Fairness and Causality

Shira Mitchell
Statistician at Mathematica Policy Research

April 10, 2018

Setup

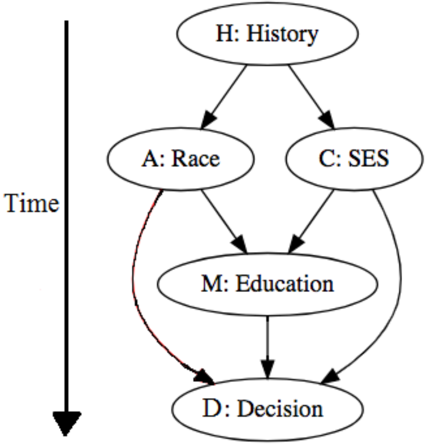
Suppose $\mathrm{P}(\text{hired}|\text{black}) < \mathrm{P}(\text{hired}|\text{white})$.

1. Is this disparity unfair?
2. Can we reduce this disparity?

To answer, we need to explain: $\mathrm{P}(\text{hired}|\text{white}) - \mathrm{P}(\text{hired}|\text{black})$.

Goal: explain disparity

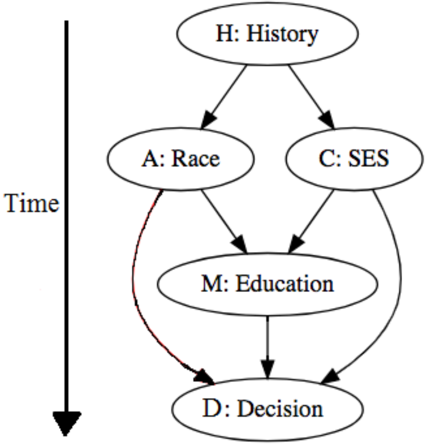
Consider a very simplified world (VanderWeele and Robinson, 2014):



Vaguely speaking: arrows represent possible causal relationships

Goal: explain disparity

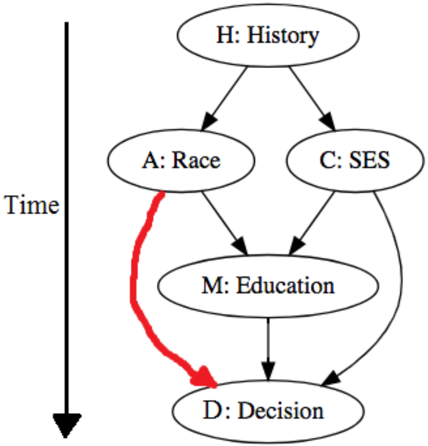
Q: What can explain $\mathrm{P}(\mathrm{D} = \text{hired}|A = \text{white}) - \mathrm{P}(\mathrm{D} = \text{hired}|A = \text{black})$?



Goal: explain disparity

What can explain $\mathrm{P}(D = \text{hired}|A = \text{white}) - \mathrm{P}(D = \text{hired}|A = \text{black})$?

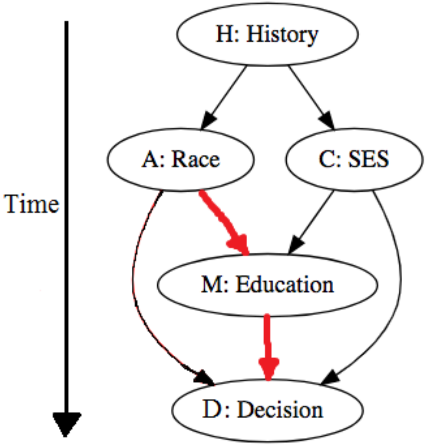
A direct effect (prejudice)



Goal: explain disparity

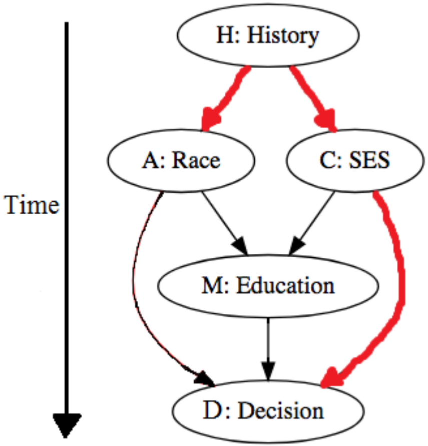
What can explain $\mathrm{P}(\mathrm{D} = \mathrm{hired}|A = \mathrm{white}) - \mathrm{P}(\mathrm{D} = \mathrm{hired}|A = \mathrm{black})$?

An indirect effect through education



Goal: explain disparity

What can explain $\mathrm{P}(\mathrm{D} = \text{hired}|A = \text{white}) - \mathrm{P}(\mathrm{D} = \text{hired}|A = \text{black})$?
Correlation through history (not a causal path from race to decision)



Goal: explain disparity

Zhang and Bareinboim (2018) show how to decompose the disparity:

$$\begin{aligned}\text{Disparity} &\equiv \mathsf{P}(\mathrm{D} = \text{hired}|\mathsf{A} = \text{white}) - \mathsf{P}(\mathrm{D} = \text{hired}|\mathsf{A} = \text{black})\\ &= \text{direct effect}\\ &\quad + \text{indirect effect through education}\\ &\quad + \text{correlation through history}\end{aligned}$$

and when/how we can estimate each piece.

Back to our questions

$$\begin{aligned}\text{Disparity} = {} & \text{direct effect} \\ & + \text{indirect effect through education} \\ & + \text{correlation through history}\end{aligned}$$

1. Is this disparity unfair?

Person A All 3 are unfair, I don't need the decomposition to say "yes".

Person B The indirect effect through education is ok, so I need the decomposition to answer.

Q: Do you agree with Person A or B?

2. Can we reduce this disparity?

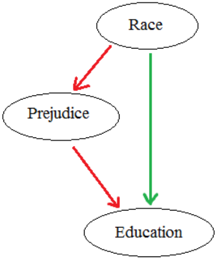
Use the decomposition to see where to focus policy/activism.

Fair paths: resolving variables?

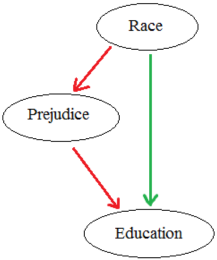
Person B	A hiring process should be allowed to use education, that variable is fair game.
Kilbertus et al. (2018)	To Person B, education is a *resolving variable*. Paths are fair if through resolving variables.
Nabi and Shpitser (2018)	*Any path* can be fair or unfair.

Q: Is the Nabi and Shpitser (2018) definition more flexible?

Nabi and Shpitser (2018) Can decide that this path is ok, but this is not.

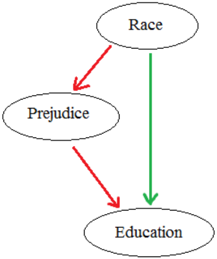


Nabi and Shpitser (2018) Can decide that this path is ok, but this is not.

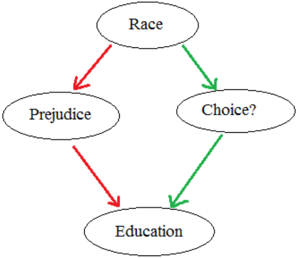


Kilbertus et al. (2018) Can't.

Nabi and Shpitser (2018) Can decide that this path is ok, but this is not.



Kilbertus et al. (2018) trick: add a new variable, call it resolving, and also decide that this path is ok, but this is not.

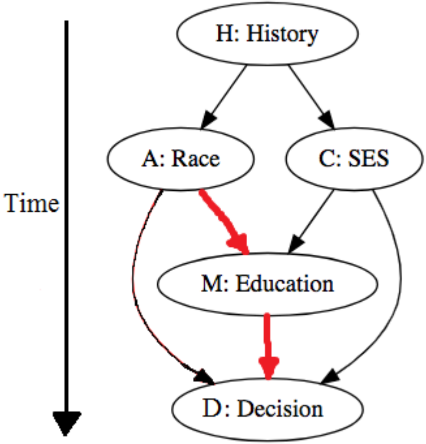


Fair paths: resolving variables?

But maybe prejudice and choice aren't measured. We need to decide if using education in hiring is fair.

Person A Meritocracy strengthens existing social/economic hierarchies.

Person B But come on, an employer should be allowed to use education.



What do the arrows mean?

Fix a variable order (say, by time): H, A, C, M, D.

Probability of particular values:
$P(H = h, A = a, C = c, M = m, D = d)$, abbreviated $P(h, a, c, m, d)$.

By rules of probability:
$P(h, a, c, m, d) = P(h)\ P(a|h)\ P(c|h, a)\ P(m|h, a, c)\ P(d|h, a, c, m)$

Suppose:

- Given history of exposure to policies, class is independent of race.
- Given a person's race and class, education and hiring are independent of history.

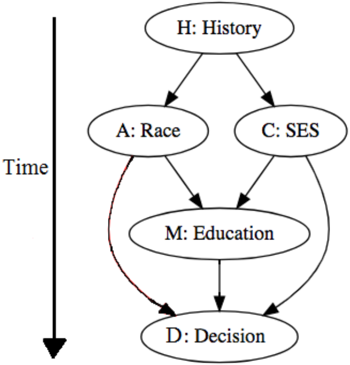
$P(h, a, c, m, d) = P(h)\ P(a|h)\ P(c|h, a)\ P(m|h, a, c)\ P(d|h, a, c, m)$

What do the arrows mean?

Draw these arrows to get a *directed acyclic graph* (DAG, or *Bayesian Network*):

$$P(h, a, c, m, d) = P(h)\ P(a|h)\ P(c|h, a)\ P(m|h, a, c)\ P(d|h, a, c, m)$$

This gives our graph:



So far, nothing causal yet.

What do the arrows mean?

Causal Bayesian Networks: tell us what happen under interventions.

Suppose I add "PhD" to a person's resume.
What is $P(h, a, c, m, d \mid do(M = PhD))$?

Q: How is this different from $P(h, a, c, m, d \mid M = PhD)$?

What do the arrows mean?

Causal Bayesian Networks: tell us what happen under interventions.

Suppose I add "PhD" to a person's resume.
What is $P(h, a, c, m, d \mid do(M = PhD))$?

Q: How is this different from $P(h, a, c, m, d \mid M = PhD)$?
Doing versus seeing.

What do the arrows mean?

$P(h, a, c, m, d \mid do(M = \mathrm{PhD}))$ abbreviated as $P_{\mathrm{PhD}}(h, a, c, m, d)$.

Our graph is a *Causal* Bayesian Network if

1. It still gives the factorization:

$$P_{\mathrm{PhD}}(h, a, c, m, d) = P_{\mathrm{PhD}}(h)\ P_{\mathrm{PhD}}(a|h)\ P_{\mathrm{PhD}}(c|h)\ P_{\mathrm{PhD}}(m|a, c)\ P_{\mathrm{PhD}}(d|a, c, m)$$

2. $P_{\mathrm{PhD}}(\mathrm{PhD}|a, c) = 1$, i.e. we succeeded in setting education to PhD.
3. For all other variables, P and P_{PhD} are the same as long as we condition on the variable's parents.

Given these, the factorization is
$P_{\mathrm{PhD}}(h, a, c, d) = P(h)\ P(a|h)\ P(c|h)\ P(d|a, c, \mathrm{PhD})$

(Pearl, 2009, p.23-24)

References

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Scholkopf, B. (2018). Avoiding discrimination through causal reasoning. *31st Conference on Neural Information Processing Systems (NIPS 2017)*.

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. *Proceeding of the 32nd AAAI Conference on Artificial Intelligence*.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, second edition edition.

VanderWeele, T. J. and Robinson, W. R. (2014). On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology*, 25(4):473–484.

Zhang, J. and Bareinboim, E. (2018). Fairness in decision-making - the causal explanation formula. *Proceeding of the 32nd AAAI Conference on Artificial Intelligence*.