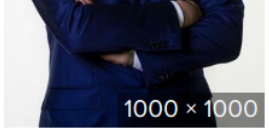
Barocas, Solon, and Andrew D. Selbst. "Big data's disparate impact." Calif. L. Rev. 104 (2016): 671.

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

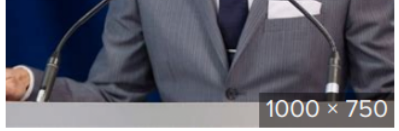
Data reflects past biases, not intended outcomes



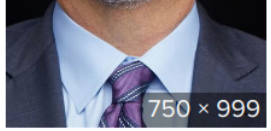


Croatian Doctor To...
croatiaweek.com

Lufthansa CEO Says Brit...
skift.com



'The ideal match': Lululemon...
business.financialpost.com



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"

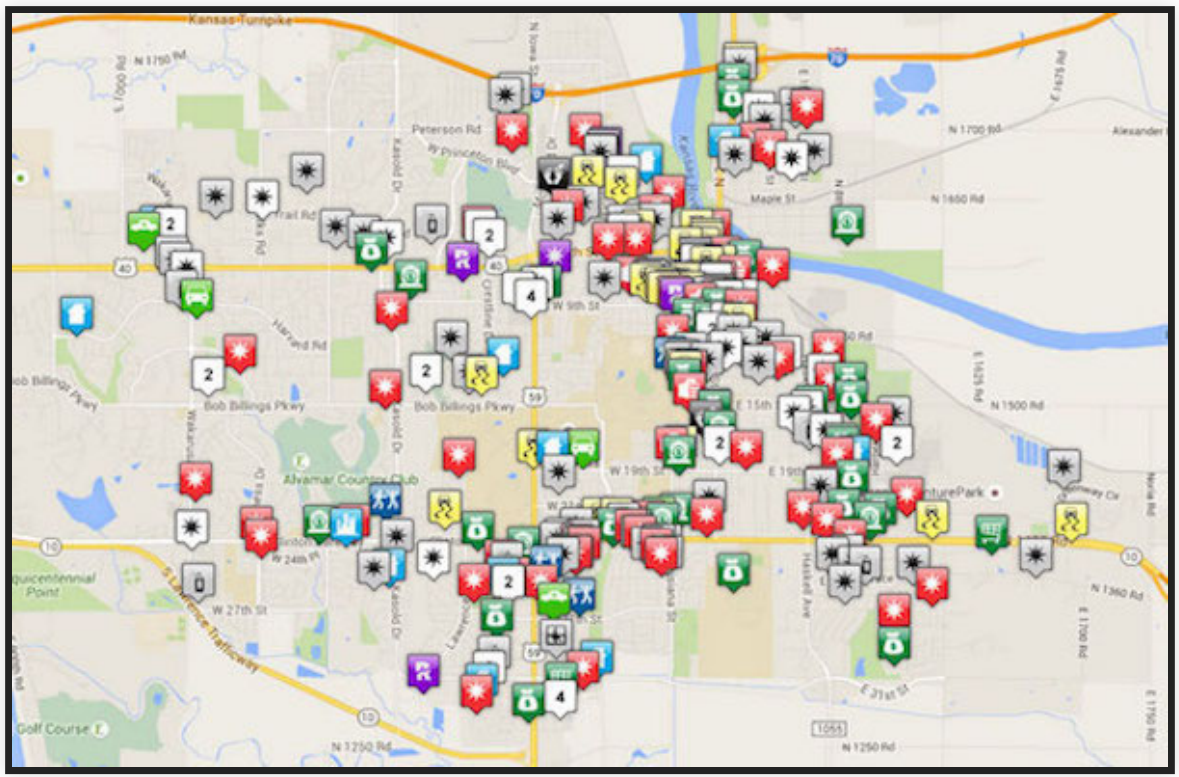
By James Vincent | Oct 10, 2018, 7:09am EDT

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

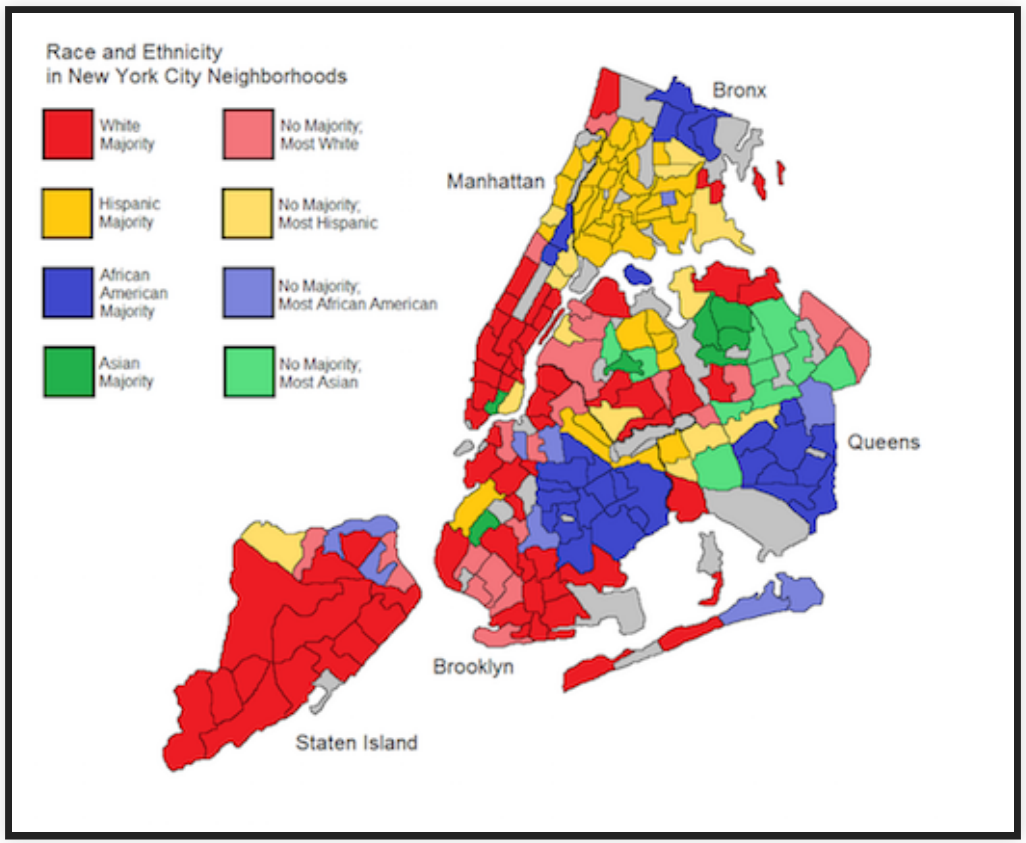
Speaker notes

- 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

Tweet

PROXIES

Features correlate with protected attributes



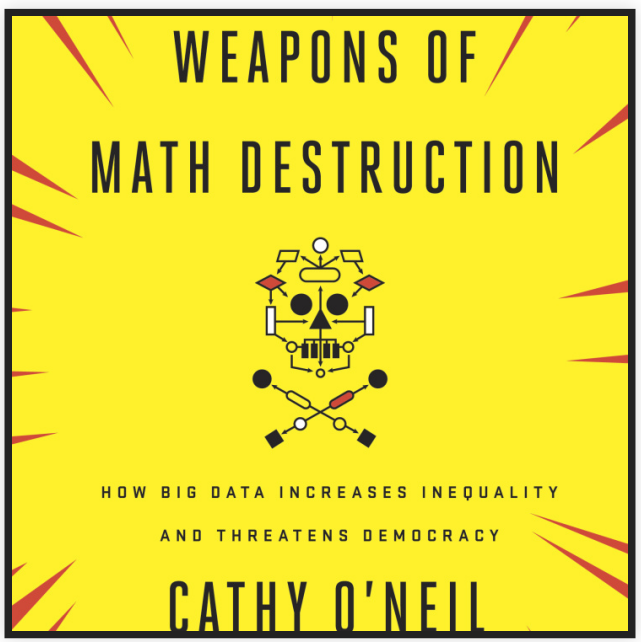
Speaker notes

- 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

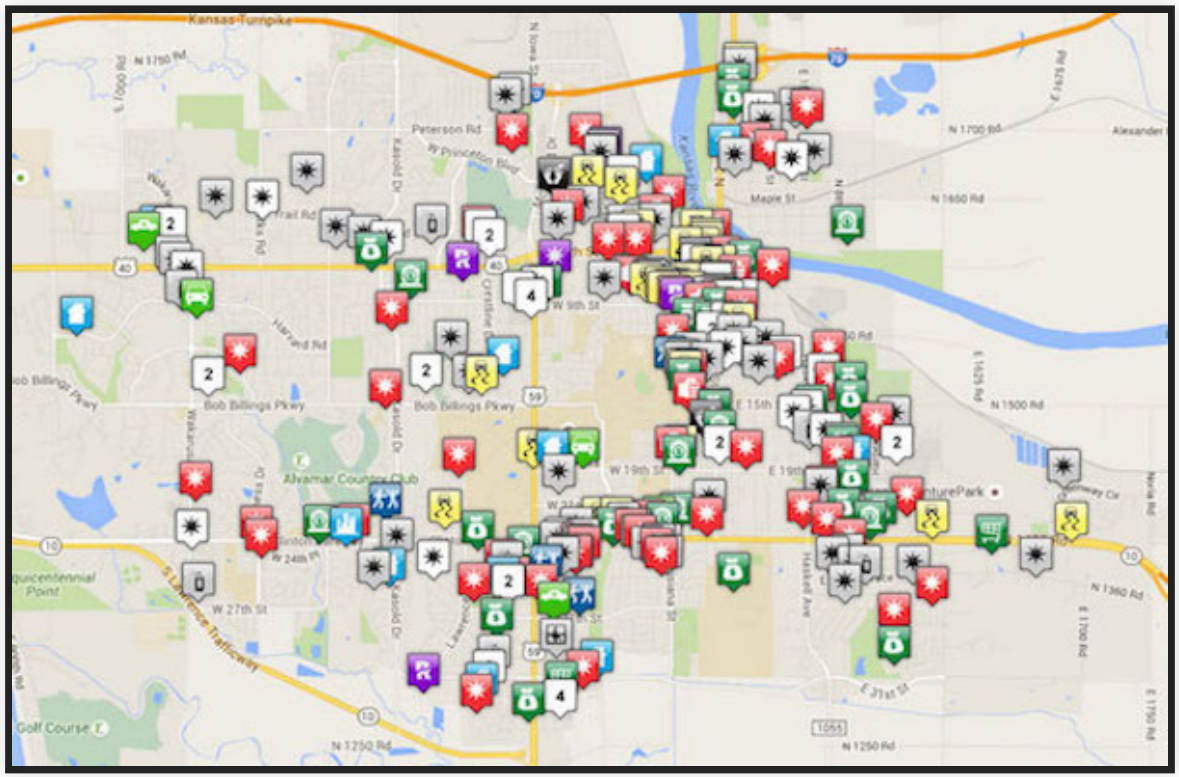
- Classification: Is this student likely to succeed?
- Features: GPA, SAT, race, gender, household income, city, etc.,
- **Discuss:** Historical bias? Skewed sample? Tainted examples? Limited features? Sample size disparity? Proxies?

MASSIVE POTENTIAL DAMAGE



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

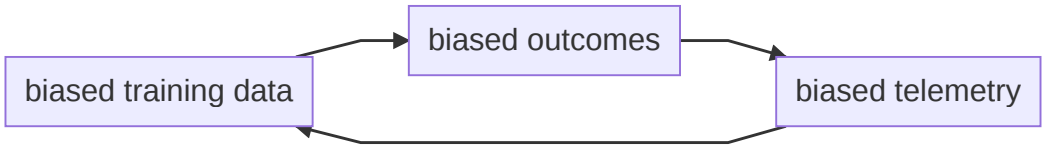
EXAMPLE: PREDICTIVE POLICING



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.

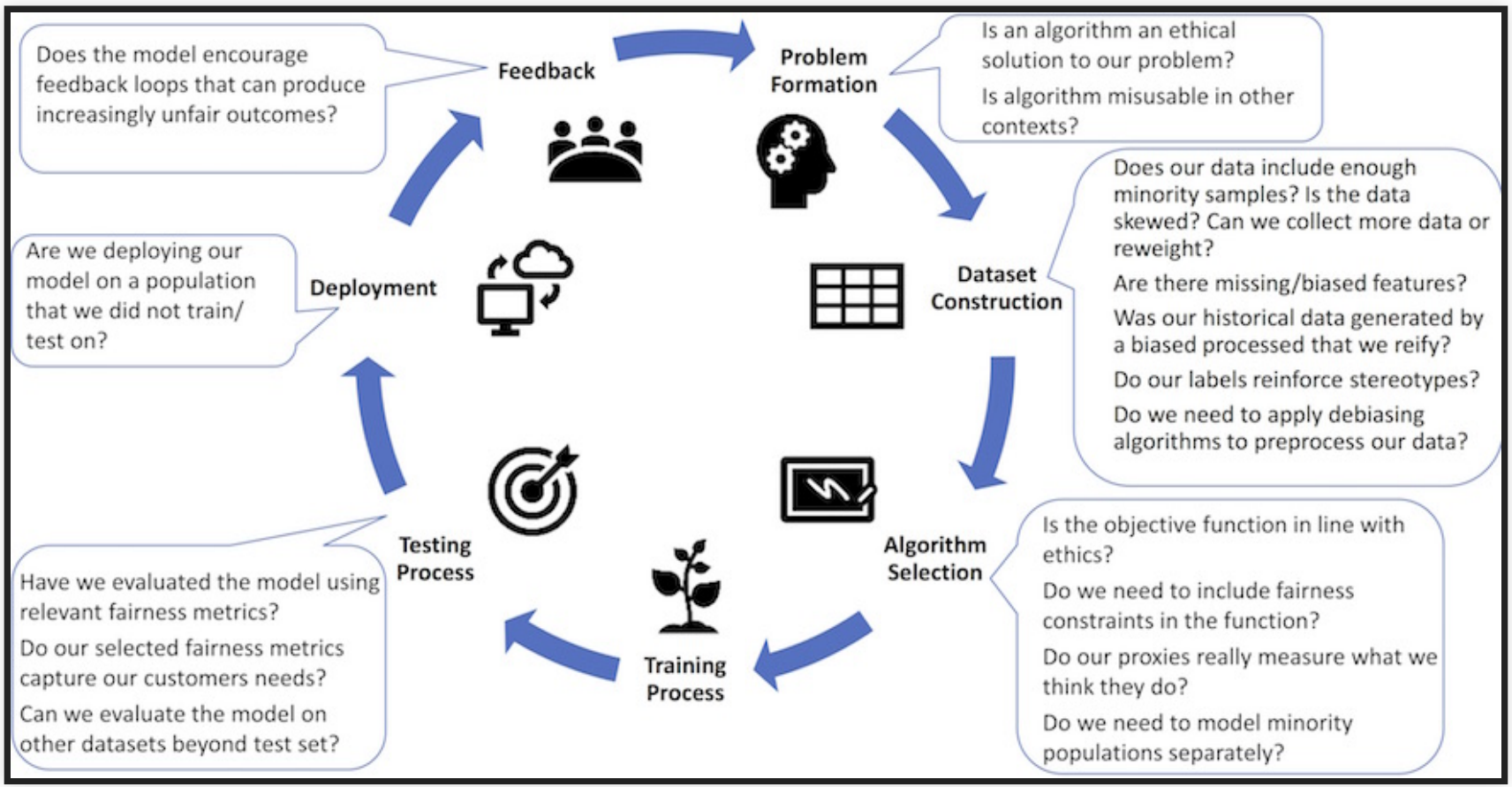
"WEAPONS OF MATH DESTRUCTION"

- 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.

BUILDING FAIR ML SYSTEMS

FAIRNESS MUST BE CONSIDERED THROUGHOUT THE ML LIFECYCLE!



Bennett et al., Fairness-aware Machine Learning, WSDM Tutorial (2019).

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

THE ROLE OF SOFTWARE ENGINEERS

- Whole system perspective
- Requirements engineering, identifying stakeholders
- Tradeoff decisions among conflicting goals
- Interaction and interface design
- Infrastructure for evaluating model quality and fairness offline and in production
- Monitoring
- System-wide mitigations (in model and beyond model)

BEST PRACTICES: TASK DEFINITION

- Clearly define the task & model's intended effects
- Try to identify and document unintended effects & biases
- Clearly define any fairness requirements
- *Involve diverse stakeholders & multiple perspectives*
- Refine the task definition & be willing to abort

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)

BEST PRACTICES: CHOOSING A DATA SOURCE

- Think critically before collecting any data
- Check for biases in data source selection process
- Try to identify societal biases present in data source
- Check for biases in cultural context of data source
- Check that data source matches deployment context
- Check for biases in
    - technology used to collect the data
    - humans involved in collecting data
    - sampling strategy
- *Ensure sufficient representation of subpopulations*
- Check that collection process itself is fair & ethical

How can we achieve fairness without putting a tax on already disadvantaged populations?

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)

BEST PRACTICES: LABELING AND PREPROCESSING

- Check for biases introduced by
  - discarding data
  - bucketing values
  - preprocessing software
  - labeling/annotation software
  - human labelers
- Data/concept drift?

Auditing? Measuring bias?

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)

BEST PRACTICES: MODEL DEFINITION AND TRAINING

- Clearly define all assumptions about model
- Try to identify biases present in assumptions
- Check whether model structure introduces biases
- Check objective function for unintended effects
- Consider including “fairness” in objective function

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)

BEST PRACTICES: TESTING & DEPLOYMENT

- Check that test data matches deployment context
- Ensure test data has sufficient representation
- Continue to involve diverse stakeholders
- Revisit all fairness requirements
- Use metrics to check that requirements are met

- Continually monitor
  - match between training data, test data, and instances you encounter in deployment
  - fairness metrics
  - population shifts
  - user reports & user complaints
- Invite diverse stakeholders to audit system for biases

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)

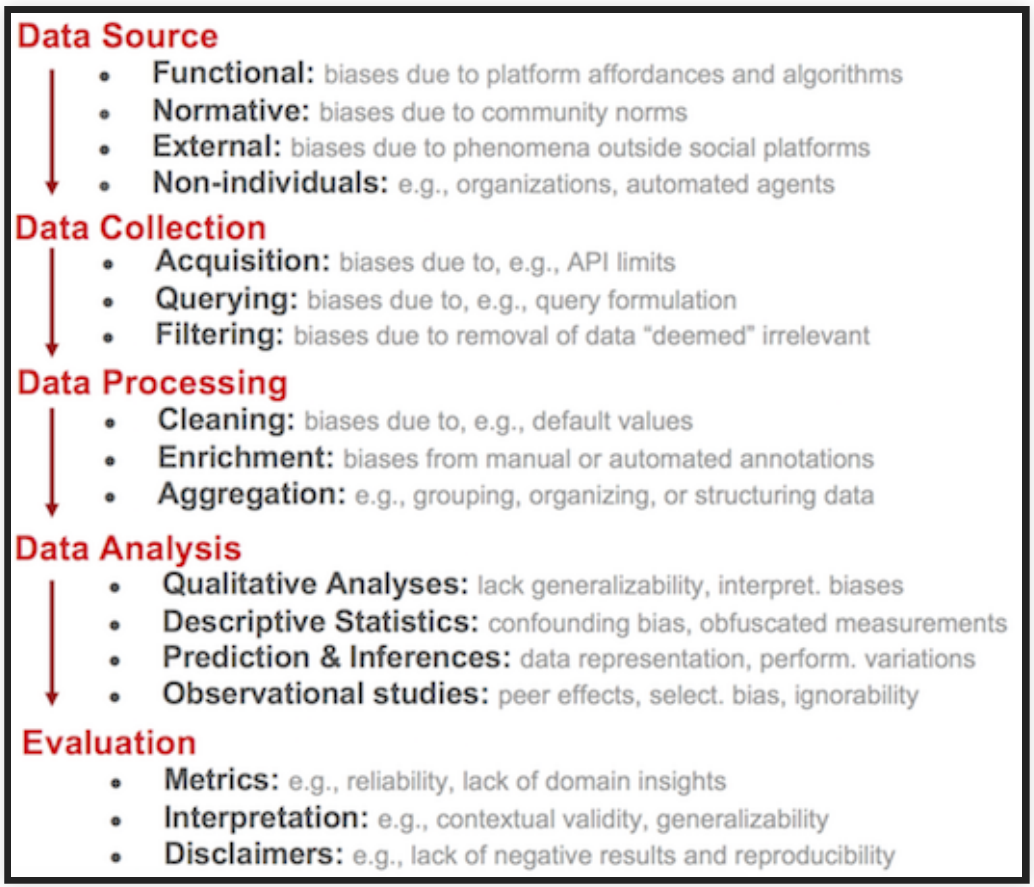
DATASET CONSTRUCTION FOR FAIRNESS

FLEXIBILITY IN DATA COLLECTION

- Data science education often assumes data as given
- In industry most have control over data collection and curation (65%)
- Most address fairness issues by collecting more data (73%)

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)

Bias can be introduced at any stage of the data pipeline



Bennett et al., Fairness-aware Machine Learning, WSDM Tutorial (2019).

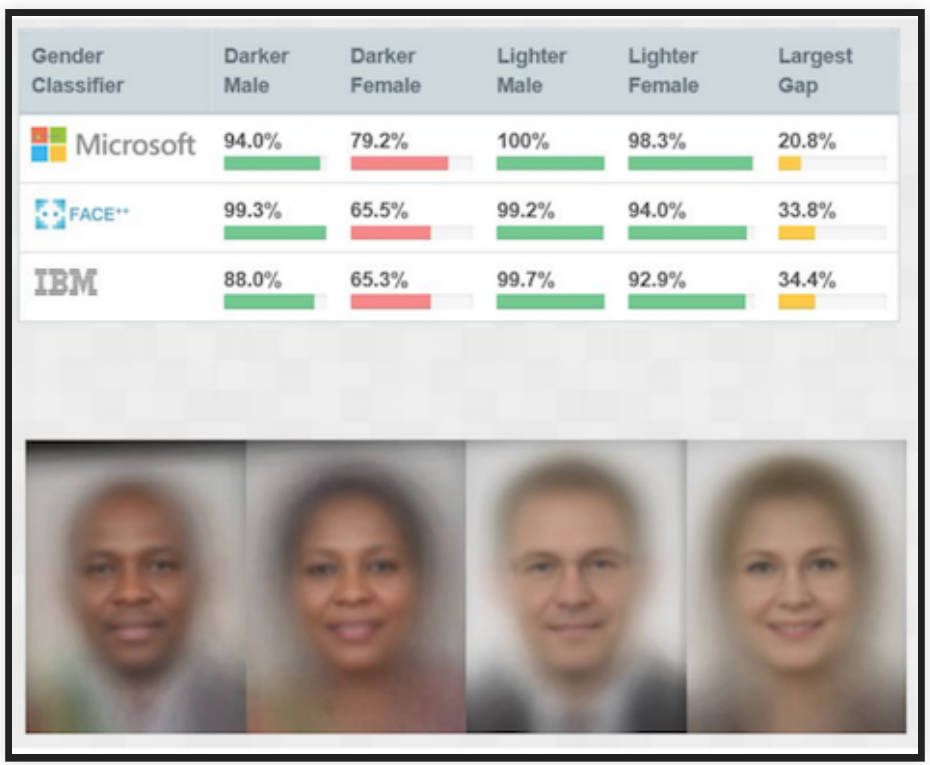
TYPES OF DATA BIAS

- **Population bias**
- **Behavioral bias**
- Content production bias
- Linking bias
- Temporal bias

Olteanu et al., Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, Olteanu et al., Frontiers in Big Data (2019).

POPULATION BIAS

- Differences in demographics between a dataset vs a target population
- Example: Does the Twitter demographics represent the general population?
- In many tasks, datasets should match the target population
- But some tasks require equal representation for fairness



BEHAVIORAL BIAS

- Differences in user behavior across platforms or social contexts

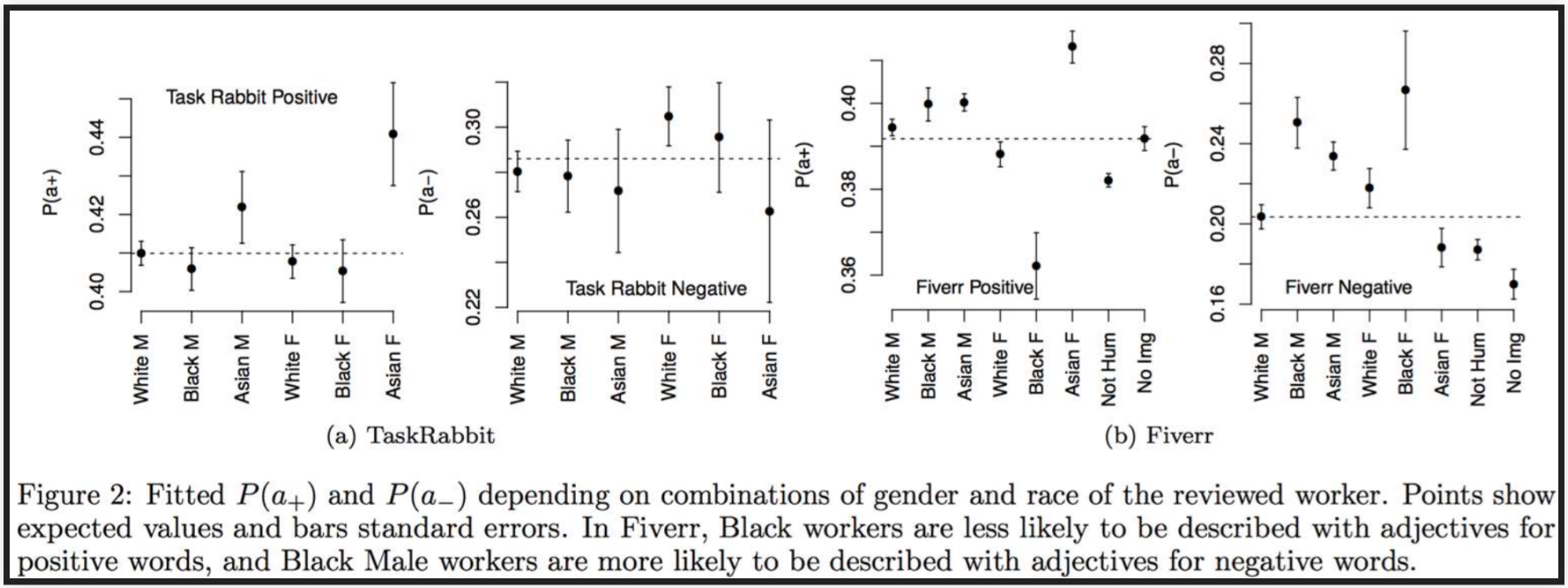


Figure 2: Fitted $P(a_+)$ and $P(a_-)$ depending on combinations of gender and race of the reviewed worker. Points show expected values and bars standard errors. In Fiverr, Black workers are less likely to be described with adjectives for positive words, and Black Male workers are more likely to be described with adjectives for negative words.

Example: Freelancing platforms (Fiverr vs TaskRabbit): Bias against certain minority groups on different platforms

Bias in Online Freelance Marketplaces, Hannak et al., CSCW (2017).

FAIRENESS-AWARE DATA COLLECTION

- Address population bias
  - Does the dataset reflect the demographics in the target population?
- Address under- & over-representation issues
  - Ensure sufficient amount of data for all groups to avoid being treated as "outliers" by ML
  - But also avoid over-representation of certain groups (e.g., remove historical data)
- Data augmentation: Synthesize data for minority groups
  - Observed: "He is a doctor" -> synthesize "She is a doctor"
- Fairness-aware active learning
  - Collect more data for groups with highest error rates

Bennett et al., Fairness-aware Machine Learning, WSDM Tutorial (2019).

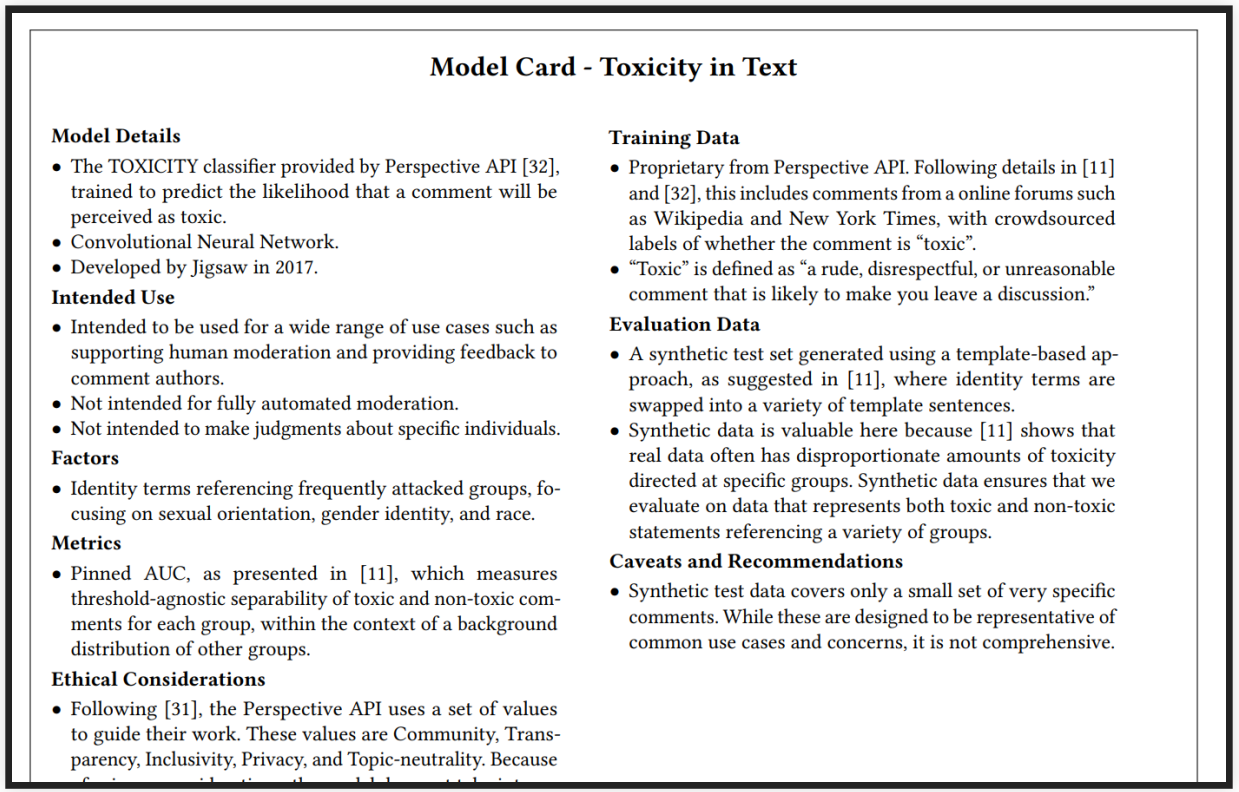
DATA SHEETS

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%

- A process for documenting datasets
- Based on common practice in the electronics industry, medicine
- Purpose, provenance, creation, composition, distribution: Does the dataset relate to people? Does the dataset identify any subpopulations?

Datasheets for Dataset, Gebru et al., (2019).

MODEL CARDS



Model Card - Toxicity in Text

Model Details

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

Intended Use

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

Factors

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

Metrics

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

Ethical Considerations

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because

Training Data

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

Evaluation Data

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

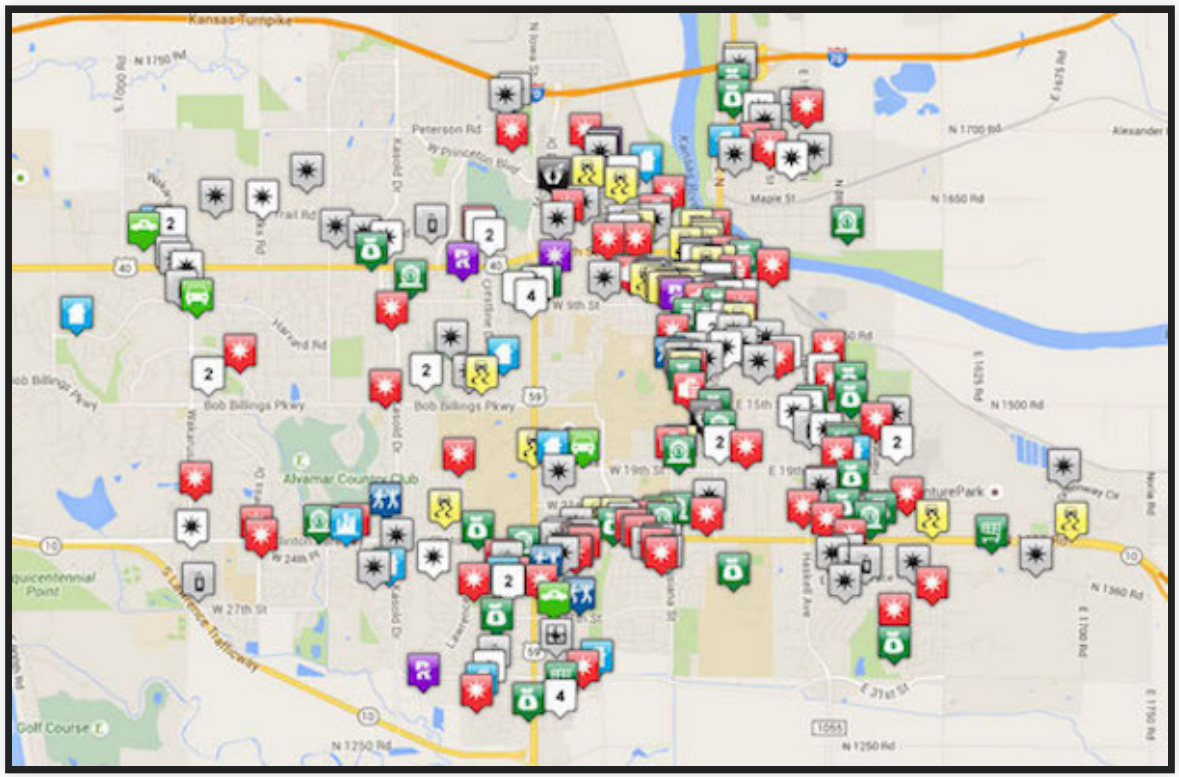
Caveats and Recommendations

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

see also https://modelcards.withgoogle.com/about

Mitchell, Margaret, et al. "Model cards for model reporting." In Proceedings of the Conference on fairness, accountability, and transparency, pp. 220-229. 2019.

EXERCISE: CRIME MAP



How can we modify an existing dataset or change the data collection process to reduce the effects the feedback loop?

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
- Addressing fairness requirements engineering and throughout the entire ML pipeline
- Data bias & data collection for fairness

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