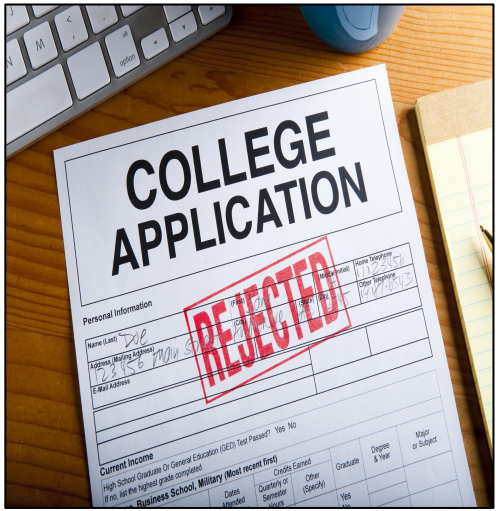
Path-Specific Counterfactual Fairness

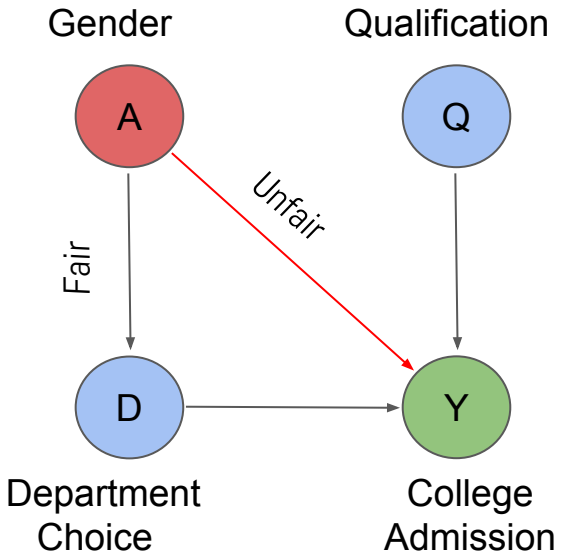
Silvia Chiappa csilvia@google.com

DeepMind

Motivation

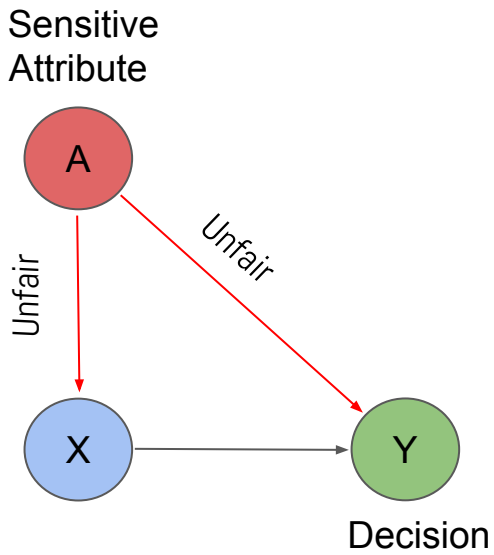
Design ML decision system from data in which a sensitive attribute A affects the decision Y through fair and unfair pathways





Most Existing Approaches considering Unfairness in the Training Data

All pathways are unfair

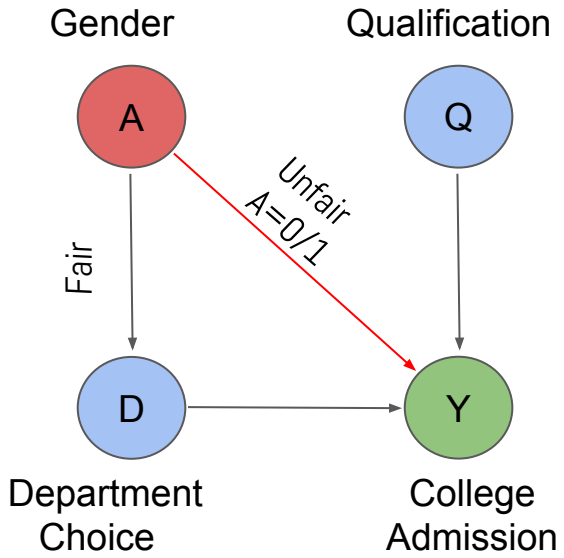


WHY?

Cannot compute the influence separately along the different pathways with standard Machine Learning

…. but we can with causal inference from observations!

Path-Specific Fairness



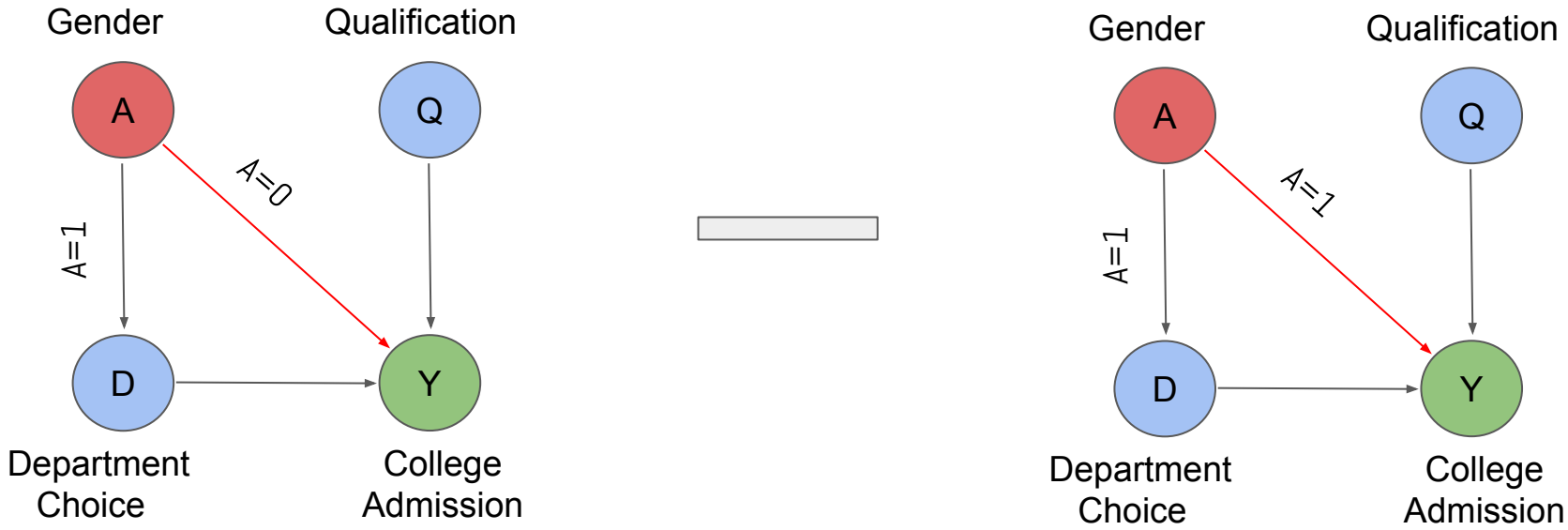
Making female (A=1) candidates males (A=0) along A ⟶ Y should not change the admission decisions.

We want Y_0_fair, defined as the random variable resulting from constrained conditioning of Y on A=1 along A ⟶ D ⟶ Y, to be independent on A.

Simulated intervention (action) A=0/1 along A ⟶ Y.

Fair Inference on Outcomes (FIO), Nabi and Shpitser 2018

Path-Specific Effect (PSE)

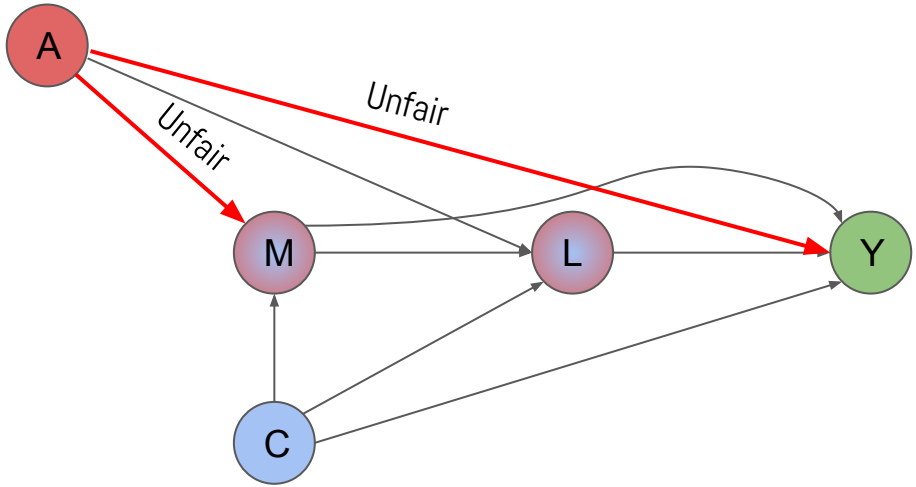


Do constrained training such that the PSE is enforced to be small.

Fair Inference on Outcomes, Nabi and Shpitser 2018

Issues:

- Integrate out L and M when taking a decision (at test time): Loss in performance



- Does not necessarily ensure each individual is treated fairly

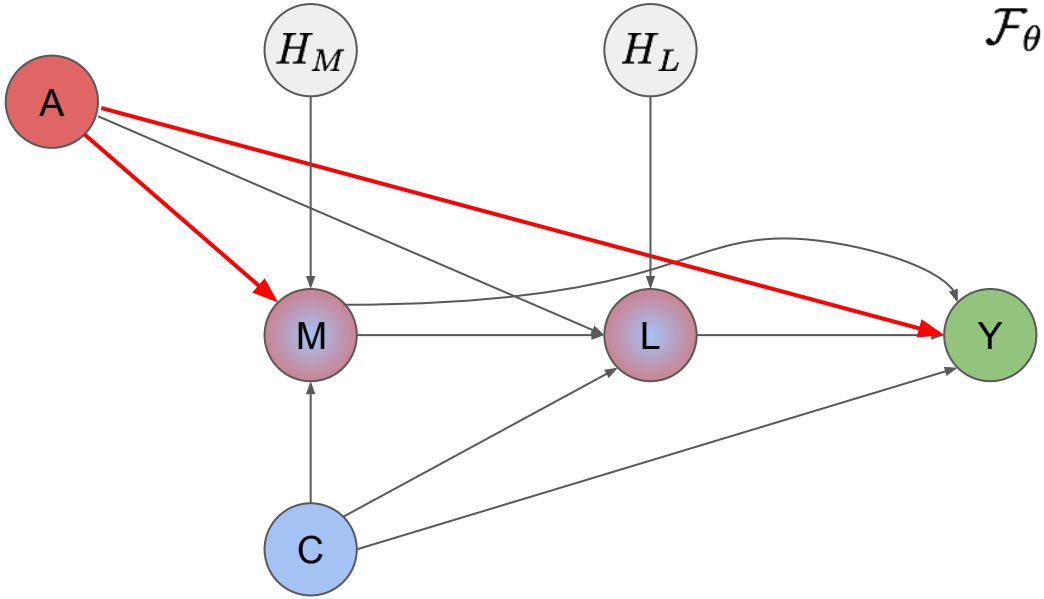
Path-Specific Counterfactual Fairness

- Training Time: Train in an unconstrained way.
- Test Time: Correct admission decision of a specific female individual by pretending that the individual is male (A=0) along the unfair pathways.

Path-Specific Counterfactual Fairness

Introduce latent space:

Capture all non-sensitive information



Train as a penalized VAE

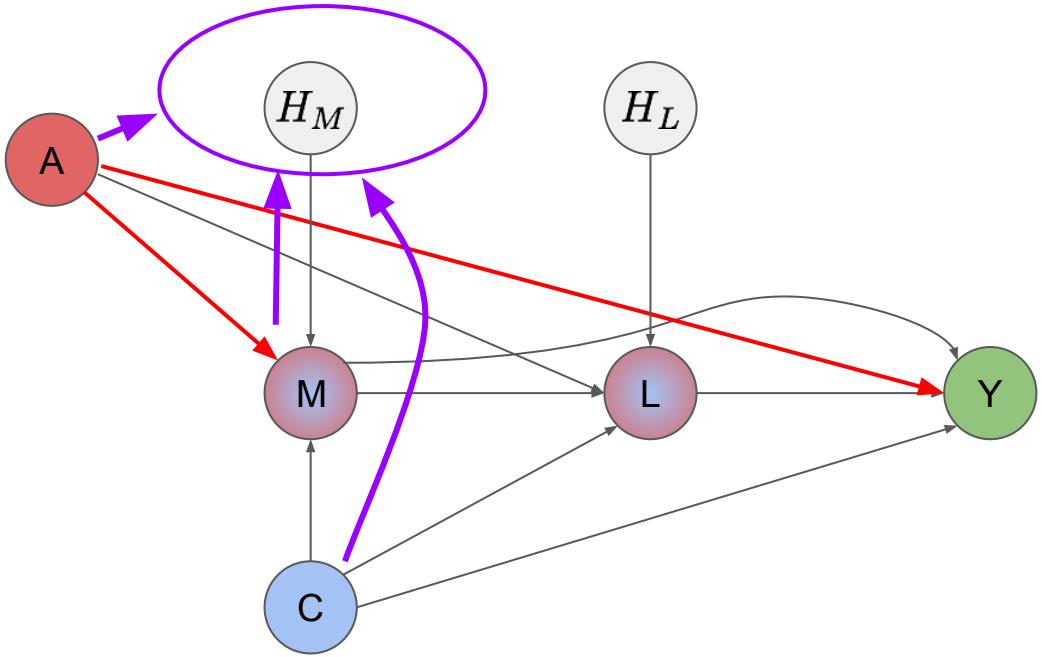
$$\mathcal{F}_{\theta,\phi} - \beta \mathcal{L}_{\text{MMD}}(A = 0, A = 1)$$

$$q(H_M|M, A, C)$$
$$q(H_L|L, M, A, C)$$

Path-Specific Counterfactual Fairness

Introduce latent space:

Capture all non-sensitive information

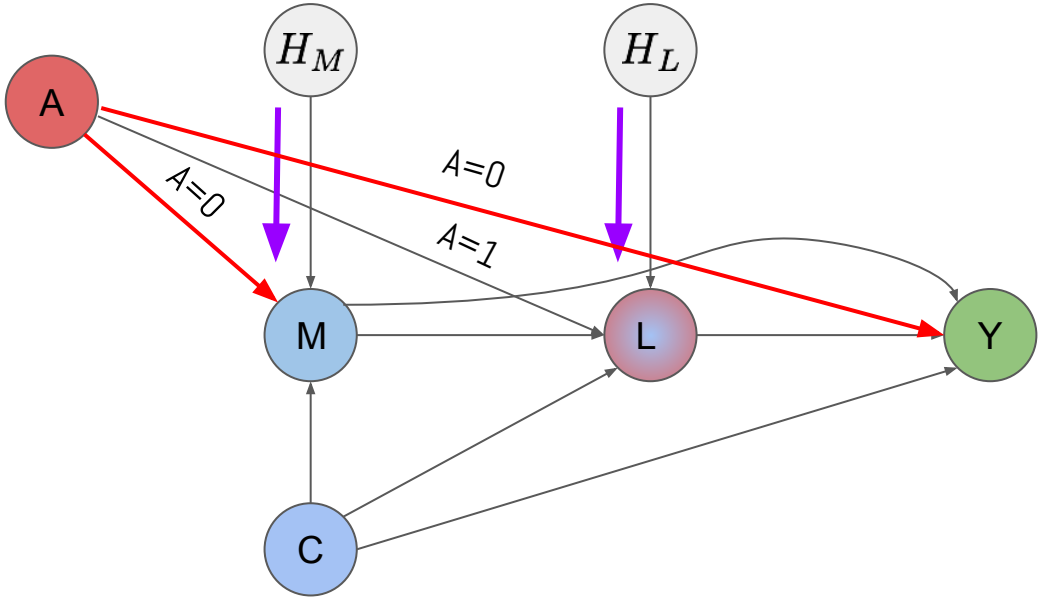


Test Time I: (For each female individual) estimate the latent variables given the observations

Path-Specific Counterfactual Fairness

Introduce latent space:

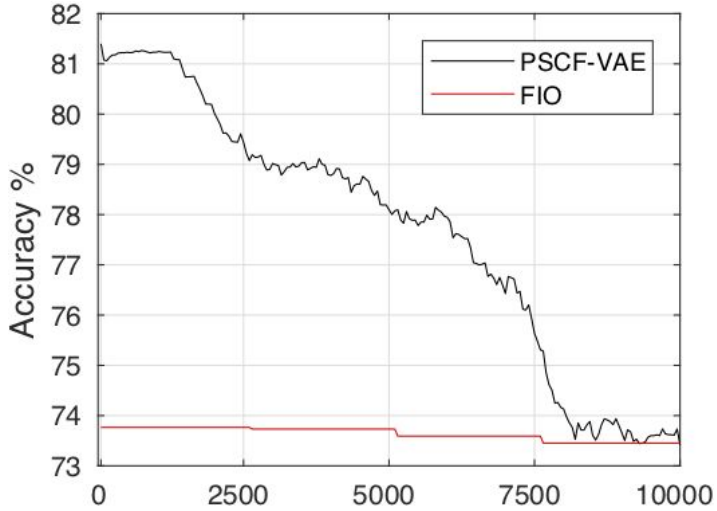
Capture all non-sensitive information



Test Time II: Project back using A=0 along unfair pathways
---> Generate corrected version of M and L and Y.

Results

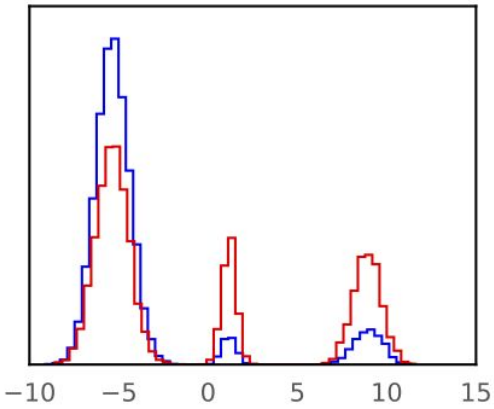
Adult Dataset

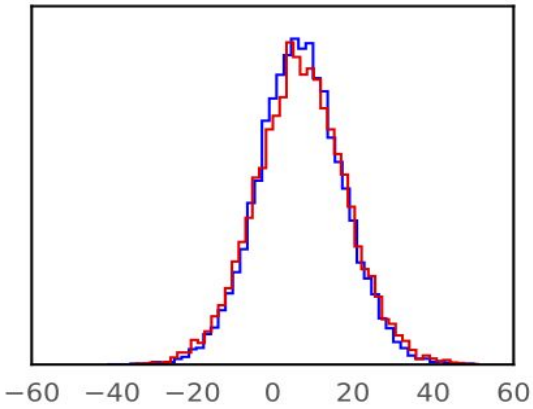


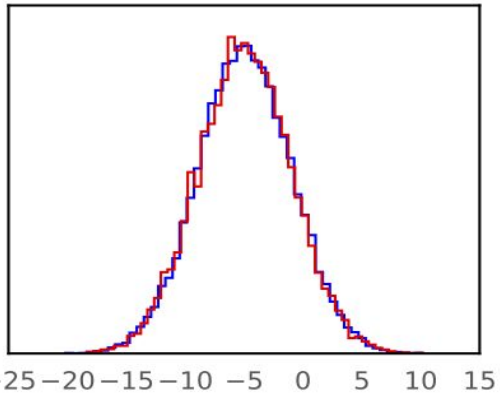
German Dataset

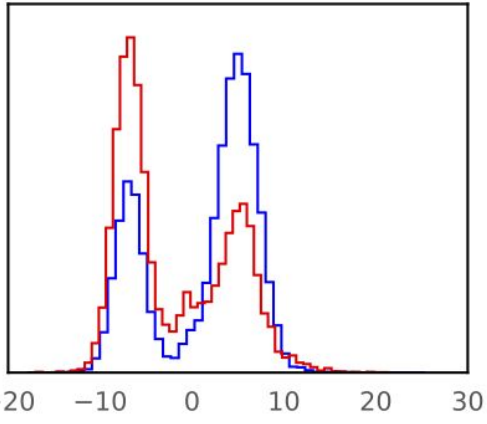
- PSCF: From 76% to 76% accuracy irrespective of penalization
- Nabi and Shpitser: From 70% to 70% accuracy irrespective of constraint

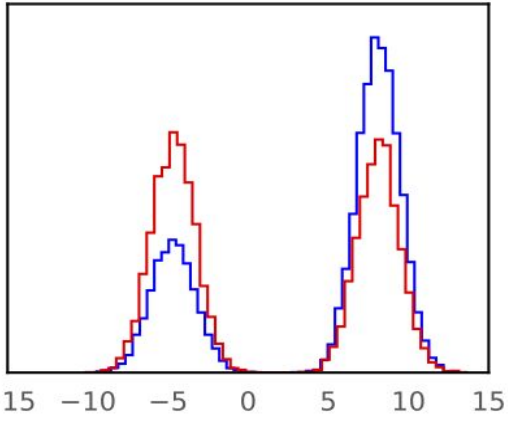
Latent Space

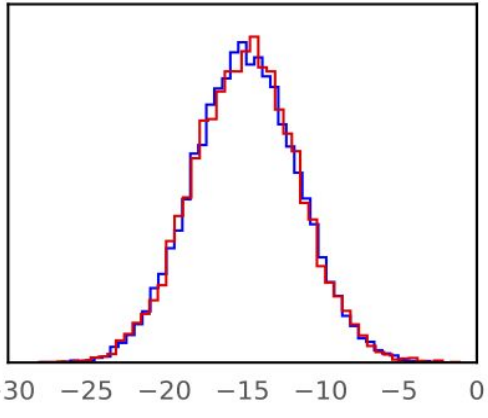












Relation to Other Work

- Counterfactual Fairness, Kusner at al. 2017
  - Does not deal with path-specific unfairness.
  - Does not enforce independence in latent space.
  - Does not provide a method for inferring latent variables.
- Fair Inference on Outcomes, Nabi and Shpitser 2018
  - Integrate out L and M: Loss in performance.
  - Difficult to explicitly compute the path-specific effect.
  - Group-fairness.
- Avoiding Discrimination through Causal Reasoning, Kilbertus 2017
  - Linear relationships among variables.
  - Group-fairness.