A causal and counterfactual view of (un)fairness in automated decision making

Razieh Nabi, PhD

Department of Biostatistics and Bioinformatics
Rollins School of Public Health
Emory University

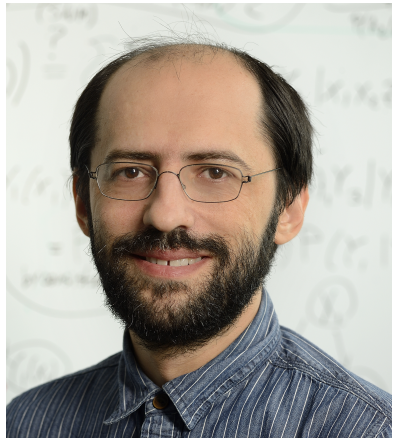
razieh.nabi@emory.edu

September 18, 2022

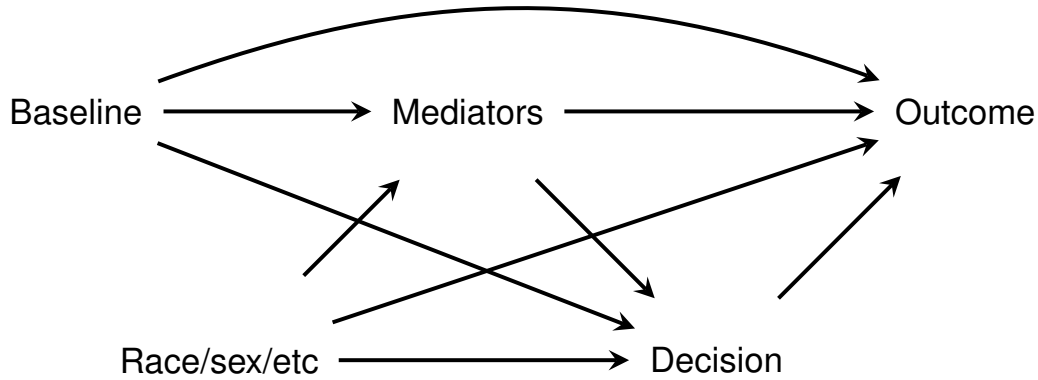
Co-authors

Daniel Malinsky
Department of Biostatistics
Columbia University

Ilya Shpitser
Department of Computer Science
Johns Hopkins University

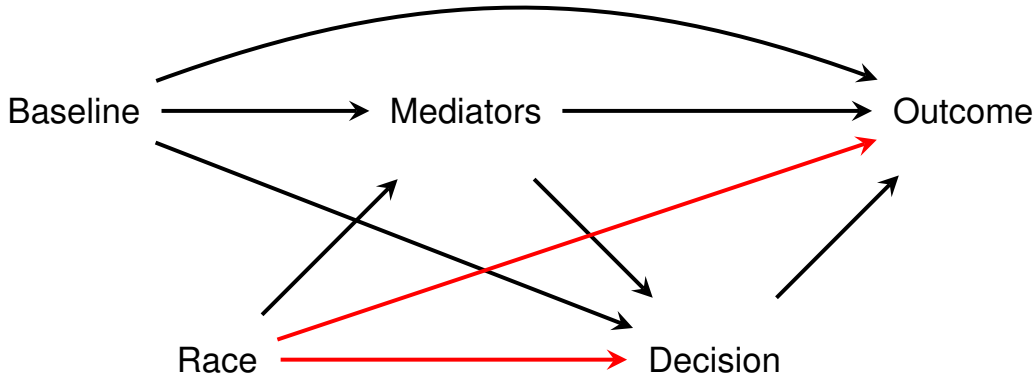


Example: child welfare



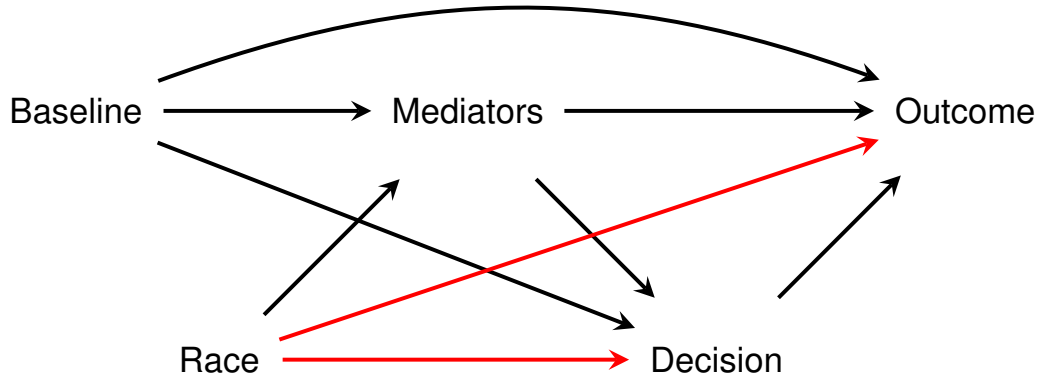
- Decision to dispatch case-worker may depend on all available information, and optimal decision would minimize negative outcomes (e.g. child separation and/or hospitalization).
- Unconstrained optimal decision-making may lead to unacceptable racial disparities.
- Ignoring race information is insufficient: dependence due to proxies

Our perspective



- In a “fairer world,” certain (discriminatory or unjust) mechanisms would be absent.
- This corresponds to the absence of some path-specific causal effects (RN and Shpitser, 2018).

Our perspective



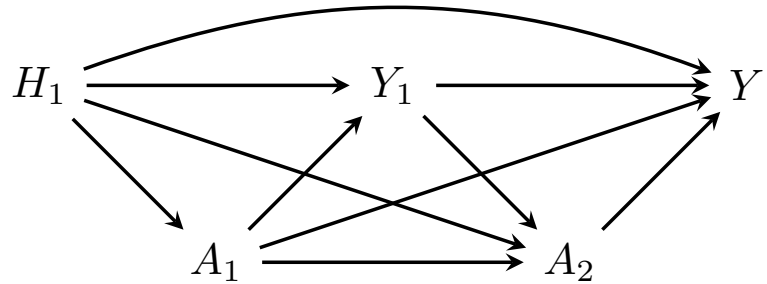
- In a “fairer world,” certain (discriminatory or unjust) mechanisms would be absent.
- This corresponds to the absence of some path-specific causal effects (RN and Shpitser, 2018).
- Approximate the “nearest fair world” and learn optimal policies there (RN, Malinsky, Shpitser, 2019)
- Must sacrifice some optimality to make decisions fairly.

Sequential decision making

Decision rule: $f_{A_i} : \mathcal{H}_i \mapsto \mathcal{A}_i$

Policy: $f_A = \{f_{A_1}, f_{A_2}\}$

(dynamic treatment regimes)

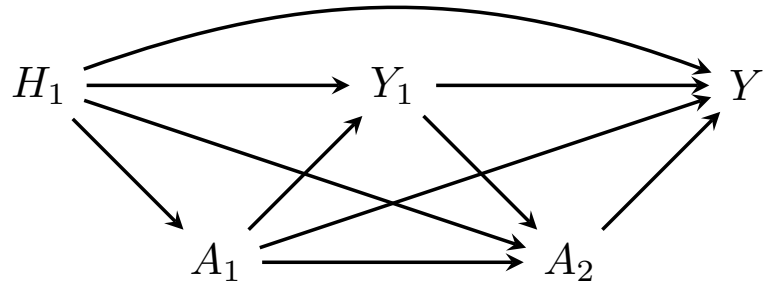


Sequential decision making

Decision rule: $f_{A_i} : \mathcal{H}_i \mapsto \mathcal{A}_i$

Policy: $f_A = \{f_{A_1}, f_{A_2}\}$

(dynamic treatment regimes)



- Counterfactual response under f_A is denoted by $Y(f_A)$
- Optimal policy: $f_A^* := \arg\max_{f_A} \mathbb{E}[Y(f_A)]$
- Fairness concerns arise since $H_1 = \{X, S, M\}$

Main methodological questions

1. How to express fairness principles mathematically?
2. How to construct a fair distribution/world?
3. How to lean optimal policies in the fair distribution/world?

- RN and I. Shpitser, *Fair Inference on Outcomes,* AAAI 2018.
- RN, D. Malinsky, and I. Shpitser, *Learning Optimal Fair Policies,* ICML 2019.
- RN, D. Malinsky, and I. Shpitser, *Optimal Training of Fair Predictive Models,* CLeaR 2022.

1. Fairness notions

Definitions of (un)fairness

Associative measures of (un)fairness:

- Disparate impact, predictive parity, equalized odds, etc

Definitions of (un)fairness

Associative measures of (un)fairness:

- Disparate impact, predictive parity, equalized odds, etc

- Ignore the underlying DGP
- Ignore the role of policy makers, bioethicists, legal experts (one-fits-all definitions)
- Not adaptable to use context-specific information

Definitions of (un)fairness

Associative measures of (un)fairness:

- Disparate impact, predictive parity, equalized odds, etc

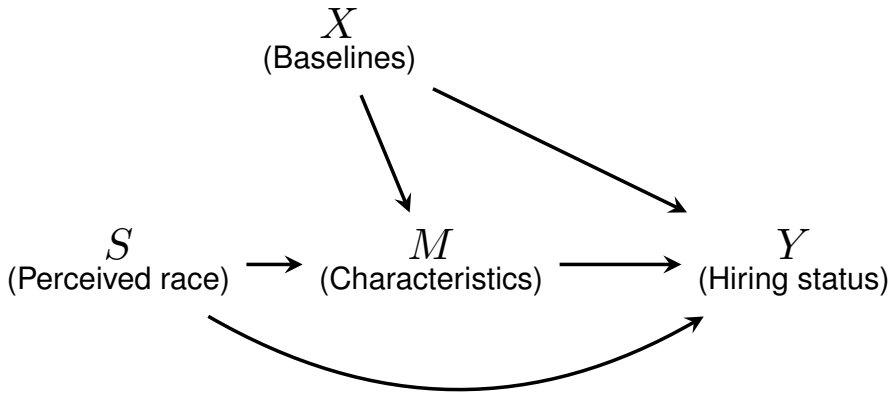
- Ignore the underlying DGP
- Ignore the role of policy makers, bioethicists, legal experts (one-fits-all definitions)
- Not adaptable to use context-specific information

Desirable definition of (un)fairness should:

- Use context-specific information
- Listen to causal relations
- Keep experts in the loop

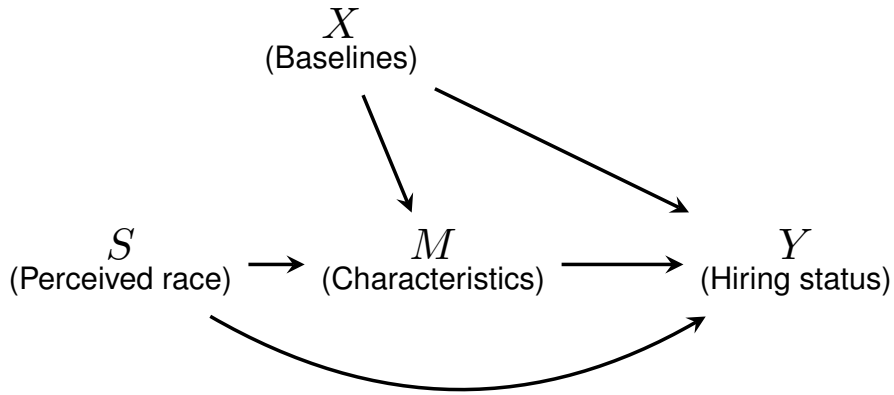
Mathematical expression of a legal passage

Would the employer have taken the same action had the employee been of a different race and everything else had remained the same?



Mathematical expression of a legal passage

Would the employer have taken the same action had the employee been of a different race and everything else had remained the same?

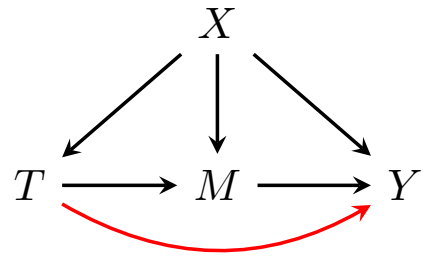


Name-swapping experiments to evaluate racism in hiring:

- African American: $S = 1$, Caucasian: $S = 0$,
- $Y(1, M(0))$: hiring a Caucasian with an African American sounding name
- $Y(0)$: hiring a Caucasian
- Direct effect: $\mathbb{E}\big[Y\big(1, M(0)\big)\big] - \mathbb{E}\big[Y(0)\big]$

Identification and estimation of the *direct effect*

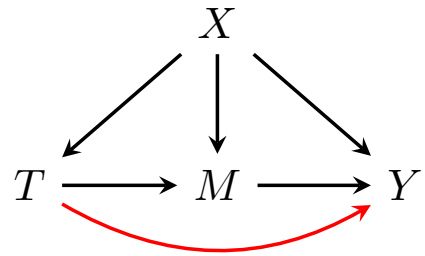
Direct effect $= \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0)]$



$$\mathbb{E}[Y(1, M(0))] = \mathbb{E}\Big[\sum_m \mathbb{E}[Y \mid x, m, T = 1] \times p(M = m \mid x, T = 0)\Big] = g(P_Z)$$

Identification and estimation of the *direct effect*

Direct effect $= \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0)]$



$$\mathbb{E}[Y(1, M(0))] = \mathbb{E}\Big[\sum_m \mathbb{E}[Y \mid x, m, T = 1] \times p(M = m \mid x, T = 0)\Big] = g(P_Z)$$

Plugin estimator: $\mathbb{P}_n\Big(\sum_m \widehat{\mathbb{E}}[Y \mid x_i, m, T = 1] \times \widehat{p}(M = m \mid x_i, T = 0)\Big)$

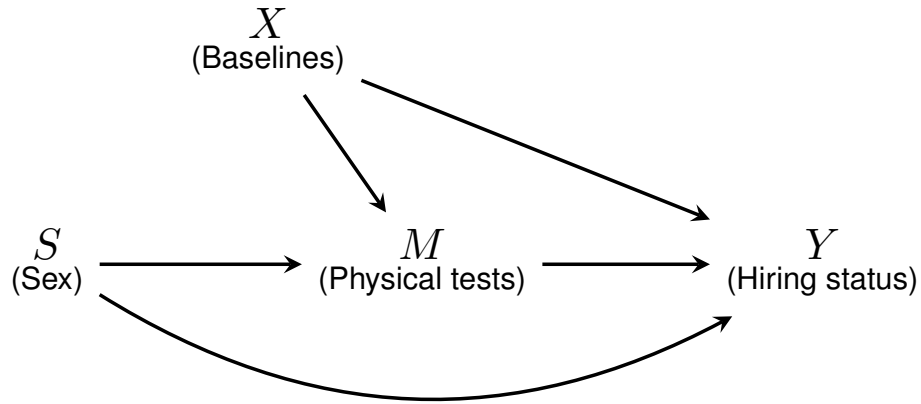
Inverse probability weighting: $p(T \mid X),\ p(M \mid X, T)$

Mixed estimator: $p(T \mid X),\ \mathbb{E}[Y \mid X, T, M]$

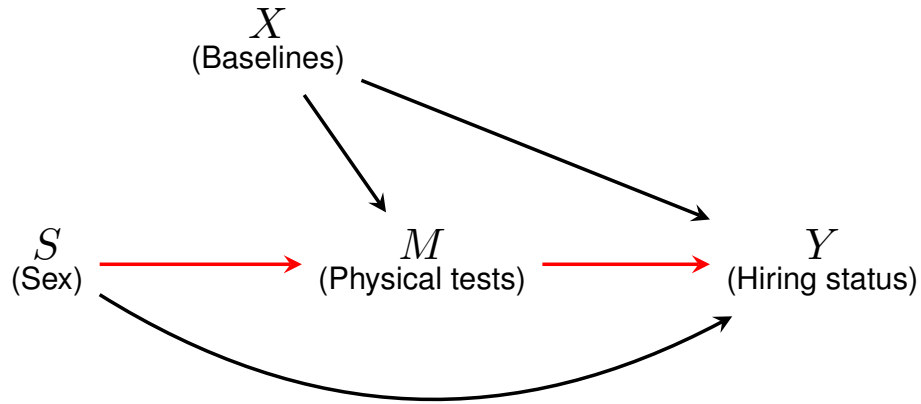
Augmented IPW: $p(T \mid X),\ p(M \mid X, T),\ \mathbb{E}[Y \mid X, T, M]$ (triply robust)

(Tchetgen Tchetgen and Shpitser, Annals of statistics, 2012)

Is unfairness always about the direct effect of S on Y?

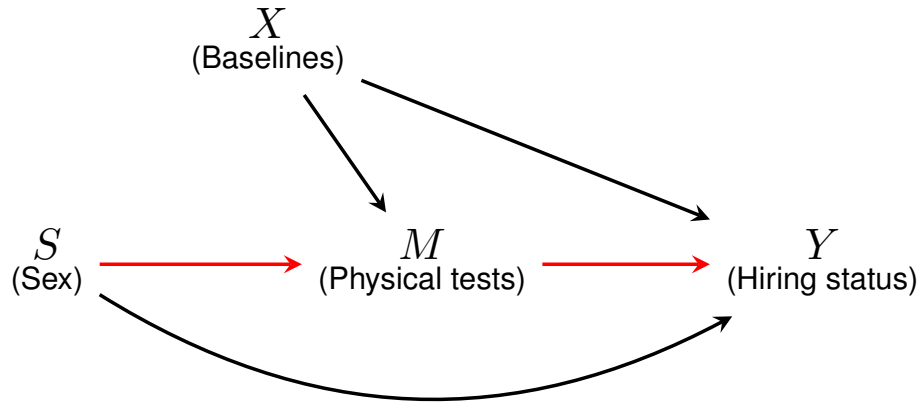


Is unfairness always about the direct effect of S on Y?



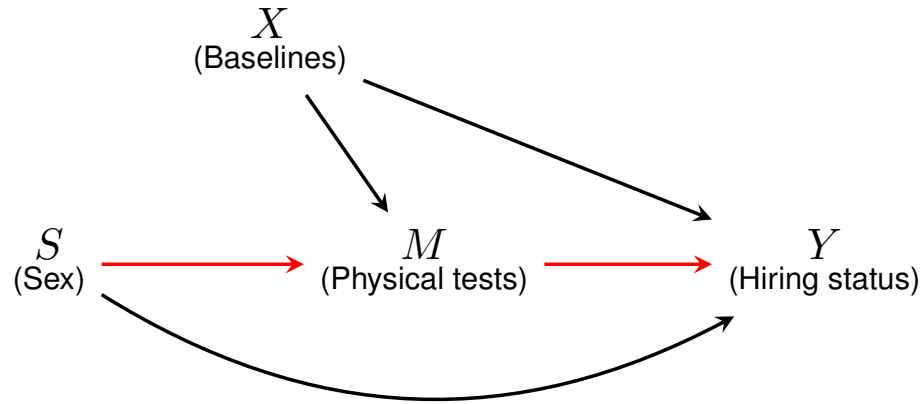
- Y : hiring a fire fighter
  - $S \rightarrow M \rightarrow Y$ ✓

Is unfairness always about the direct effect of S on Y?



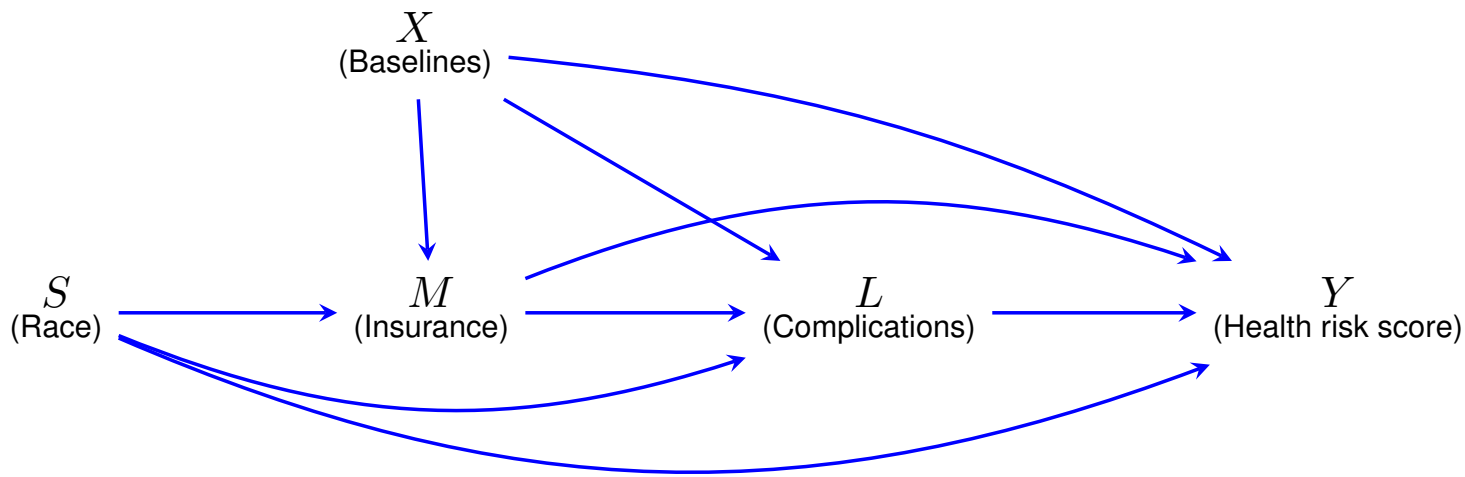
- Y : hiring a fire fighter
  - $S \rightarrow M \rightarrow Y$ ✓
- Y : hiring an accountant
  - $S \rightarrow M \rightarrow Y$ ✗

Is unfairness always about the direct effect of S on Y?

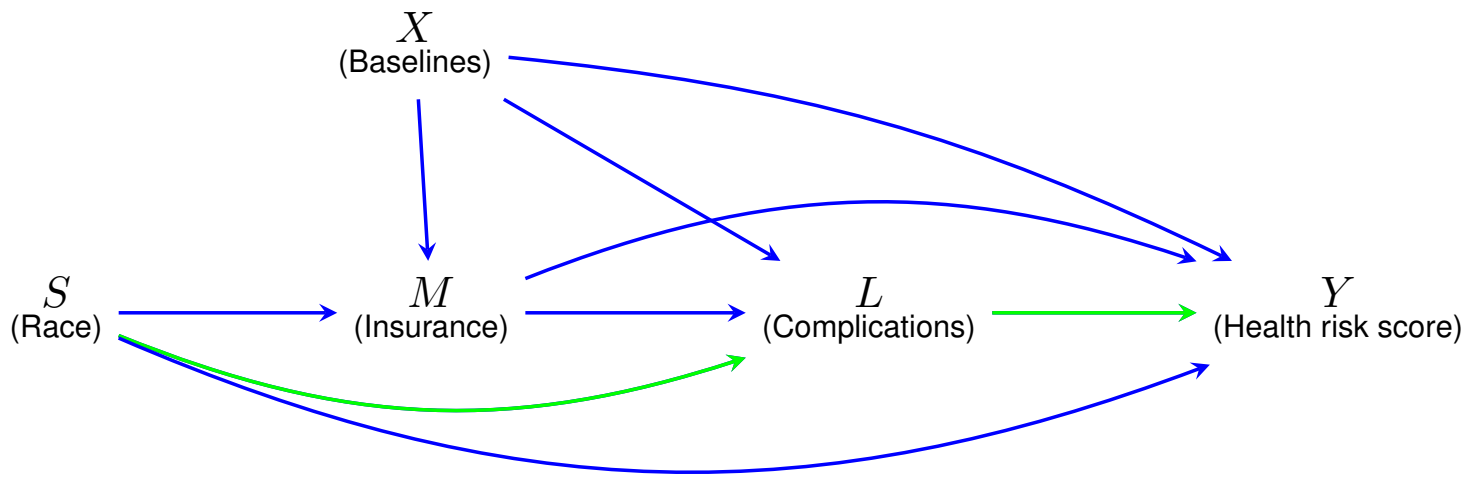


- Y : hiring a fire fighter
  - $S \rightarrow M \rightarrow Y$ ✓
- Y : hiring an accountant
  - $S \rightarrow M \rightarrow Y$ ✗
- So the answer is NO! Definition should be context-specifific.

From mediation to arbitrary path-specific effects

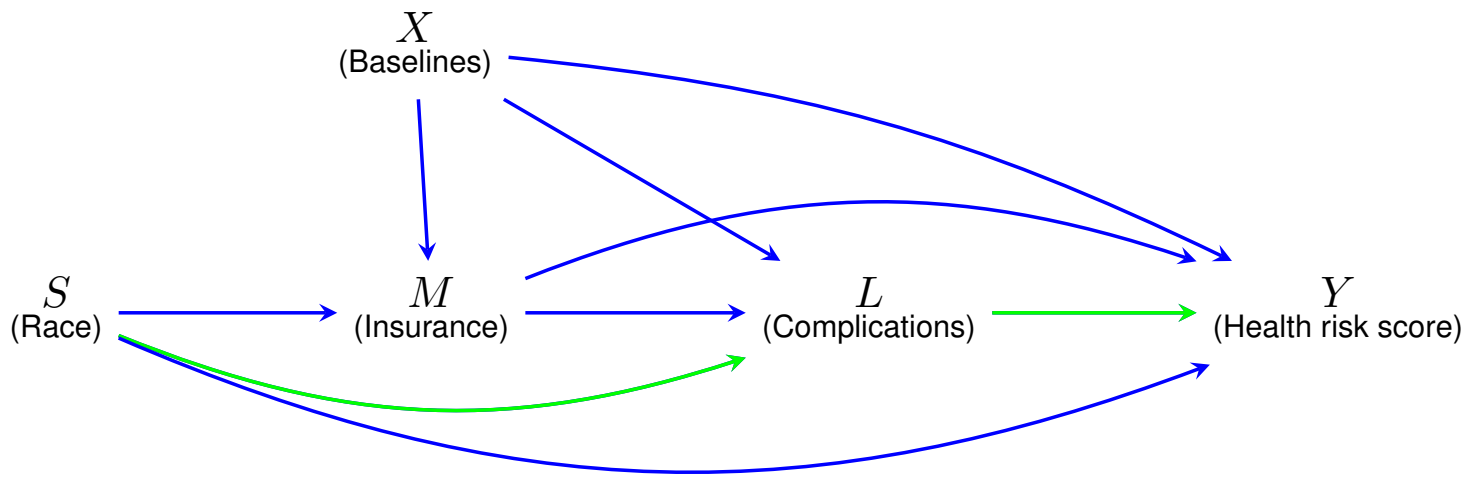


From mediation to arbitrary path-specific effects



- $S \rightarrow Y$ ✗
- $S \rightarrow M \rightarrow Y$ ✗
- $S \rightarrow M \rightarrow L \rightarrow Y$ ✗
- $S \rightarrow L \rightarrow Y$ ✓

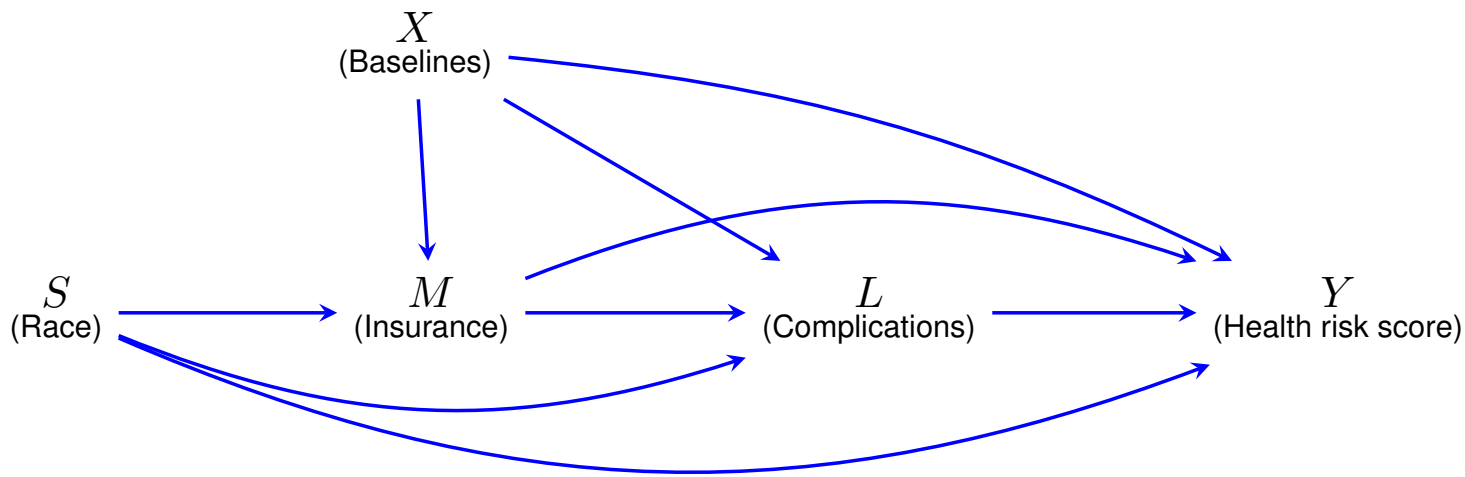
From mediation to arbitrary path-specific effects



- $S \rightarrow Y$ ✗
- $S \rightarrow M \rightarrow Y$ ✗
- $S \rightarrow M \rightarrow L \rightarrow Y$ ✗
- $S \rightarrow L \rightarrow Y$ ✓

$$\mathbb{E}\left[Y\left(s, M(s), L\left(s', M(s)\right)\right)\right] = g(P_Z)$$

From mediation to arbitrary path-specific effects



- Path-specific effect (PSE):
  - Along pathways of interest, all nodes behave as if $S = s$,
  - Along all other pathways, nodes behave as if $S = s'$.

- Identification and estimation of PSEs:
  (Shpitser, Tchetgen Tchetgen, VanderWeele, Avin, Pearl, Robins, Richardson, Malinsky, and more)

Our definition of fairness

- ACE = $\text{PSE}^{\text{unfair}}$ + PSE^{fair}
- $\text{PSE}^{\text{unfair}}$: effect of S on Y along unfair causal pathways

  (RN and Shpitser, Fair Inference on outcomes, AAAI, 2018.)
- Determining unfair pathways is a domain specific issue
  - This is a feature not a bug.

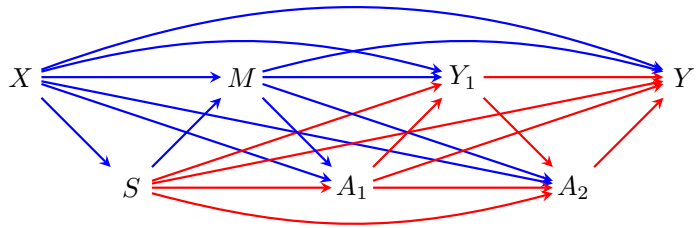
Our definition of fairness

- ACE = $\text{PSE}^{\text{unfair}}$ + PSE^{fair}
- $\text{PSE}^{\text{unfair}}$: effect of S on Y along unfair causal pathways

  (RN and Shpitser, Fair Inference on outcomes, AAAI, 2018.)
- Determining unfair pathways is a domain specific issue
  - This is a feature not a bug.
- Is this sufficient for defining fairness in automated descision making?

Fairness in automated decision making

Fairness in automated decision making



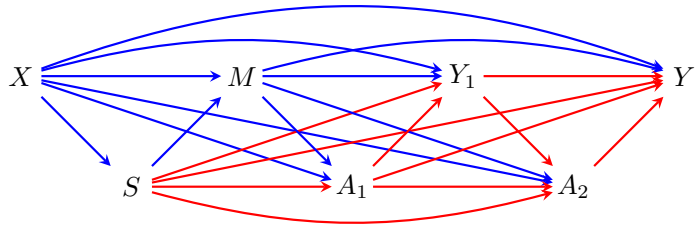
- **Retrospective bias**:
  bias in historical data used as input to learning procedure.

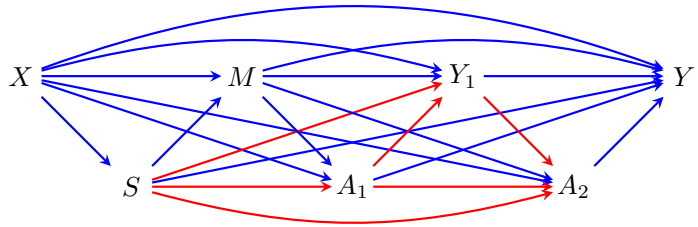
  Example: unfair paths from S to Y:
  $\{S \to Y, S \to Y_1 \to \cdots \to Y,$
  $S \to A_1 \to \ldots \to Y,$
  $S \to A_2 \to \ldots \to Y\}.$

$$\text{PSE}^{sy} = g_1(P_Z)$$

$$Z = \{X, S, M, A_1, ..., A_K, Y_1, ..., Y_K\}$$

Fairness in automated decision making





- **Retrospective bias**:
  bias in historical data used as input to learning procedure.

  Example: unfair paths from S to Y:
  $\{S \to Y, S \to Y_1 \to \cdots \to Y,$
  $S \to A_1 \to \ldots \to Y,$
  $S \to A_2 \to \ldots \to Y\}.$

- **Prospective bias**:
  functional form of policy depends on sensitive features.

  Example: unfair paths from S to A_1, A_2:
  $\{S \to A_1\},$
  and
  $\{S \to A_2, \quad S \to A_1 \to \ldots \to A_2\}$

$$\text{PSE}^{sy} = g_1(P_Z)$$

$$\text{PSE}^{sa_k} = g_k(P_Z)$$

$$Z = \{X, S, M, A_1, ..., A_K, Y_1, ..., Y_K\}$$

2. Fair world

Defining a fair world/distribution

- $p(Z)$: observed (unfair) distribution
- A set of identified unfair PSEs denoted by $g_j(P_Z)\ \forall j \in \{1, ..., J\}$
- $p^*(Z)$: fair distribution
  - A distribution where unfair effects are null
  - Close to $p(Z)$ via Kullback-Leibler divergence

Defining a fair world/distribution

- $p(Z)$: observed (unfair) distribution
- A set of identified unfair PSEs denoted by $g_j(P_Z)\ \forall j \in \{1, ..., J\}$
- $p^*(Z)$: fair distribution
  - A distribution where unfair effects are null
  - Close to $p(Z)$ via Kullback-Leibler divergence

- Give *lower/upper tolerance bounds* $\epsilon_j^-, \epsilon_j^+$, $p^*(Z)$ is defined as:

$$p^*(Z) \equiv \arg\min_q \ D_{KL}(p \,||\, q)$$

$$\text{subject to } \ \epsilon_j^- \leq g_j(P_Z) \leq \epsilon_j^+, \ \ \forall j \in \{1, ..., J\},$$

Approximating the fair world with finite samples

- Assume n iid samples $\sim p(Z)$
- Likelihood function: $\mathcal{L}(Z; \alpha)$
- Denote the estimator for $g(P_Z)$ via $\widehat{g}(P_Z)$

Approximating the fair world with finite samples

- Assume n iid samples $\sim p(Z)$
- Likelihood function: $\mathcal{L}(Z;\alpha)$
- Denote the estimator for $g(P_Z)$ via $\widehat{g}(P_Z)$

- Approximate $p^*(Z)$ by solving:

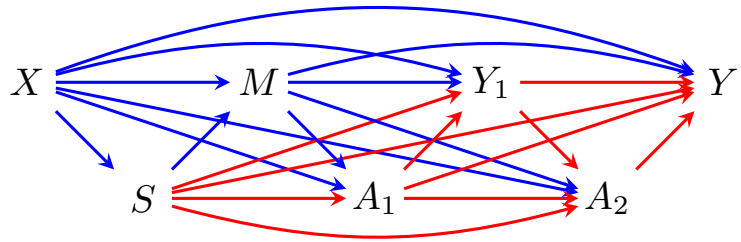
$$\widehat{\alpha} = \arg\max_{\alpha} \quad \mathcal{L}(Z;\alpha)$$
$$\text{such that} \quad \epsilon_j^- \leq \widehat{g}_j(P_Z) \leq \epsilon_j^+, \; j = 1, \ldots, J.$$

Example: a two-stage decision point

Approximate $p^*(Z)$ by solving:

$$\widehat{\alpha} = \arg\max_{\alpha} \quad \mathcal{L}(Z;\alpha)$$

$$\text{s.t.} \quad \epsilon_j^- \leq \widehat{g}_j(P_Z) \leq \epsilon_j^+, j = 1, \ldots, J.$$

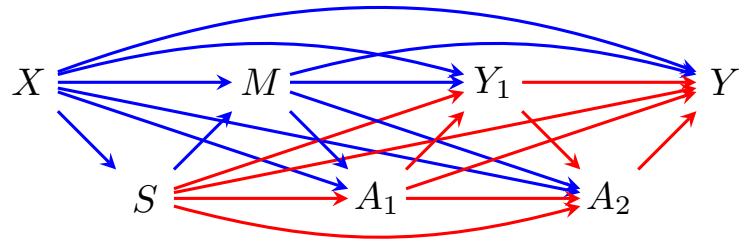


Example: a two-stage decision point

Approximate $p^*(Z)$ by solving:

$$\widehat{\alpha} = \arg\max_{\alpha} \quad \mathcal{L}(Z;\alpha)$$

$$\text{s.t.} \quad \epsilon_j^- \leq \widehat{g}_j(P_Z) \leq \epsilon_j^+, j = 1, \ldots, J.$$



Consistent estimators of PSE^{sy} and PSE^{sa_k}:

$$\widehat{g}^{sy}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\,\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} Y_n,$$

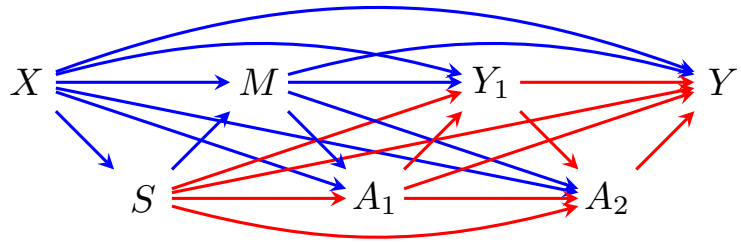
$$\widehat{g}^{sa_k}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\,\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} A_{kn}, \quad k = 1, 2.$$

Example: a two-stage decision point

Approximate $p^*(Z)$ by solving:

$$\widehat{\alpha} = \arg\max_{\alpha} \quad \mathcal{L}(Z;\alpha)$$

$$\text{s.t.} \quad \epsilon_j^- \leq \widehat{g}_j(P_Z) \leq \epsilon_j^+, j = 1, \ldots, J.$$



Consistent estimators of PSE^{sy} and PSE^{sa_k}:

$$\widehat{g}^{sy}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\,\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} Y_n,$$

$$\widehat{g}^{sa_k}(Z) = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{\mathbb{I}(S_n = s)}{p(S_n|X_n)}\,\frac{p(M_n|s', X_n)}{p(M_n|s, X_n)} - \frac{\mathbb{I}(S_n = s')}{p(S_n|X_n)}\right\} A_{kn}, \quad k = 1, 2.$$

Constraints involve $p(S \mid X; \alpha_s)$ and $p(M \mid S, X; \alpha_m)$ models.

Breaking the cycle of injustice

- Let $p^*(M \mid S, X; \alpha_m)$ and $p^*(S \mid X; \alpha_s)$ be the constrained models chosen to satisfy $\text{PSE}^{sy} = \text{PSE}^{sa_1} = \text{PSE}^{sa_2} = 0$

Breaking the cycle of injustice

- Let $p^*(M \mid S, X; \alpha_m)$ and $p^*(S \mid X; \alpha_s)$ be the constrained models chosen to satisfy $\text{PSE}^{sy} = \text{PSE}^{sa_1} = \text{PSE}^{sa_2} = 0$
- Let $\widetilde{p}(Z)$ be the joint distribution induced by $p^*(M|S, X; \alpha_m)$ and $p^*(S|X; \alpha_s)$:

$$\widetilde{p}(Z) \equiv p(X)p^*(S|X; \alpha_s)\, p^*(M|S, X; \alpha_m) \prod_{k=1}^{K} p(A_k|H_k)p(Y_k|A_k, H_k).$$

Breaking the cycle of injustice

- Let $p^*(M \mid S, X; \alpha_m)$ and $p^*(S \mid X; \alpha_s)$ be the constrained models chosen to satisfy $\text{PSE}^{sy} = \text{PSE}^{sa_1} = \text{PSE}^{sa_2} = 0$
- Let $\widetilde{p}(Z)$ be the joint distribution induced by $p^*(M|S, X; \alpha_m)$ and $p^*(S|X; \alpha_s)$:

$$\widetilde{p}(Z) \equiv p(X)p^*(S|X; \alpha_s)\, p^*(M|S, X; \alpha_m) \prod_{k=1}^{K} p(A_k|H_k)p(Y_k|A_k, H_k).$$

- Then PSE^{sy} and PSE^{sa_k} taken wrt $\widetilde{p}(Z)$ are also zero.

  $\implies$ constraining the S and M models induces a "fair distribution" no matter how A_k or Y_k are determined.
- This enables us to choose our optimal decision rules without restricting the policy space and also allowing for the mechanisms that determine outcomes to remain outside the control of the policy-maker.

3. Optimal policies in the fair world

Three strategies for policy estimation

We consider three strategies for estimating the optimal policy:

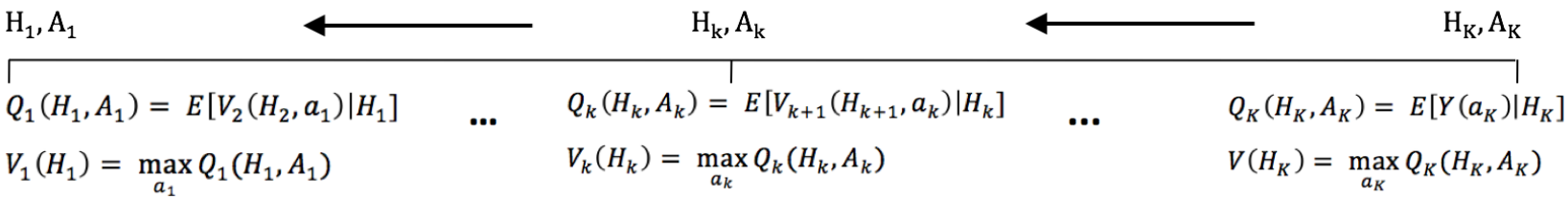
- Q-learning
- Value search
- G-estimation

In each case, we must modify these procedures to operate wrt the fair distribution.

We focus on Q-learning and value search in this talk.

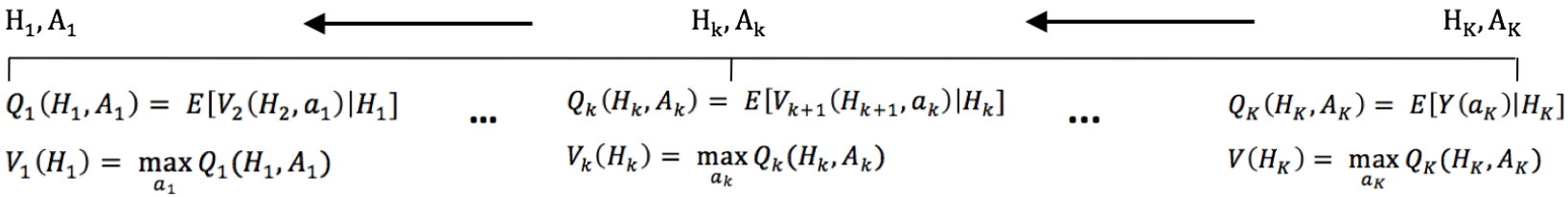
Optimal fair policy: Q-learning

- **Unfair world**: expectations wrt to $p(Z)$



Optimal fair policy: Q-learning

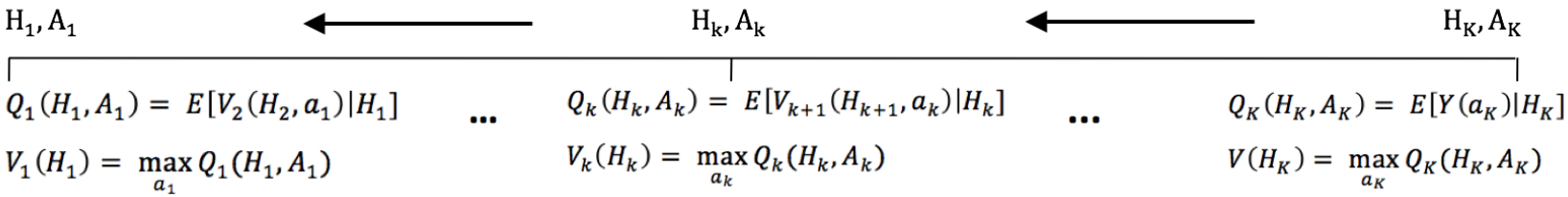
- **Unfair world**: expectations wrt to $p(Z)$



- **Optimal policy**: $f^*_{A_k} = \arg\max_{a_k} Q_k(H_k, a_k; \beta_k)$

Optimal fair policy: Q-learning

- **Unfair world**: expectations wrt to $p(Z)$



- **Optimal policy**: $f^*_{A_k} = \arg\max_{a_k} Q_k(H_k, a_k; \beta_k)$

- **Fair world** expectations wrt to $p^*(Z)$

$$\widetilde{Q}_k(H_k \setminus \{M, S\}, A_k) = \frac{1}{Z_k} \sum_{m,s} Q^*_k(H_k, A_k)\, p^*(m|X, s; \alpha_m)\, p^*(s|X; \alpha_s),$$

for $k = 1, \ldots, K$.

Optimal fair policy: Value search

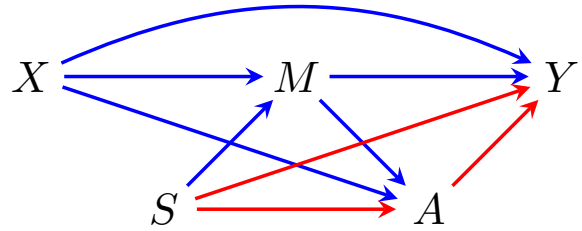
- **Optimal policy**: $f_A^* = \arg\max_{f_A} \mathbb{E}[Y(f_A)]$
- **Unfair world**: expectations wrt to $p(Z)$

$$\mathbb{E}[Y(f_A)] = \mathbb{E}\Big[\frac{\mathbb{I}(A_1 = f_{A_1}(H_1))}{p(A_1|H_1;\psi)} \times \frac{\mathbb{I}(A_2 = f_{A_2}(H_2))}{p(A_2|H_2;\psi)} \times Y\Big],$$

- **Fair world**: expectations wrt to $p^*(Z)$

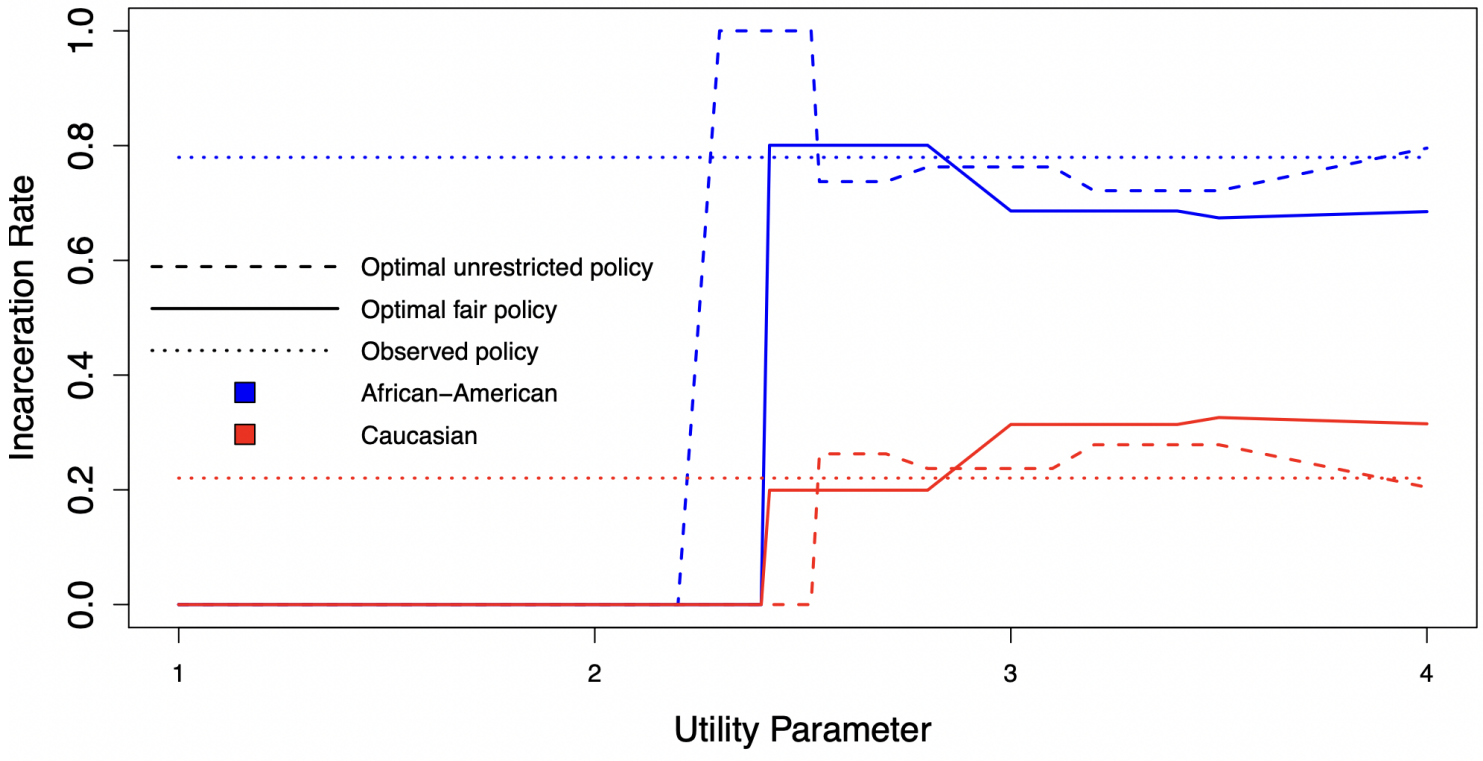
$$\widetilde{\mathbb{E}}[Y(f_A)] = \frac{1}{Z}\sum_{m,s} \mathbb{E}[Y(f_A)]\; p^*(m|X,s;\alpha_m)\; p^*(s|X;\alpha_s)$$

COMPAS application



- S: race, X: other demographics, M: prior convictions
- A: incarceration (based on risk of recidivism)
- Heuristic utility: $Y \equiv (1-A) \times \{\theta R + (1-R)\} - A$
  - R: whether or not recidivism occurred in a span of two years
  - Negative utility (social, economical costs) associated with incarceration $A = 1$.
  - Some cost to releasing individuals who go on to reoffend (i.e., for whom $A = 0$ and $R = 1$) controlled by θ
  - Positive utility associated with releasing individuals who do not go on to recidivate (i.e., for whom $A = 0$ and $R = 0$)

COMPAS application ctd.



Question: What would be the resulting difference in pre-trial incarceration rate under a “fair” vs. unconstrained optimal policy?

Result: “fair” vs. unconstrained policies differ, and incarceration rates depend crucially on the utility function.

Summary

1. **How to express fairness principles mathematically?**
   - The approach we take requires substantive *ethical* input from experts and/or the public.
   - We also require specifying a causal model (based on domain knowledge or causal structure learning).
   - Dealing with unidentified causal effects (use of bounds)

2. **How to modify statistical procedures to reduce unfair effects?**
   - Constrained MLE (hybrid likelihood)
   - Developing more robust constrained optimization methods to use data as efficiently as possible

3. **How to generalize and deploy these modified algorithms?**
   - Be mindful of the fact that samples are collected in p and not p^*
   - Find more effective approaches to map instances between p and p^*

Relevant papers:

- RN and Shpitser, Fair Inference on Outcomes, AAAI 2018.
- RN, Malinsly, and Shpitser, Learning Optimal Fair Policies, ICML 2019.
- RN, Malinsly, and Shpitser, Optimal Training of Fair Predictive Models, CLeaR 2022.

Sincerely,

Razieh Nabi, PhD (she/her/hers)
Rollins Assistant Professor
Department of Biostatistics and Bioinformatics
Emory University
razieh.nabi@emory.edu
https://raziehnabi.com