Causal Inference in Algorithmic Fairness

Sepehr Torab Parhiz

Winter 2019

What is the problem?

- Algorithms are now used set parole, give loans, and set insurance premiums
- Back when humans made these decisions, biases such as racism influenced how individuals from different groups were treated
- Algorithms tend to exhibit similar prejudices

Preliminaries & Notation

- Causal model defined with triplet (U, V, F)
  - U: unobservable variables
  - V: observable variables
  - F: set of structural assignments in a Structural Causal Model
- A Fairness Problem
  - Y is the real outcome, $\hat{Y}$ the predicted outcome
  - A is the protected attribute, X are the rest of the features
- Intervention: $P(Y | do(X = x))$, Counterfactual: $V_i(v_j, u)$

Case Study: COMPAS

- COMPAS predicts whether defendants will commit crimes after release
- Judges use COMPAS score to set bail and parole terms
- ProPublica reported that COMPAS was significantly biased against black individuals
- In reality, black and white individuals with similar background reoffend at the same rates

Fairness Criteria

- **Equalized odds**: Predict the the real outcome with the same probability for all values of A
    - $P(\hat{Y} = y \,|\, A = a, Y = y) = P(\hat{Y} = y \,|\, A = a', Y = y)$
    - In other words: $\hat{Y} \perp\!\!\!\perp A \,|\, Y$
- **Calibration**: For a predicted outcome, the probability of that outcome being true should be the same for all values of A
    - $P(Y = y \,|\, A = a, \hat{Y} = y) = P(Y = y \,|\, A = a', \hat{Y} = y)$
    - In other words: $Y \perp\!\!\!\perp A \,|\, \hat{Y}$

Incompatibility of Fairness Measures

- COMPAS used *calibration* as a fairness criterion
- ProPublica considered COMPAS to be unfair, using *equalised odds* as a measure
- Both measures are reasonable to some extent
- Kleinberg et al. showed that these two constraints cannot be simultaneously satisfied

Kleinberg, Jon M. et al. (2017)

Fairness Criteria for Populations & Individuals

- **Demographic Parity/Disparate Impact**
  - $P(\hat{Y} = y \mid A = a) = P(\hat{Y} = y \mid A = a')$ **Or** $\hat{Y} \perp\!\!\!\perp A$
  - It can cause discrimination, undermines calibration and equalized odds

- **Individual Fairness**

$$P(\hat{Y}^{(i)} = y \mid X^{(i)}, A^{(i)}) \approx P(\hat{Y}^{(j)} = y \mid X^{(j)}, A^{(j)}) \approx \text{, **if** } d(i, j) \approx 0$$

  - $d(\,.\,,\,.\,)$ is a task-specific similarity metric between individuals
  - Fixes DP's discrimination, but choosing $d(\,.\,,\,.\,)$ is a hard problem

Kamiran, F., & Calders, T. (2009) **Dwork, C. et al. (2012).**

Why does Causality matter in Fairness?

- Many ideas and statements in ethics and law are causal in nature
  - Agency & Egalitarianism in Justice
- Unfairness: Experiencing different outcomes due to caused by factors that are out of one's control
- Causal Inference is the superior platform for dealing with confounders and inherent biases

Shortcomings of Conventional Fairness Measures

- If A and Y are not independent, true outcomes are biased themselves
    - The judges may be prejudiced toward minorities
- Both calibration and equalized odds fail to mitigate inherent bias in the data
- This is a problem that Causal Inference can solve

Bareinboim, Elias and Judea Pearl. “Causal inference and the data-fusion problem.”

Inherent Bias: Gender Bias & Simpson's Paradox

- Berkeley Admissions: a lower percentage of women were accepted to graduate programs compared to men
- If we control for department of choice, unfairness disappears.
- Women applied to the most selective programs
- Bickel et al. concluded that socialization caused women to apply to more crowded, less funded departments
- Pearl extensively analyzes this example from a causal standpoint

Counterfactual Fairness

- A predictor $\hat{Y}$ satisfies *counterfactual fairness (CF)* if:

$$P(\hat{Y}(a, U) = y \,|\, X = x, A = a) = P(\hat{Y}(a', U) = y \,|\, X = x, A = a)$$

- "...other things being equal, our prediction would not have changed in the parallel world where only A would have changed."

- "...we purposely avoid making use of any information concerning the structural equation for Y in model. This is motivated by the fact that Y must not make use of $\hat{Y}$ at test time.

Kusner, Matt J. et al. (2017)

Counterfactual Fairness vs. Conventional Fairness Criteria

- CF satisfies Demographic Parity if the predictor is independent of A and a function of U and X
- Two different individuals can be compared as counterfactual version of each other
- "...the counterfactual version of individual i ... is in reality an observed case j in a sample of controls, such that i and j are close..."
- "close" is the similarity metric used in Individual Fairness
- But here the fairness condition only holds for matched pairs

Counterfactual Fairness vs. Conventional Fairness Criteria

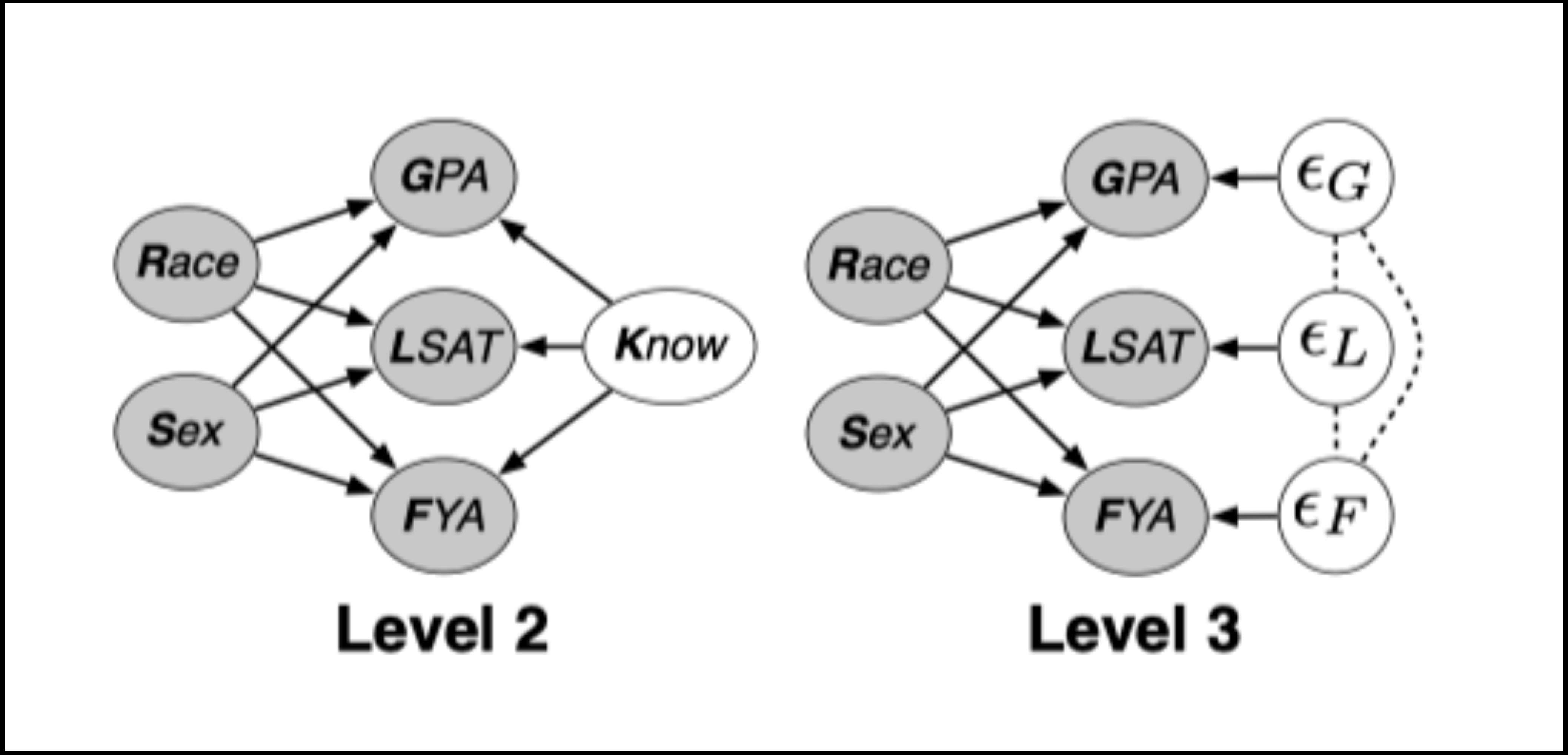
- Y and A are independent if there are no causal paths exist between them
- Consider a counterfactually fair predictor, only trained with non-descendants of A
  - It can be shown that it will respect both Equalized Odds & Calibration

Counterfactual Fairness: Three levels of Causal Modeling

- Three levels of assumptions of increasing strength
- Level 1: Build predictor using only the observable non-descendants of A
- Level 2: Latent variables that act as non-deterministic causes of observable variables
- Level 3: Fully deterministic model with latents
    - Treat $P(V_i | pa_i)$ as an additive error model, $V_i = f_i(pa_i) + e_i$
    - e_i as an in put to $\hat{Y}$

Counterfactual Fairness: Applications

	Full	Unaware	Fair K	Fair Add
RMSE	0.873	0.894	0.929	0.918



Kusner, M.J. (2017)

τ-Controlled Counterfactual Privilege

- Our goal is to assign (binary) interventions z to maximize the sum of expected outcomes over individuals subject to a maximum budget B

$$\mathbf{z}^\star \equiv \text{argmax}_{\mathbf{z}} \sum_{i=1}^{n} \mathbb{E}[Y_i(\mathbf{z}) \mid A_i = a_i, X_i = x_i],$$
$$s.t., \sum_{i=1}^{n} z_i \leq B$$

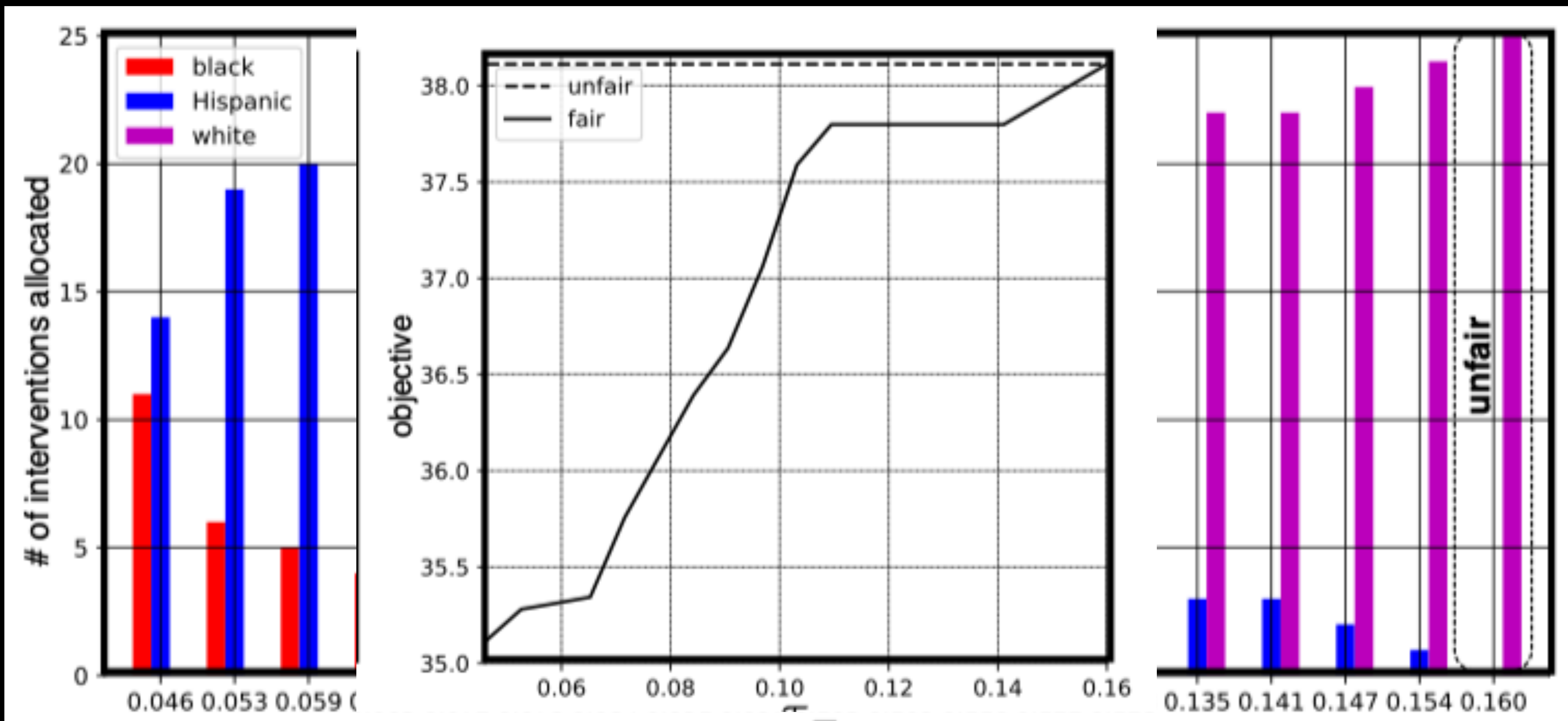
$$\underbrace{\mathbb{E}_{\mathcal{M}^\prec}[Y_i(a_i, \mathbf{z}) \mid A_i = a_i, X_i^\prec = x_i^\prec] - \mathbb{E}_{\mathcal{M}^\prec}[Y_i(a', \mathbf{z}) \mid A_i = a_i, X_i^\prec = x_i^\prec]}_{G_{ia'}} < \tau,$$

$X_i^\prec$ is the subset of X_i that are non-descendants of A_i in the causal graph, and $\mathcal{M}^\prec$ is a causal model that excludes all observed non-descendants of A but Y

The Fair Optimization Problem

- Formulated as a mixed-integer-linear-program
- Also a set of neighbors are considered for spillover effect
- $$\max_{z_1,\ldots,z_n} \sum_{i=1}^{n} \mathbb{E}[Y_i(\mathbf{z}) \mid A_i = a_i, X_i = x_i]$$
  $$s.t., \sum_{i=1}^{n} z_i \leq B$$
  $$G_{ia'} \leq \tau \quad \forall a' \in \mathcal{A},\ i \in \{1,\ldots,n\},$$

Counterfactual Fairness: Applications



Kusner, M.J. (2018)

Causal Fairness: Interventions

- Enforcing fairness by constraining interventional distributions

$$P(\hat{Y}|do(A=a)) = P(\hat{Y}|do(A=a'))$$

- A family of causal models are created as to minimize total effect of A on $\hat{Y}$

Kilbertus, Niki et al. (2017).

Kilbertus et al. on Counterfactual Fairness

- "It requires modeling counterfactuals on a per individual level, which is a delicate task..."
- "...Even determining the effect of race at the group level is difficult."
- An interesting argument among Causality researchers:
  - "Can we estimate causal effects for causes that we cannot understand in the real world?"

Pearl, J. (2018). "Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes". https://ftp.cs.ucla.edu/pub/stat_ser/r483-reprint.pdf

Judea Pearl @yudapearl · 15 Dec 2018

5/n
that DEFINES potential outcomes. This explains why @_MiguelHernan depicts it as black magic when I assert that an ideal intervention is defined as a property of one's model. This conceptual barrier continues to impede communication until ..(YES)... a metamorphosis occurs...

2 13

Miguel Hernán
@_MiguelHernan

Following

Replying to @yudapearl

No, not black magic. Just magic.

Black magic is evil. Your magical thinking is simply extrascientific.

By the way, your belief that hopelessly ill-defined causal questions can bring us the answer to life, the universe, and everything reminds me of this:

https://twitter.com/_MiguelHernan/status/1074164493419200512

Judea Pearl*

Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes

Miguel Hernán @_MiguelHernan · 16 May 2018

We cannot estimate "the causal effect of obesity" because we don't know what that means.

For the causal effect of A to be well defined, we need a common understanding of the interventions that we would use to change A. Otherwise, the effect is undefined.

Miguel Hernán
@_MiguelHernan

Following

Pearl believes that any causal effect we can name must also exist.

To him, the meaning of “the causal effect of A on death” is self-evident. He says we can quantify, say, the causal effect of race or the causal effect of obesity.

I don't think we can.

https://twitter.com/_MiguelHernan/status/996876145953050624

Loftus et al. on Kilbertus et al.

- On the criterion: "...not realistic if X is a descendant of A in the causal graph, since in this case no single individual will keep X at a fixed level as A hypothetically varies."

- "...it is perfectly possible that $\hat{Y}$ is highly discriminatory in a counterfactual sense and yet satisfies the purely interventional criterion..."

- Example: consider structural assignment $Y = f(A, U_Y)$ such that:

$P(U_Y = 0) = P(U_y = 1) = \frac{1}{2}$ **and** $f(a,1) = 1, f(a,0) = 1 - a$ **for** $a \in \{0,1\}$

- Then: $P(Y = 1 \mid do(A = 1)) = P(Y = 1 \mid do(A = 0)) = \frac{1}{2}$

- Even though for every individual: $Y(a, u_Y) = 1 - Y(1 - a, u_Y)$

The Causal Explanation Formula

- Explain discrimination by breaking up causal effects to 3 categories
- Partition discrimination to direct and indirect
- Alternative notions of causal unfairness
  - Zhang et al. consider $A \leftarrow X \rightarrow \hat{Y}$ as spurious discrimination
  - But Loftus et al. believe that then X is protected too

Zhang, J., & Bareinboim, E. (2018). Fairness in Decision-Making

The Causal Explanation Formula

- Direct discrimination/Disparate treatment
    - Enforces procedural fairness
    - "...The equality of treatments that prohibits the use of the protected attribute in the decision process."
- Indirect discrimination/Disparate impact
    - Enforces outcome fairness
    - "...the equality of outcomes among protected groups."
    - "...occurs if a facially neutral practice has an adverse impact..."

The Causal Explanation Formula

- Direct discrimination $A \rightarrow Y$
- Indirect discrimination
  - Indirect causal discrimination $A \rightarrow M \rightarrow Y$
  - Indirect spurious discrimination $A \leftarrow Z \rightarrow Y$
- Zhang et al. argue that none of the existing measures are capable of detecting all three types of discrimination

The Causal Explanation Formula

- Recap: Demographic parity: $P(\hat{Y} = y \mid A = a) = P(\hat{Y} \mid A = a')$
- Total Variation $TV_{a,a'}(y) = P(\hat{Y} = y \mid A = a) - P(\hat{Y} = y \mid A = a')$
- Recap: Counterfactual fairness:

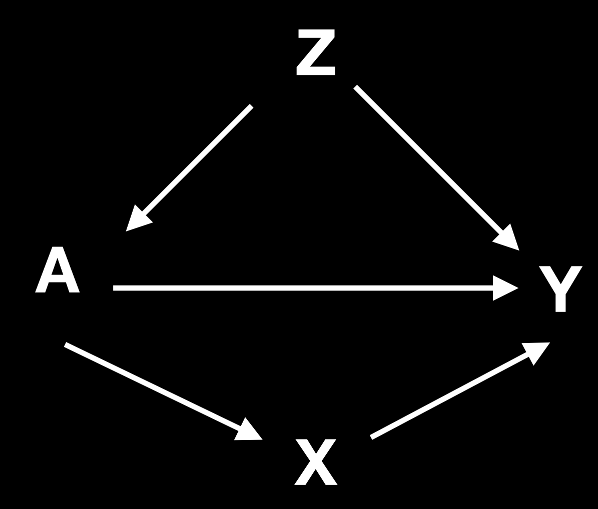
$$P(\hat{Y}(a, U) = y \mid X = x, A = a) = P(\hat{Y}(a', U) = y \mid X = x, A = a)$$

- Effect of Treatment on the Treated

$$ETT_{a,a'}(y) = P(\hat{Y}(a, U) = y \mid X = x, A = a) - P(\hat{Y}(a', U) = y \mid X = x, A = a)$$

Toy Example: Hiring

- Y: Hiring decision, 1 for hired and 0 for not hired
- A: Religious belief, 1 for believer and 0 for non-believer
- Z: Educational background, 1 for high, 0 for low
- X: Location of the applicant, 1 for close to religious institutes, 0 for distant
- Assume Z has a negative effect on A
- Assume Z has a positive effect on X



Toy Example: Hiring

- An applicant sues the company for discrimination
- The court notices that $TV_{a,a'}(y) = 1$ where $Y = 1$
- The company argues the disparity is mainly caused by Z
- How can we verify this claim?
- Neither ETT, nor TE will show any unfairness
    - These two along, other measures, only account for direct and indirect effects

Decomposing TV to Counterfactual Measures

- Counterfactual Direct Effect (Ctf-DE)

$$DE_{a,a'}(y|A) = P(Y(a', U), X(a, U)) - P(Y(a, U)|A)$$

- $DE_{a,a'}(y|A)$ captures existence of disparate treatment
- It is proved that if $DE_{a,a'}(y|A) \neq 0$, there is a direct path connecting A and Y

Decomposing TV to Counterfactual Measures

- Counterfactual Indirect Effect (Ctf-IE)

$$IE_{a,a'}(y|A) = P(Y(a,U), X(a',U)) - P(Y(a,U)|A)$$

- "For $A = a$, $IE_{a,a'}(y|A = a)$ measures changes in the probability of the outcome Y would be y had A been a , while changing X to whatever level it would have obtained had A been a' , in particular, for the individuals that (naturally) have $A = a$."

- It is proved that if $IE_{a,a'}(y|A) \neq 0$, there is a indirect path connecting A and Y

Decomposing TV to Counterfactual Measures

- Counterfactual Spurious Effect (Ctf-SE)

$$SE_{a,a'}(y) = P(Y(a, U) | A = a') - P(y | A = a)$$

- " $SE_{a,a'}(y)$ measures the difference in outcome $Y = y$ had A been a for the individuals that would naturally choose A to be a versus a'.
- It is proved that if $SE_{a,a'}(y) \neq 0$, there is a back-door path connecting A and Y

Decomposing TV to Counterfactual Measures

- Total disparity (TV) experienced by the individuals naturally attaining a' relative to the ones attaining a equals to the disparity experienced due to the spurious discrimination minus the advantage the ones attaining a' would have gained had they been a

$$TV_{a,a'}(y) = SE_{a,a'}(y) - ETT_{a,a'}(y)$$

$$TV_{a,a'}(y) = ETT_{a',a}(y) - SE_{a',a}(y)$$

$$ETT_{a,a'}(y) = DE_{a,a'}(y \mid A = a) - IE_{a',a}(y \mid A = a)$$

Causal Explanation Formula

- Total disparity experienced by the individuals who have naturally attained a' (relative to a) equals to the disparity experienced associated with spurious discrimination, plus the advantage it lost due to indirect discrimination, minus the advantage it would have gained without direct discrimination

$$TV_{a,a'}(y) = SE_{a,a'}(y) + IE_{a,a'}(y \mid A = a') - DE_{a',a}(y \mid A = a')$$

$$TV_{a,a'}(y) = DE_{a,a'}(y \mid A = a) - SE_{a',a}(y) + IE_{a',a}(y \mid A = a)$$

Causal Explanation Formula for Linear Models

$$IE_{a,a'}(Y|A=\alpha) = \gamma_{yx.\alpha z}\gamma_{x\alpha.z}(a'-a)$$

$$DE_{a,a'}(Y|A=\alpha) = \gamma_{ya.zx}(a'-a)$$

$$SE_{a,a'}(Y|A) = \gamma_{\alpha z}(\gamma_{yz.\alpha x} + \gamma_{yx.\alpha z}\gamma_{xz.\alpha})(a'-a)$$

- γ are the corresponding (partial) regression coefficient.

$$TV_{a,a'}(Y) = SE_{a,a'}(Y) + IE_{a,a'}(Y|A=\alpha) + DE_{a,a'}(Y|A=\alpha)$$

Applications: Designing Reparatory Policies

- Companies and universities are required to fix unfairness if outcome disparity persists

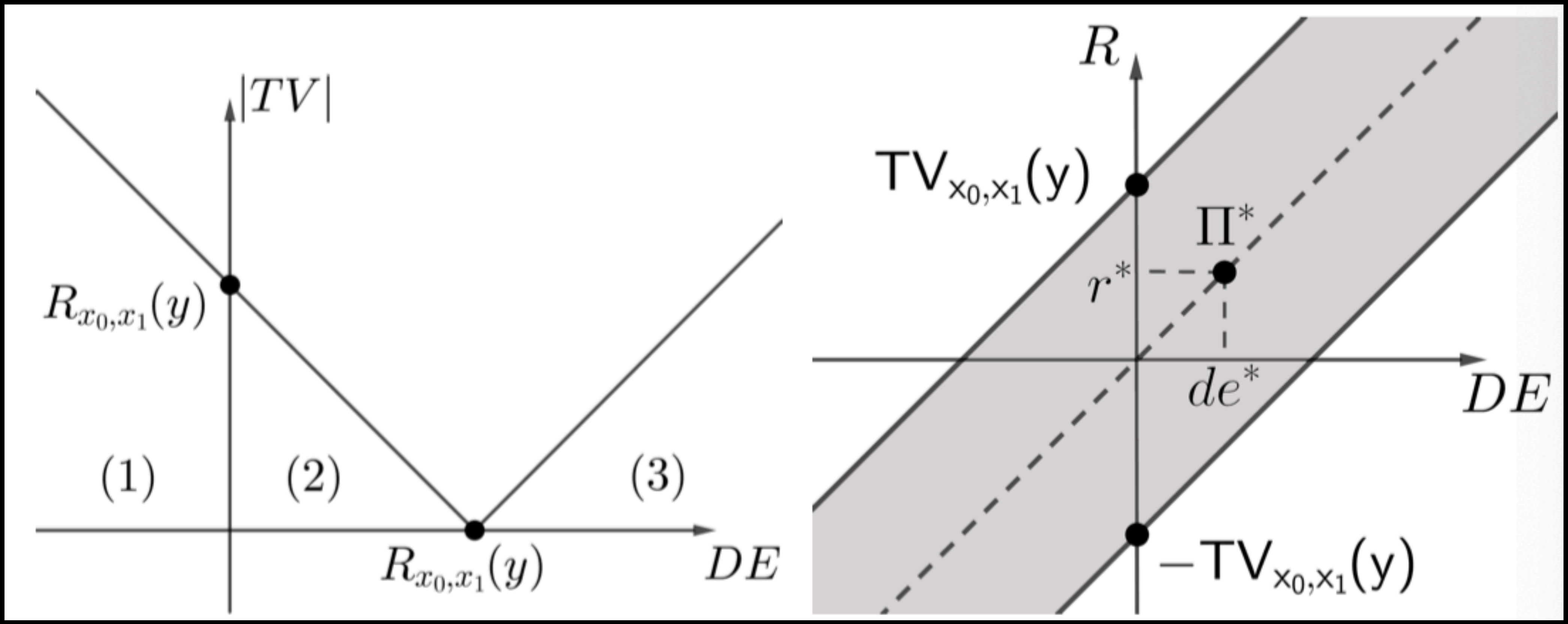
- Affirmative action: Compensate previous discrimination by providing opportunities for members of the protected group

- A trade-off between Procedural and Outcome fairness

- Zhang et al. advocate for “narrowly tailoring” of affirmative action, to not introduce *reverse discrimination*

Applications: Designing Reparatory Policies

- Residual disparity: $R_{a,a'}(y) = SE_{a,a'}(y) + IE_{a,a'}(y \mid a')$
- Fix positive $R_{a,a'}(y)$ and manipulate $DE_{a',a}(y)$ so as to minimize total disparity $|TV_{a,a'}(y)| = |R_{a,a'}(y) - DE_{a',a}(y \mid a')|$



- Narrow tailoring is satisfied only if: $DE_{a',a}(y \mid A = a') \in [0, R_{a,a'}(y)]$

What is Next?

- A paradigm for composition of causal measures of fairness
- Formalizing Interventions as a tool in policy-making
- Manipulating the non-manipulable causes

- Racial Bias and In-group Bias in Judicial Decisions: Evidence from Virtual Reality Courtrooms. Samantha Bielen, Wim Marneffe, Naci H. Mocan. December 2018

- "We shot videos of criminal trials using 3D Virtual Reality (VR) technology, prosecuted by actual prosecutors and defended by actual defense attorneys in an actual courtroom."

- "...allows us to replace white defendants in the courtroom with individuals who have Middle Eastern or North African descent in a real-life environment. We alter only the race of the defendants in these trials, holding all activity in the courtroom constant".

- "...significant overall racial bias in conviction decisions against minorities"

https://www.semanticscholar.org/paper/Racial-Bias-and-In-group-Bias-in-Judicial-Evidence-Bielen-Mocan/4aad052ab688a8b8e829caaeeede59ee05493b9e

Bibliography

- Russell, C., Kusner, M.J., Loftus, J.R., & Silva, R. (2017). When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. NIPS.
- Loftus, J.R., Russell, C., Kusner, M.J., & Silva, R. (2018). Causal Reasoning for Algorithmic Fairness. CoRR, abs/1805.05859.
- Kusner, M.J., Russell, C., Loftus, J.R., & Silva, R. (2018). Causal Interventions for Fairness. CoRR, abs/1806.02380.
- Kusner, M.J., Loftus, J.R., Russell, C., & Silva, R. (2017). Counterfactual Fairness. NIPS.
- Kleinberg, J.M., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. ITCS.
- Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. Proceedings of the National Academy of Sciences of the United States of America, 113 27, 7345-52.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R.S. (2012). Fairness through awareness. ITCS.
- Kamiran, F., & Calders, T. (2009). Classifying without discriminating. 2009 2nd International Conference on Computer, Control and Communication, 1-6.
- Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley. Science, 187(4175):398–404, 1975.
- Judea Pearl. Causality: Models, Reasoning and Inference. Cambridge University Press, 2000.
- Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding Discrimination through Causal Reasoning. NIPS.
- VanderWeele, T.J., & Robinson, W.R. (2014). On the causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiology, 25 4, 473-84.
- Nabi, R., & Shpitser, I. (2018). Fair Inference on Outcomes. Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence, 2018, 1931-1940.
- Zhang, J., & Bareinboim, E. (2018). Non-Parametric Path Analysis in Structural Causal Models.
- Zhang, J., & Bareinboim, E. (2018). Fairness in Decision-Making - The Causal Explanation Formula. AAAI.
- Zhang, J., & Bareinboim, E. (2018). Equality of Opportunity in Classification : A Causal Approach.
- Bielen, S., Mocan, N., Eren, O., Kantor, S., Kitchens, C., Isaac, M.C., Unel, B., Voigt, S., & Willage, B. (2018). Racial Bias and In-group Bias in Judicial Decisions: Evidence from Virtual Reality Courtrooms.