

Statistical aspects of algorithmic fairness

Joshua Loftus

New York University

10/14/2019

Joint work with

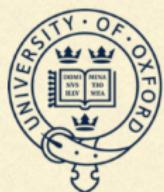
Matt Kusner^{1→3,4}

Chris Russell^{2,4}

Ricardo Silva^{3,4}



¹Oxford, NYU, ²Surrey, ³UCL, ⁴Alan Turing Institute



The
Alan Turing
Institute

HIGH-CRIME AREA MONITORING

PROBLEM

- Parking lot thefts
- Frequent drug trafficking
- Looting threats
- Dark alleys
- Multiple locations
- Limited force
- Suspect identification
- Evidence gathering

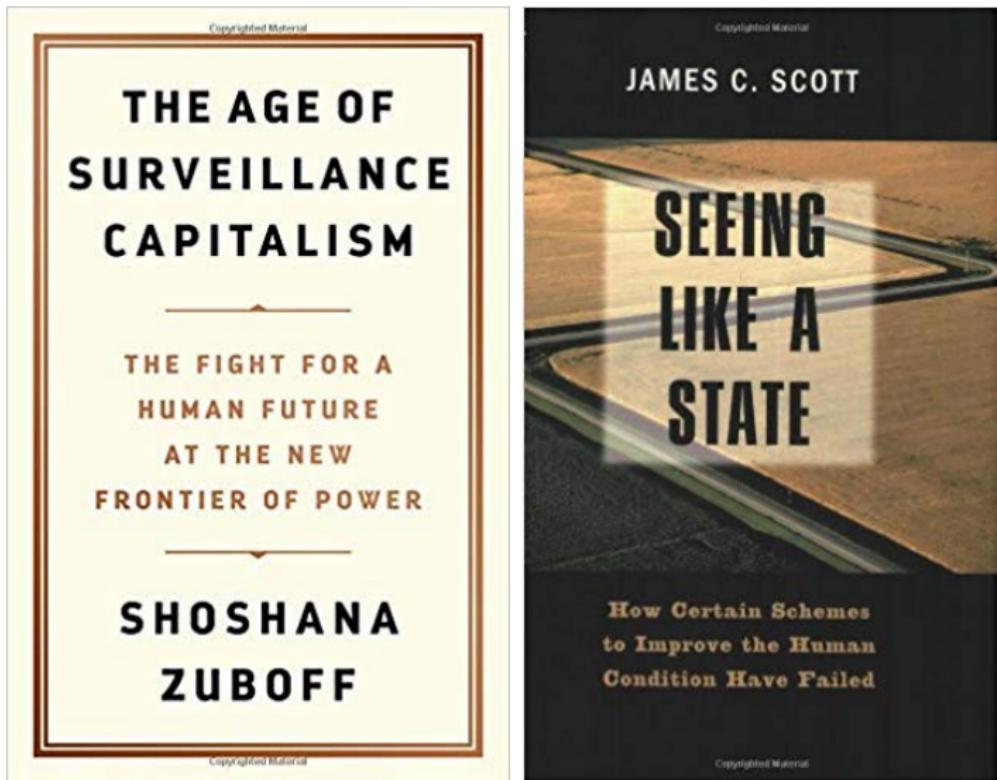
SOLUTION

- Visible deterrent
- Man or unmanned
- Towable portability & rapid deployment
- Single officer
- Generator power
- Thermal & visible surveillance cameras and recording
- Extra spotlights



Source: flir.com “SkyWatch”

Maybe closer to the opposite of fair by default...



Technological crisis

"New products and services, including those that incorporate or utilize artificial intelligence and machine learning, can raise new or exacerbate existing ethical, technological, legal, and other challenges, which may negatively affect our brands and demand for our products and services and adversely affect our revenues and operating results"

TOM SIMONITE BUSINESS 02.11.19 07:00 AM

GOOGLE AND MICROSOFT WARN THAT AI MAY DO DUMB THINGS



© ALYSSA FOOTE

Source: WIRED article, Feb. 2019

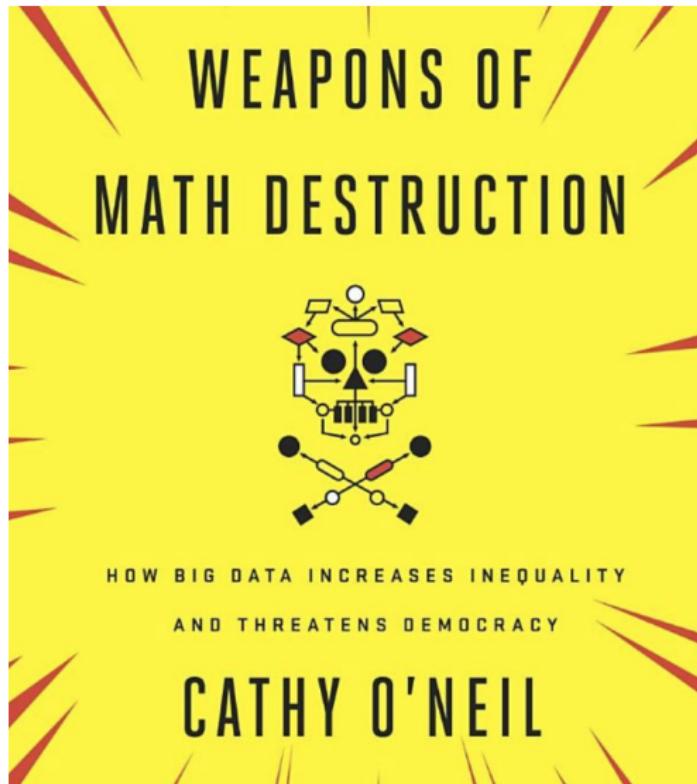
Outline

- ▶ **Algorithmic fairness** background
- ▶ “**Impossibility**” of fairness
- ▶ **Causality** as a path forward

Much of the technical material here can be found in a survey (preprint) with my collaborators, Causal Reasoning for Algorithmic Fairness (2018), and references therein

Selective history of fairness in machine learning

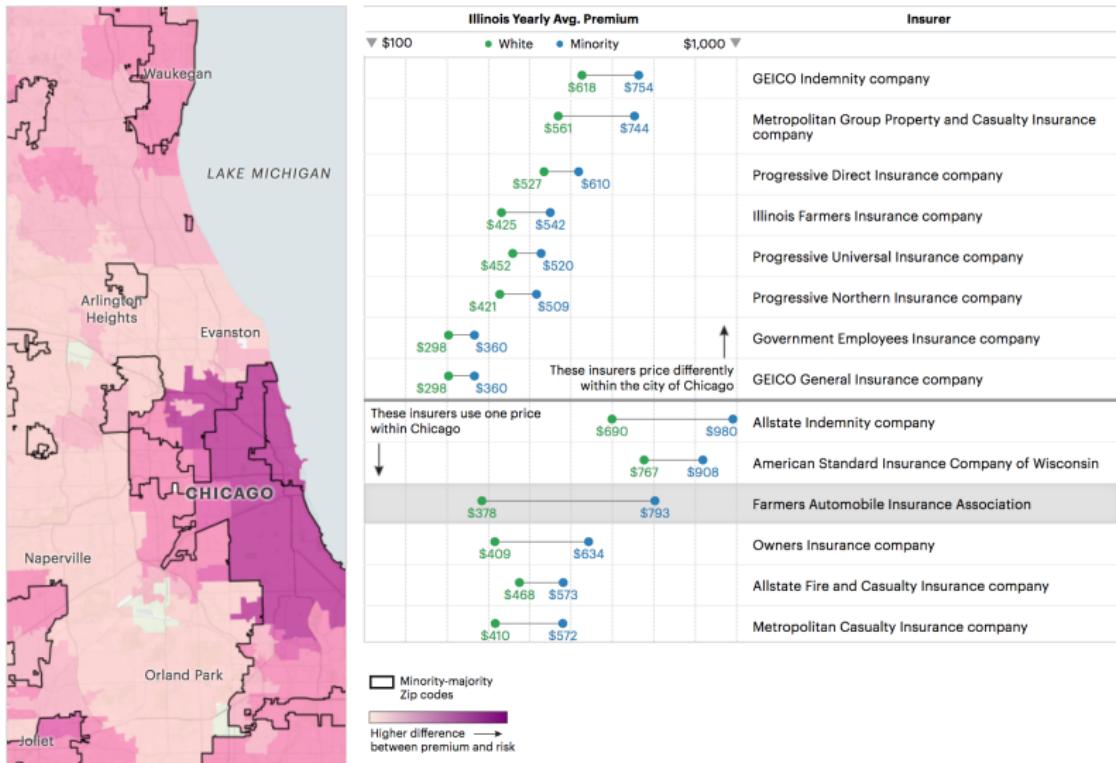
- ▶ Pre-history: non-automated decisions (50 Years of Test (Un)fairness, Hutchinson & Mitchell)
- ▶ Fairness, Accountability, and Transparency (FAT) workshop in ML conferences (2014: 2+0 organizers, 2016: 6+11)
- ▶ 2016: ProPublica series on Machine bias (COMPAS). Weapons of Math Destruction by Cathy O'Neil, ...
- ▶ 2017: more fairness papers in NeurIPS (causal approach takes off) and Kate Crawford's invited talk, more books, ...
- ▶ 2018: FAT* conference with hundreds of attendees, even more books, NYT vs FB takes off, Google walkout, ...
- ▶ 2019: Google AI ethics board, Algorithmic Accountability Act (impact assessments), presidential candidates with AI policies,
...



- ▶ Admissions
- ▶ Credit
- ▶ Employment
- ▶ Insurance
- ▶ Healthcare

Unique dangers: scale,
opacity, biased data, faux
objectivity

Discrimination based on non-sensitive variables? Location...



Source

Overt discrimination

- ▶ Failure to apply relevant laws/regulations
- ▶ Poor design/incentives
- ▶ Rationalized "statistical discrimination"

"Unintentional" disparate impact

- ▶ Non-diverse workplaces
- ▶ Bias introduced by data collection/use
- ▶ Unfair world: many predictor variables correlated with both outcome and protected attributes (race, sex, etc)

(Not an exhaustive list)

A narrow, “internal” focus

Here are some questions I won’t address today:

- ▶ What about privacy or other (ethical) objectives?
- ▶ Which attributes should be protected?
- ▶ How are such attributes (race, gender) defined or measured?
- ▶ What if the premise of the system is fundamentally unfair?
- ▶ Will market dynamics or Goodhart’s/Campbell’s law invalidate all this (eventually)?

Focus on an oversimplified model for a tiny part of the process/system in which these questions are embedded

Also off-topic: unsupervised fairness

Here is a poem written by Google Translate on the topic of gender. It is the result of translating [Turkish sentences](#) using the gender-neutral “o” to English (and inspired by [this](#) Facebook post).

Gender by Google Translate

he is a soldier
she's a teacher
he is a doctor
she is a nurse

he is a writer
he is a dog
she is a nanny
it is a cat

Source

Supervised learning

- ▶ Outcome variable \mathbf{Y} - consider as a score for decisions, or $\mathbf{Y} = 1$ as the desirable decision
- ▶ Sensitive/protected attribute(s) \mathbf{A} - race, gender, ...
- ▶ Other predictors \mathbf{X} - not sensitive (*prima facie*)

Machine learning task: learn a function $f(\mathbf{X}, \mathbf{A})$ from (labeled) training data to predict values of $\hat{\mathbf{Y}} = f(\mathbf{X}, \mathbf{A})$ on (unlabeled/future) test data (by minimizing some loss function that measures closeness of $\hat{\mathbf{Y}}$ to \mathbf{Y} on the training data)

What would it mean for such function to be fair with respect to \mathbf{A} ?

A causal perspective on fairness

Ethicists and social choice theorists have various notions about

- ▶ the role of agency in justice
- ▶ responsibility-sensitive egalitarianism
- ▶ luck egalitarianism

which rely on causal reasoning. Roughly, it is unfair for individuals to experience different outcomes due to factors outside of their control.

The Alan Turing Institute



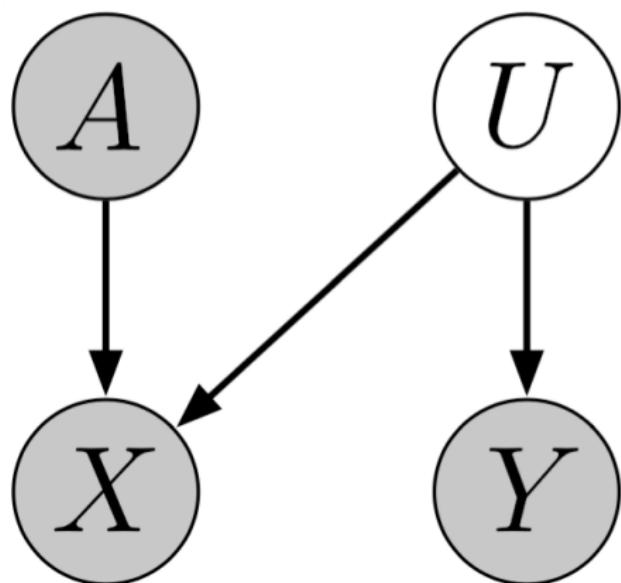
Statistical Laboratory

Basic idea

Model the relevant real-world causal relationships to separate discrimination on the basis of protected attributes from other, potentially correlated observables (and unobserved confounders)

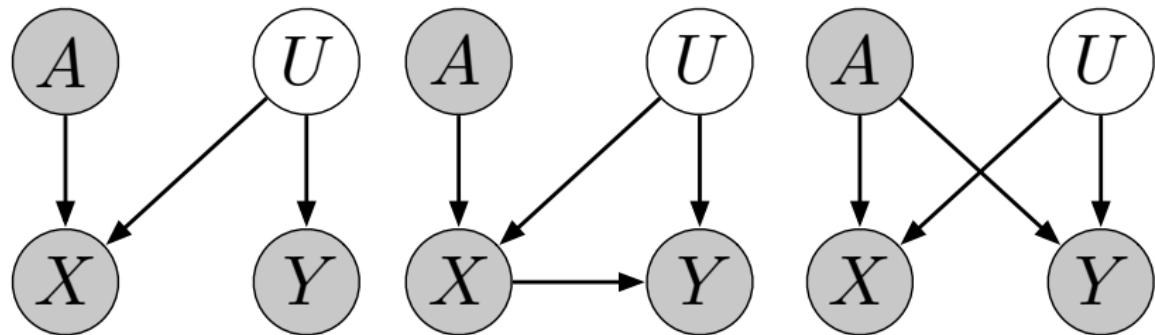
(Similar works: Pearl et al (2016) (textbook), DeDeo (2014), Kilbertus et al (2017), Johnson et al (2016), Nabi and Shpitser (2018), Zhang and Bareinboim (2017), Chiappa and Gillam (2018), and others)

DAG/SEM causal model framework



- ▶ Nodes: variables (**U** unobserved)
- ▶ Arrows: causal relationships / conditional (in)dependence
- ▶ Structural equations: functional forms of (arrow) relationships

Computing counterfactuals: follow the paths



Assume a probabilistic model for **U**, estimate parameters using observed data, change a to a' and propagate that change through the structural equations to all descendants of **A**

A causal definition of fairness

Counterfactual fairness

An estimator \hat{Y} is **counterfactually fair** if

$$\mathbb{P}(\hat{Y}_a | \mathbf{X} = x, \mathbf{A} = a) = \mathbb{P}(\hat{Y}_{a'} | \mathbf{X} = x, \mathbf{A} = a)$$

for all a' .

Proposition: structural counterfactual fairness

Any estimator \hat{Y} which is a function of only non-descendents of \mathbf{A} is counterfactually fair (sufficient condition, not necessary)

- ▶ This definition captures the intuition that the outcome should not depend in a causal way on an individual's sensitive attributes *or* other causal consequences thereof.
- ▶ With correct model of the world* it addresses root sources of unfairness.
- ▶ Next: A few *observational* definitions of fairness, a simple example, and then back to causality...

*Big assumption, but also **transparent** because the model is explicit.

(Likely violates “no causation without manipulation” unless we restrict to perception of the sensitive attribute)

Perhaps the most straightforward definition (and my favorite due to its elegant simplicity), often described as **equality of outcomes**

Demographic parity

Predictions (or decisions) are independent of **A**:

$$\mathbb{P}(\hat{\mathbf{Y}}|\mathbf{A} = 0) = \mathbb{P}(\hat{\mathbf{Y}}|\mathbf{A} = 1)$$

Equality of opportunity (Hardt et al, 2016)

The accuracy of the algorithm does not depend on \mathbf{A} :

$$\mathbb{P}(\hat{\mathbf{Y}} = 1 | \mathbf{A} = 0, \mathbf{Y} = 1) = \mathbb{P}(\hat{\mathbf{Y}} = 1 | \mathbf{A} = 1, \mathbf{Y} = 1)$$

Demographic parity but only among individuals “qualified” for the desirable outcome

Fairness definitions: Fairness through unawareness

"Equal treatment," people tend to believe such treatment is fair.

Grgic-Hlaca et al. (2016)

Prediction does not *explicitly* use **A**, i.e.

$$\hat{\mathbf{Y}} = f(\mathbf{X})$$

Unfortunately encoded in US law (despite contradicting other laws)

Dwork et al. (2012)

Similar predictions for individuals who are similar (in their unprotected attributes). If $\mathbf{X}_i \approx \mathbf{X}_{i'}$ then

$$\hat{\mathbf{Y}}(\mathbf{X}_i, \mathbf{A}_i) \approx \hat{\mathbf{Y}}(\mathbf{X}_{i'}, \mathbf{A}_{i'})$$

Continuity condition in \mathbf{X} but not \mathbf{A} . Can be related to matching approaches to causal inference

The Fundamental Contradiction of Fairness

Various works showing impossibility of simultaneously satisfying several of the different fairness definitions at once: Kleinberg et al (2016), Chouldechova (2016)

(Simplified) impossibility theorem

Unless the world is already fair, the only solutions satisfying both equal treatment (or opportunity) and equal outcomes (demographic parity) are trivial ones (e.g. jail everyone)

Many versions of this can be proven with different sets of assumptions but basically the same conclusion: some fairness definitions are contradictory

Causal models as tools for understanding fairness

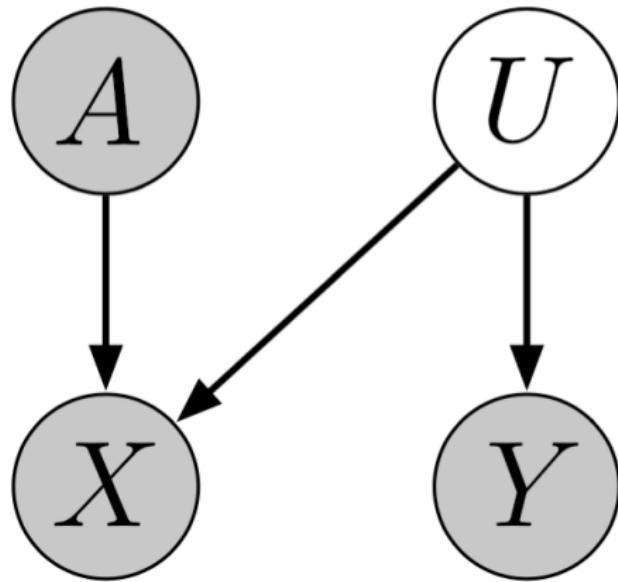
I do not advocate everyone use the counterfactual definition of fairness. Maybe there is no one “right” answer. This approach is useful for probing/understanding limitations.

Chief Justice Roberts: “The way to stop discrimination on the basis of race is to stop discriminating on the basis of race” (PICS, 2007).

Equal treatment or fairness through unawareness

Is it actually any good?

Chief Justice Roberts is *mathematically* wrong

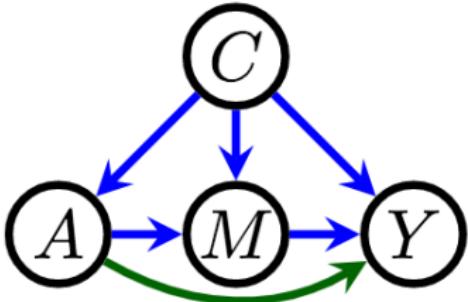


- ▶ Auto insurance risk **Y**
- ▶ Car color **X**
- ▶ Policyholder's gender **A**
- ▶ "Aggressiveness" **U**

"Equal treatment" actually *introduces* unfairness where there was none

Pathway analysis / decomposition

- ▶ Kusner et al (2017): path-dependent counterfactual fairness (supplement)
- ▶ Kilbertus et al (2017): proxies and resolving variables
- ▶ Nabi and Shpitser (2018): path specific effects, mediators, constrain parameters
- ▶ Chiappa and Gillam (2018): more flexible modeling, modify features
- ▶ Zhang and Bareinboim (2017): (counterfactual) direct, indirect, and spurious



Model for crime data where the mediator can be, e.g., prior convictions

- ▶ Enhanced capability of causal modeling to address fairness questions
- ▶ Capture *some aspects* of equal treatment, equal outcomes, equal opportunity
- ▶ Does this resolve the Fundamental Contradiction of Fairness?

In my opinion: no. People will disagree about pathways

Problem: consensus on a causal model / pathways

When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness (Russell et al, NeurIPS 2017)

- ▶ Competing causal models
- ▶ Approximate counterfactual fairness (relax equality constraint)
- ▶ Predictions approximately satisfy fairness across both (all) models
- ▶ Limitation: the more contradictory are the competing models, the more trivial the predictions (constant)
- ▶ Causal framing of fundamental contradiction

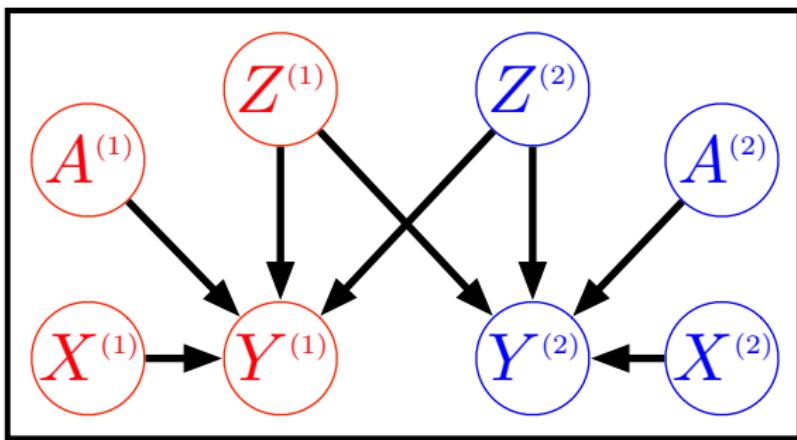
Resolving the contradiction

I think this is the right *path*. It's now about understanding the causes of unfairness well enough to reach consensus.

- ▶ Model the **intervention** (that predictions will be used for) as its own separate variable **Z**
- ▶ Causal model including **Z**, explicitly model changes in the world due to use of ML
- ▶ Potentially multiple objectives to represent interests of different stakeholders

Recent work: Making Decisions that Reduce Discriminatory Impacts
(Kusner et al, *ICML 2019*)

Example: fair intervention under interference



Counterfactual privilege

When designing an optimal, fair intervention \mathbf{Z} , instead of enforcing the equality in the definition of counterfactual fairness, we can also use an asymmetric bound on **counterfactual privilege**, for $\tau \geq 0$

$$\mathbb{E}[\hat{\mathbf{Y}}(\textcolor{red}{a}, \mathbf{Z})] - \mathbb{E}[\hat{\mathbf{Y}}(\textcolor{blue}{a'}, \mathbf{Z})] \leq \tau$$

- ▶ In practice these *asymmetric constraints* will only be active for privileged values of a (actual, left term), and inactive otherwise
- ▶ Intervention should not allocate resources in a way that helps people (in expectation) become more than τ (in terms of the outcome) units better than they would be if they were not privileged

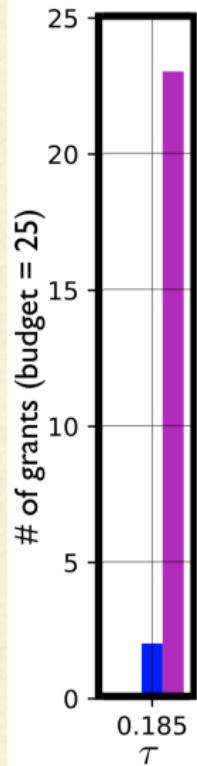
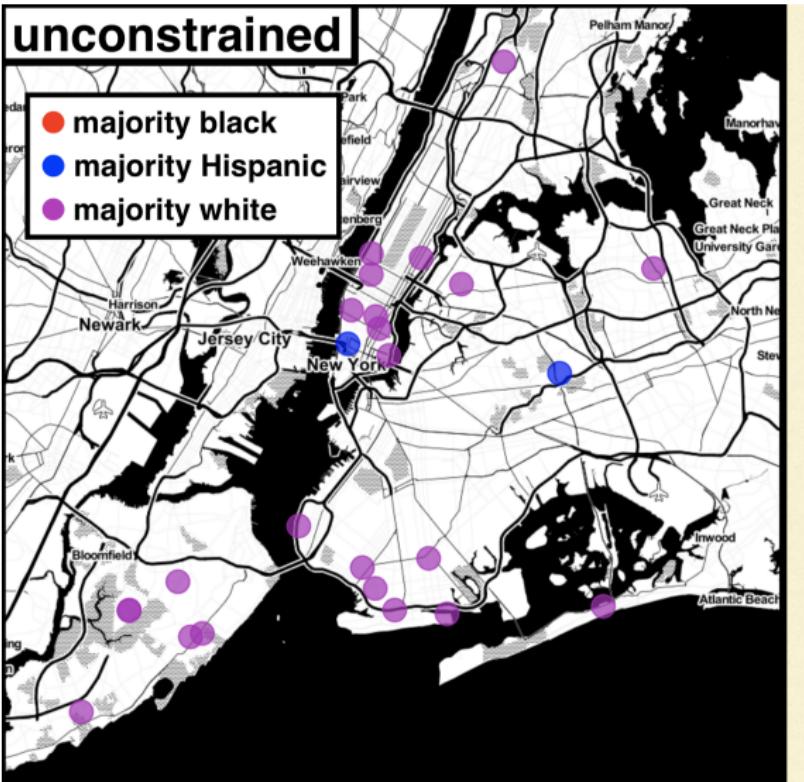
Optimal intervention under interference

- Our goal is to design optimal interventions or policies \mathbf{Z} subject to a budget constraint, e.g.

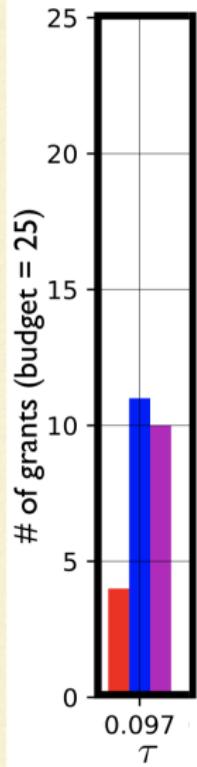
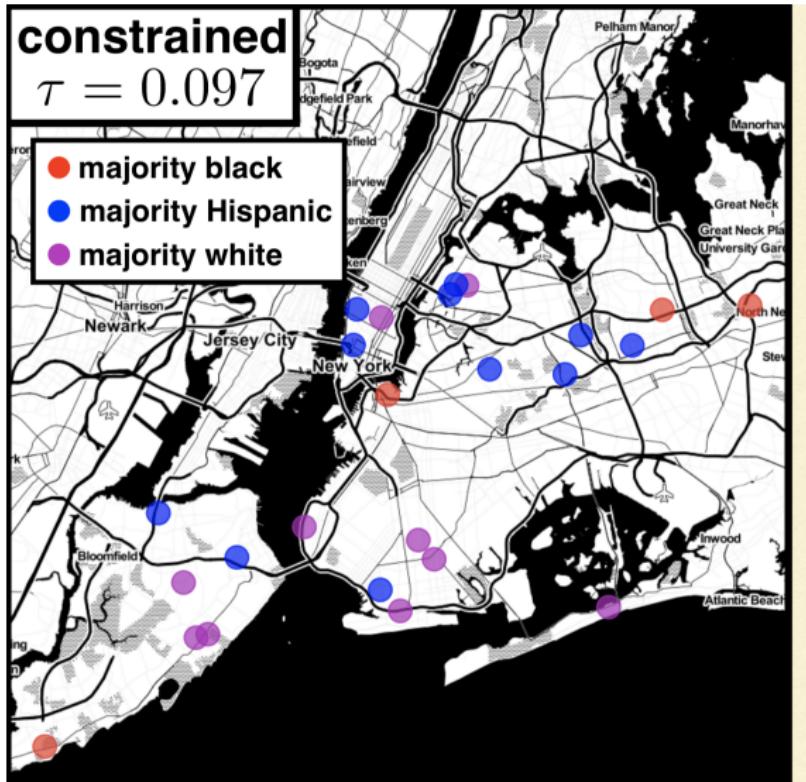
$$\mathbf{Z} = \arg \max \sum_i \mathbb{E} \left[\hat{\mathbf{Y}}^{(i)}(a^{(i)}, \mathbf{Z}) | \mathbf{A}^{(i)}, \mathbf{X}^{(i)} \right] \quad s.t. \quad \sum_i \mathbf{Z}^{(i)} \leq b$$

- Interference means $\mathbf{Y}^{(i)}$ is potentially a function of all of \mathbf{Z} and not just $\mathbf{Z}^{(i)}$
- Next two slides: optimal interventions with and without counterfactual privilege constraint

School resource allocation without fairness constraint



School resource allocation bounded counterfactual privilege



Other directions

Things I'm working on

- ▶ Drawing attention to statistical aspects of fair ML,
e.g. sampling bias in training data (equal opportunity?)
- ▶ Attacking bad ideas (from SCOTUS or elsewhere) like the “cost
of fairness” or “trade-off” between “accuracy” and fairness

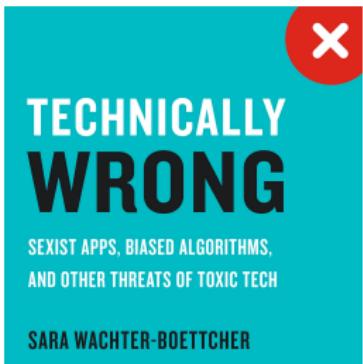
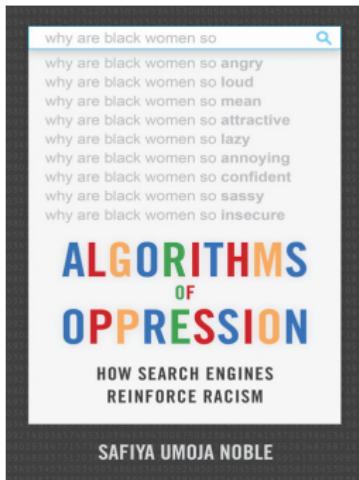
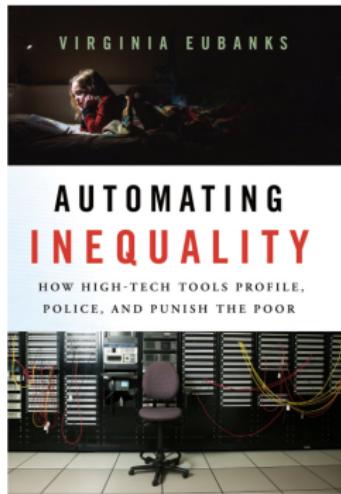
Things it would be cool if other people worked on, in general

- ▶ Interdisciplinary/STS-informed tech/quantitative research
- ▶ Empirical ethics, “big data”/ML-informed social research
- ▶ Political/economy of whether any of this is actually headed in
a good direction

Questions?

Thank you for listening!

More books full of examples: Eubanks, Noble, Wachter-Boettcher



The rise of big data policing: Surveillance, Race, and the Future of Law Enforcement by A. Ferguson