

Counterfactual Reasoning in Algorithmic Fairness

Ricardo Silva

University College London and
The Alan Turing Institute

Joint work with Matt Kusner (Warwick/Turing), Chris Russell (Sussex/Turing), and Joshua Loftus (NYU)

Fairness and Machine Learning

- The dream: if we teach machines to perform sensitive decisions, they will not suffer from human biases.
- The reality: the GIGO principle still holds, regardless of whether we are talking of statistical models or software.

The Message

- There is only so much data alone can tell you about fairness.
- I'm not talking about “just” *value judgments*.
- We should highlight the role that the *data-generating causal process* has in shaping our notions of fairness.

Nobody is Saying This is Easy

- At no point I will suggest that building a causal model is easy.
- Some *untested* and *untestable* assumptions will be needed.
- The idea is to make your assumptions as explicit as possible, hopefully being “less wrong” in the end.

The Scope of this Talk

- We consider *prediction* and *intervention* problems (more of the former).
- In prediction problems, we will have:
 - X : *features*, or attributes of an individual
 - A : the *protected attributes* of an individual
 - Y : the *target*, what we would like to predict
 - \hat{Y} : our *prediction*

Prediction Problems

- *Prediction* here means inferring a property Y that will be used for *decision* making.
- For example:
 - $Y = 1$ means “this person will default on a loan” (for the decision, “should I give this person a loan”?)
 - $Y = 1$ means “this person will commit a crime in two years” (for the decision, “should I release this convict now?”)
- We would like to predict Y in a “fair” way, meaning that our predictions should not be “biased” against particular instances of A .

Primitives

- Even if we take *the choice of what goes in A as a primitive*, it is still not obvious what we mean by being fair.
- A first idea: “ensure that \hat{Y} does not use A”.
- This is known to be unsatisfactory.

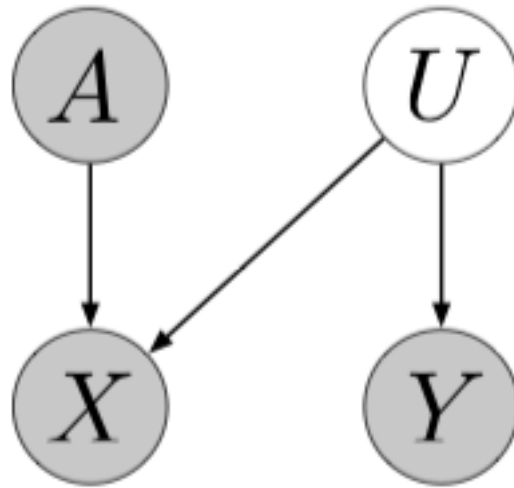
Examples

- *Equalized odds*: given the outcome Y , attribute A provides no further information about my prediction \hat{Y} .
- *Calibration*: given my prediction \hat{Y} , attribute A provides no further information about the outcome Y .
- If A is on average informative of Y , we cannot reconcile the above.
 - Remember, here we do *not* control Y (directly). We decide on predictor \hat{Y} .

Putting It in the Context of a Causal Model

- A toy model: imagine A is race, X is “owns a red car” and Y is “crashes car in one year”.
- Let’s (informally) draw a *causal diagram* showing cause-effect relationships among those. It will include a “*unobserved* trait” U measuring *aggressiveness*.
 - We will get into more formal definitions later.

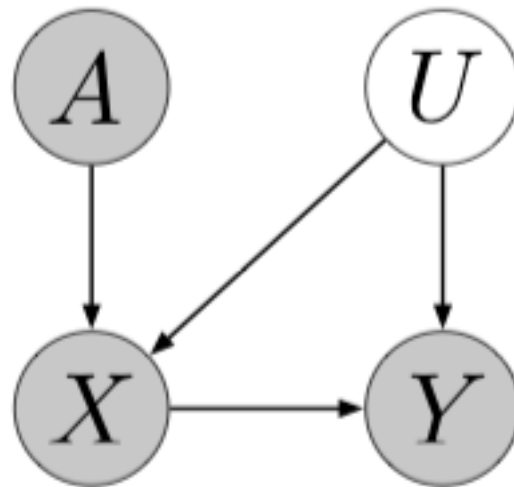
A Causal Diagram



Some Initial Conclusions

- A is not a cause of Y .
- If we build a predictor based on X , it tells us something both about A and about U .
- Hence, our predictor will be different for different values of A , which does not seem appropriate.

A Second Causal Diagram

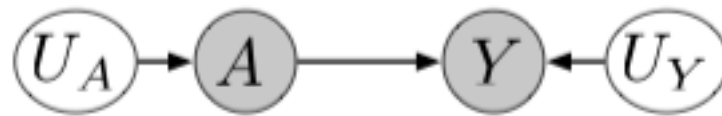


Which Conclusions?

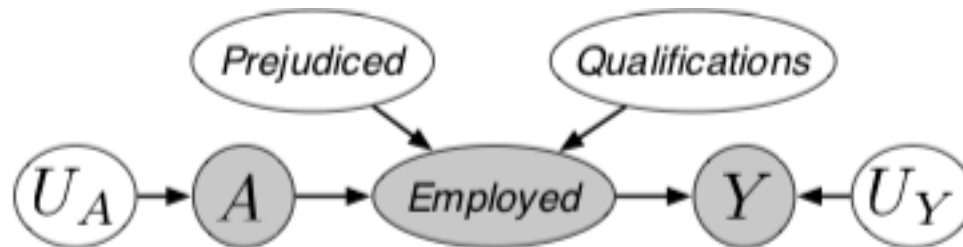
- A is now a cause of Y (indirectly).
- It is now impossible to satisfy both equalized odds and calibration simultaneously.
- Judgment call: is the *pathway* $A \rightarrow X \rightarrow Y$ “fair”?

Zooming In, with Another Example

- A here stands for race, Y for loan default.



- Same idea, augmented with a *mediator*:



A Causal Primitive: Counterfactual Fairness

- If we have some protected attribute like race, and a decision such as length of sentence, then our decision satisfies *counterfactual fairness* if
 - “*had the protected attributes (e.g., race) of the individual been different, other things being equal, the decision would have remained the same*”
- *A causal model* is necessary to infer such claims from data.

Workflow

- Regardless of the machine learning algorithm to be used, work with a domain expert to **estimate a *causal model*** of your data.
 - It's a model of the world, not of your software.
- Choose **any machine learning algorithm of interest**, any black-box that takes as inputs observed and unobserved variables in your domain.
 - **Select a set of variables based on which sets respect counterfactual fairness.**
 - If necessary, infer unobserved variables from the observed ones.

Formalizing the Idea

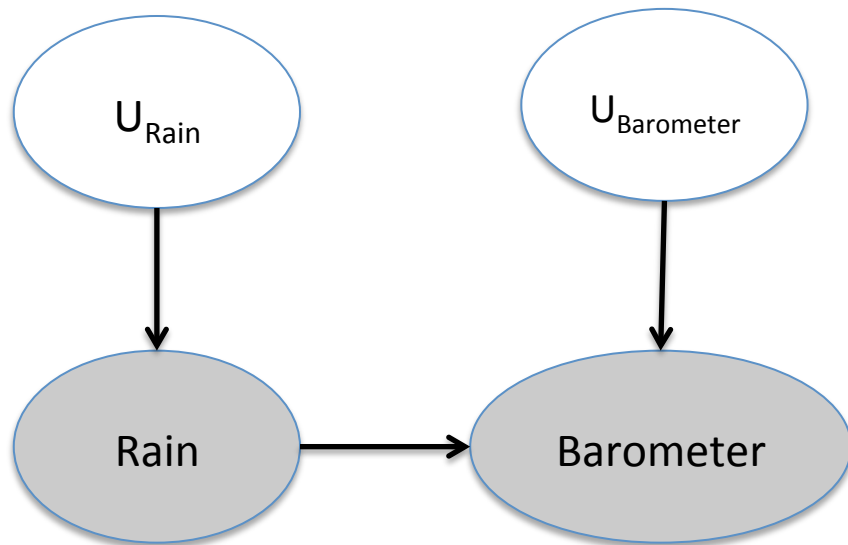
- Formal notions of counterfactuals date back at least to Jerzy Neyman in the 1920s.
- I will follow mostly the *Structural Causal Model* (SCM) framework of Judea Pearl, which has close links to the work of James Robins, and that of Spirtes, Glymour and Scheines.

Structural Causal Models

- A directed acyclic graph (DAG) postulates “direct cause-effect” pairs.
 - Each vertex in the graph is a random variable in a distribution.
- Each variable V is given an equation that deterministically defines the value of V as a function of its “parents”.
 - Such equations are postulated to be *structural*, in the sense that it follows the cause-effect direction.

The Operational Meaning

- This “DAG with equations” is causal in the sense that it must encode the *effects of a perfect intervention*.

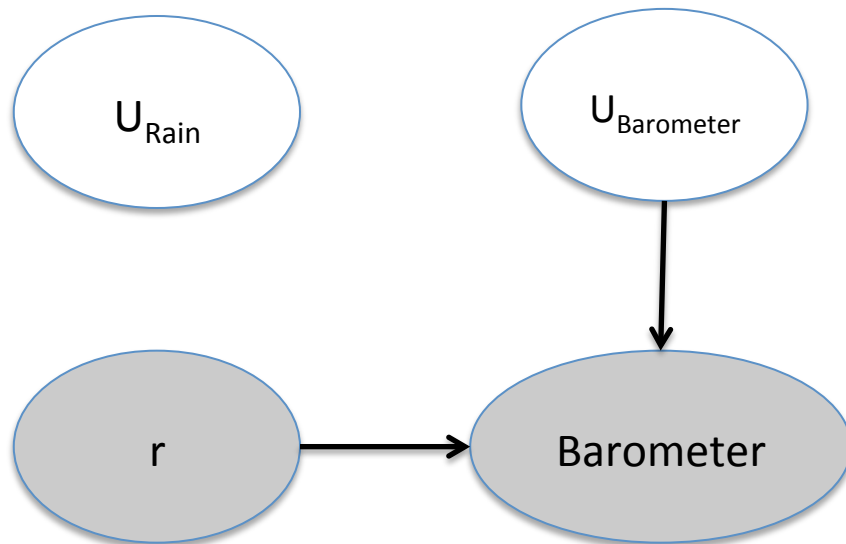


$$Rain = f_R(U_{Rain})$$

$$Barometer = f_B(Rain, U_{Barometer})$$

Interventions

- Another primitive. It is a “overriding” operator, sets a variable to a fixed value of interest. Lower case here represents constants.

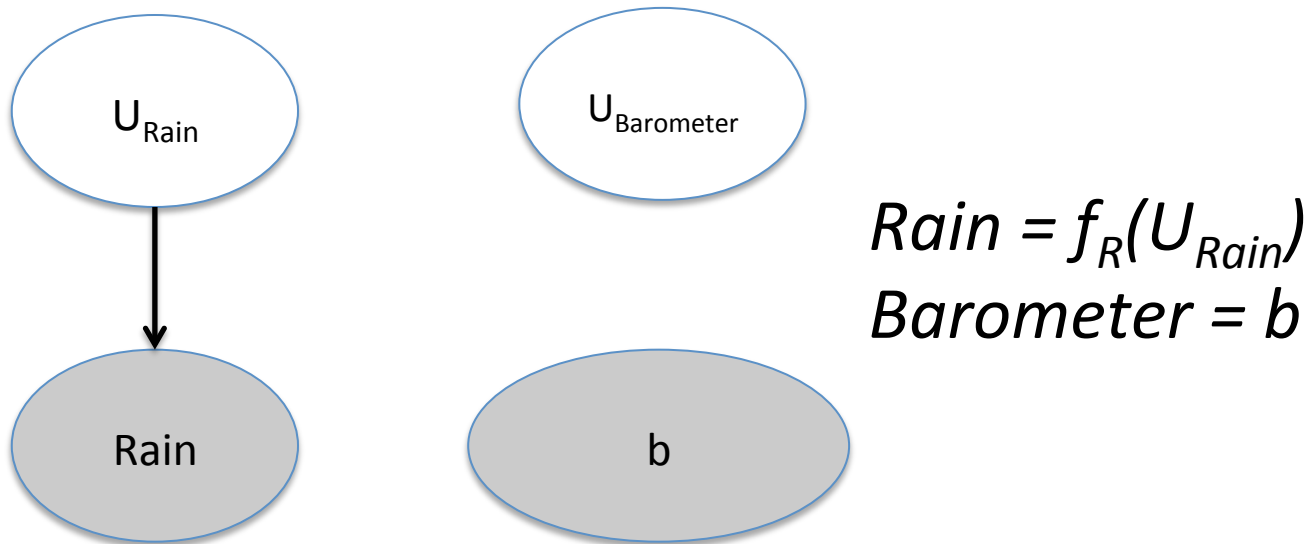


$$\text{Rain} = r$$

$$\text{Barometer} = f_B(r, U_{\text{Barometer}})$$

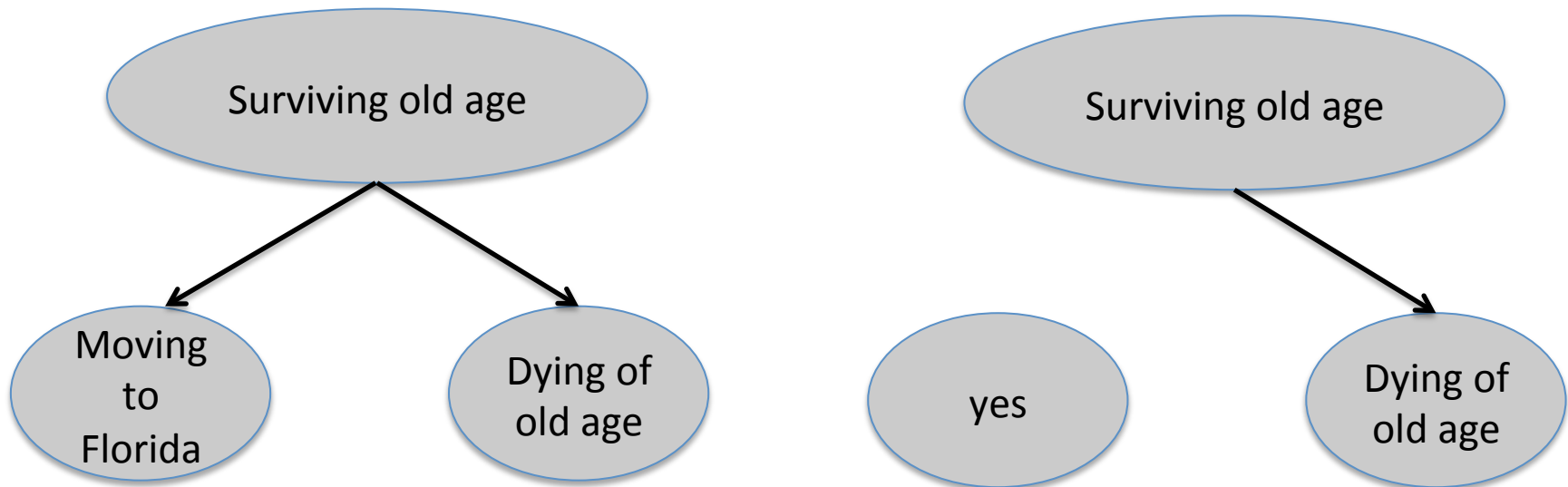
Interventions

- It is the notion of intervention that leads to the asymmetric nature of causality.



Interventions

- It is the notion of intervention that explains why “correlation is not causation”.



Notation: the “do” Operator

- We must express how “Dying of old age” varies with “Moving to Florida” in both cases.
- Traditionally, conditional probabilities can be used for that. But notice that, in our example,

$$P(\text{Dying of old age} = \text{True} \mid \text{Moving to Florida} = \text{True}) \neq P(\text{Dying of old age} = \text{True} \mid \text{Moving to Florida} = \text{False})$$

is true in the *observational* case (no intervention), but false in the *interventional* case.

Notation: the “do” Operator

- In Pearl’s calculus, this is distinguished by using the “do” operator to indicate an intervention as opposed to an observation.

$P(\text{Dying of old age} = \text{True} \mid \text{Moving to Florida} = \text{True}) \neq$
 $P(\text{Dying of old age} = \text{True} \mid \text{Moving to Florida} = \text{False})$

$P(\text{Dying of old age} = \text{True} \mid \text{do}(\text{Moving to Florida} = \text{True})) =$
 $P(\text{Dying of old age} = \text{True} \mid \text{do}(\text{Moving to Florida} = \text{False}))$

Averages Vs. Individuals

- This type of notation can be used to express whether a drug is effective or not, averaging over a population, using a *randomized controlled trial*:

$$P(\text{Healthy} = \text{True} \mid \text{do}(\text{Treatment} = \text{Drug})) \stackrel{?}{=} P(\text{Healthy} = \text{True} \mid \text{do}(\text{Treatment} = \text{Placebo}))$$

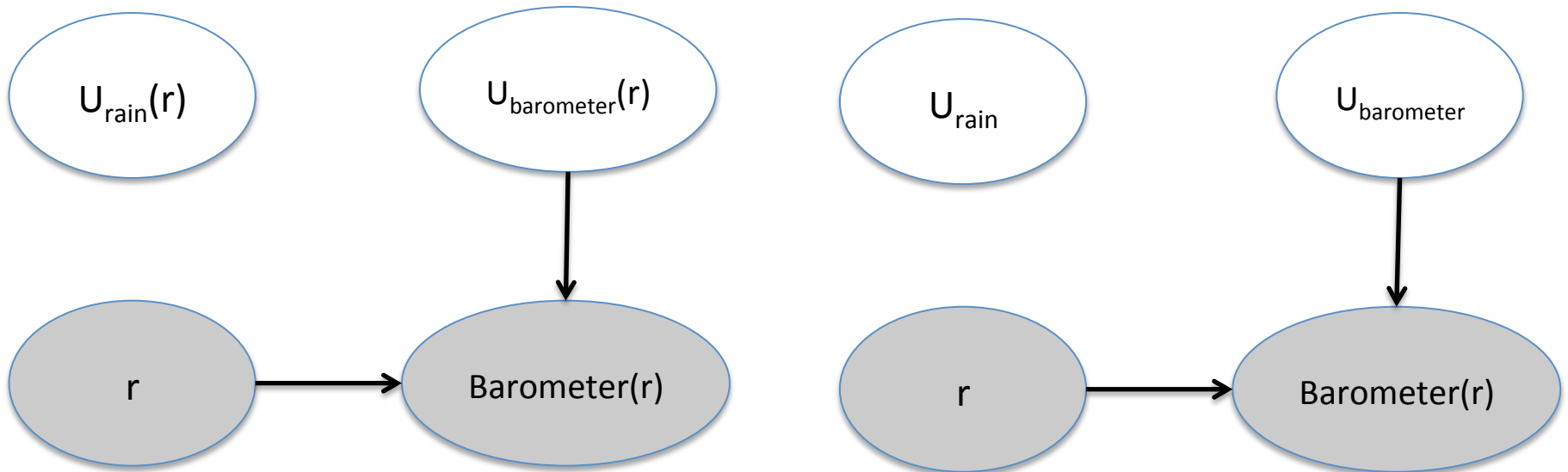
- It does not make any claims, however, on whether there is a balance of positive/negative cases that cancel out.

Notation: Counterfactual Indices

- Meant to capture individual-level variability.
- For V_j a variable in the system, and V_i a variable being intervened at value v , we use $V_j(v)$ as the counterfactual value of V_j , had V_i being set to v .
- Context will tell us which variable the value “ v ” refers to.

Example

- Notice: it is common to represent $V_j(v)$ as just V_j if V_i is not a (direct or indirect) cause of V_j .

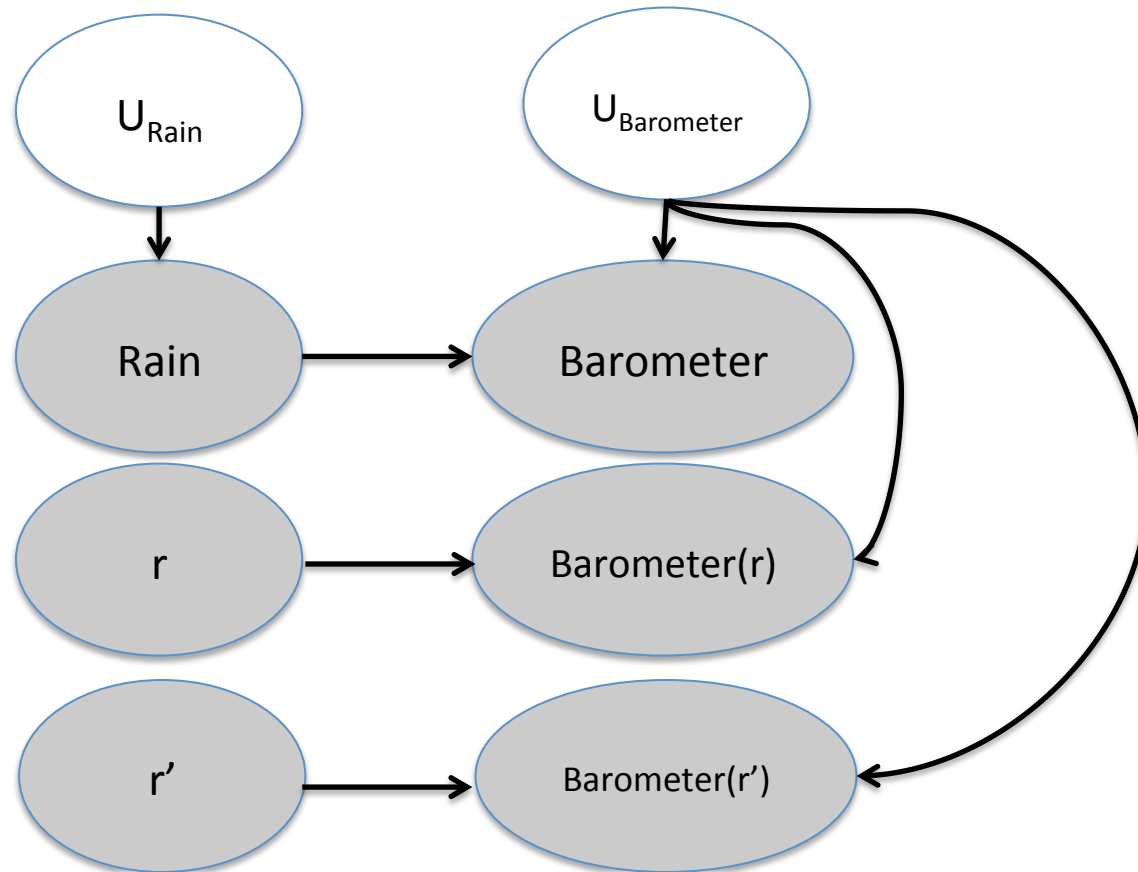


“Other Things Being Equal”

- That is,
 - A counterfactual value replaces the cause of interest
 - The counterfactual value propagates “downstream” the causal graph via the structural equations
 - Everything else remains the same (“other things being equal”), i.e., the non-descendants of the manipulated variable.


Multiple Worlds

- A counterfactual is just a different “version” of the same individual. All “versions” co-exist in one big joint distribution.



Workflow

- Regardless of the machine learning algorithm to be used, work with a domain expert to **estimate a *causal model*** of your data.
 - It's a model of the world, not of your software.
- Choose **any machine learning algorithm of interest**, any black-box that takes as inputs observed and unobserved variables in your domain.

- 
- **Select a set of variables based on which sets respect counterfactual fairness.**
 - If necessary, infer unobserved variables from the observed ones.

Back to Counterfactual Fairness

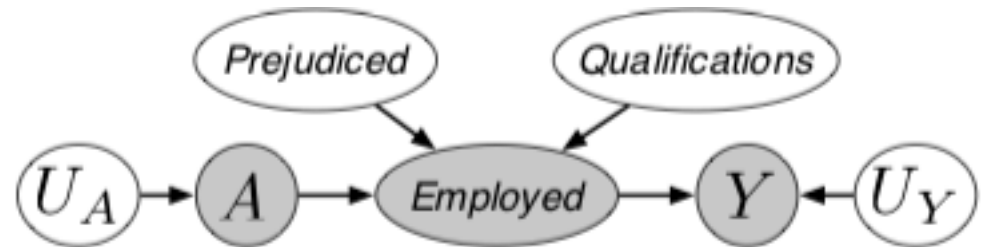
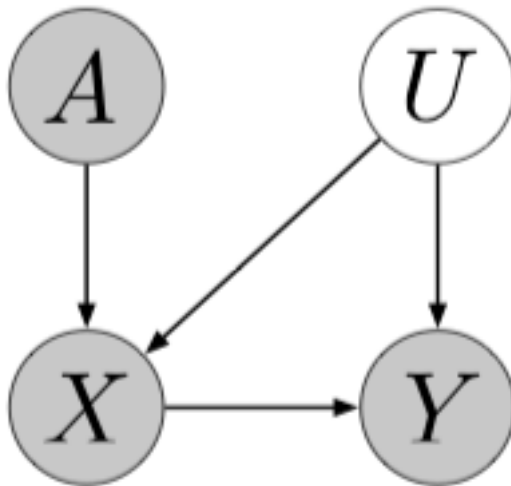
- The law of counterfactual propagation means that, if we want to ensure

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

it is sufficient (and necessary, in general) to include only the non-descendants of A in the definition of the predictor.

Examples

- *A*, *X* cannot be used. *U* can.
- *A*, *Employed* cannot be used. *Prejudiced* and *Qualifications* can.



- If it is judged that *Prejudiced* cannot be used, it should be labelled as a protected attributed.

How to Extract Unobserved Variables?

- Use the “factual” distribution to get a distribution over the unobserved variables by standard probabilistic conditioning.

$$P(\text{Unobserved} \mid \text{Observed})$$

- Monte Carlo data augmentation approach: replace each data point in your training sample by a set of training points with the unobserved variables being filled by a Monte Carlo sample.

Workflow

- Regardless of the machine learning algorithm to be used, work with a domain expert to **estimate a *causal model*** of your data.
 - It's a model of the world, not of your software.
- Choose **any machine learning algorithm of interest**, any black-box that takes as inputs observed and unobserved variables in your domain.
 - **Select a set of variables based on which sets respect counterfactual fairness.**
 - If necessary, infer unobserved variables from the observed ones.

Algorithm

- 1: **procedure** FAIRLEARNING(\mathcal{D}, \mathcal{M}) ▷ Learned parameters $\hat{\theta}$
- 2: For each data point $i \in \mathcal{D}$, sample m MCMC samples $U_1^{(i)}, \dots, U_m^{(i)} \sim P_{\mathcal{M}}(U \mid x^{(i)}, a^{(i)})$.
- 3: Let \mathcal{D}' be the augmented dataset where each point $(a^{(i)}, x^{(i)}, y^{(i)})$ in \mathcal{D} is replaced with the corresponding m points $\{(a^{(i)}, x^{(i)}, y^{(i)}, u_j^{(i)})\}$.
- 4: $\hat{\theta} \leftarrow \operatorname{argmin}_{\theta} \sum_{i' \in \mathcal{D}'} l(y^{(i')}, g_{\theta}(U^{(i')}, x_{\neq A}^{(i')}))$.
- 5: **end procedure**


Interpretation

- Extract causes of Y which are not mediators between A and Y .
- Find the “best approximation” to Y within the space of functions that exclude such mediators.
- Even if Y is “unfair” (A is a cause of it), by construction the predictor will be counterfactually fair.

Challenges

- Counterfactual fairness clarifies that algorithmic fairness in general is *not* explicitly modeling how the world becomes fairer with fair predictions.
 - Even if our decision of giving a loan is fair, it doesn't mean that in aggregate the probability of a person of a particular demographic group won't have difficulties in repaying it (A still causes Y).
- The delayed impact of fair predictions is also a research topic
 - see Liu et al., <https://arxiv.org/pdf/1803.04383.pdf>

Workflow

- 
- Regardless of the machine learning algorithm to be used, work with a domain expert to **estimate a *causal model*** of your data.
 - It's a model of the world, not of your software.
 - Choose **any machine learning algorithm of interest**, any black-box that takes as inputs observed and unobserved variables in your domain.
 - **Select a set of variables based on which sets respect counterfactual fairness.**
 - If necessary, infer unobserved variables from the observed ones.

Some Words of Caution

- Structural equations use unobserved variables.
- It is common that some of these variables are “default” choices based on some generic modeling assumption such as additive errors.

$$\text{Output} = \text{Signal} + \text{Noise}$$

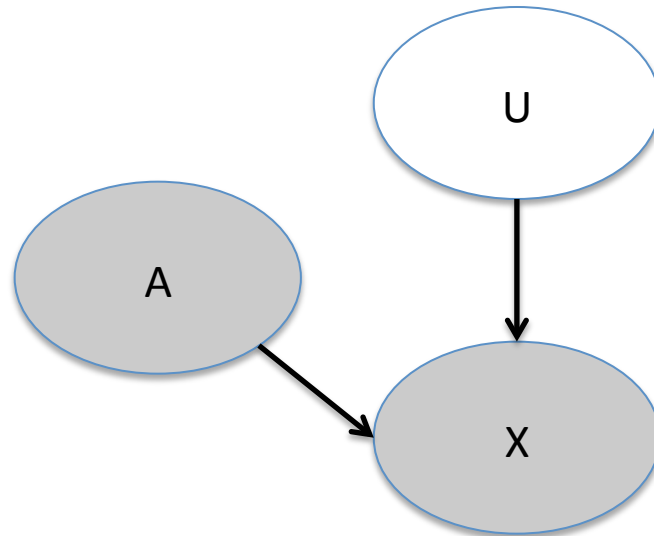
- Nature and society couldn't care less whether your mathematically convenient way of separating signal and “noise” is elegant or not.

Some Words of Caution

- That is, *there are infinitely many* structural equations $V_j = f_j(V_i, U_j)$ compatible with $P(V_i | V_j)$ and $P(V_i | \text{do}(V_j))$.
- Signal vs. noise must be determined by real-world assumptions (“simplicity” assumptions, of the Ockham’s razor type, can be sometimes adequate as long as caveats are advertised).

Some Words of Caution

- By now, there are several good papers on how to tackle fairness by generating unobserved variables which are independent of A , using assorted methods.



“Then off you go to plug-in U on a machine learning algorithm!”

However

- There are papers not causally motivated, which I find of difficult interpretation.
 - Remember: there are infinitely many ways of extracting U .
- There are papers causally motivated, but which commit themselves to a domain-free family of structural equations. OK enough, but why would you do that?
 - *Counterfactual fairness emphasizes that the causal modeling step is separate from the prediction learning process.*
- Finally, *do beware of any paper that claims to do assumption-free extraction of “causal latent factors”. Those are selling you snake oil!*

Interpretation of Counterfactuals

- But what does it mean to say “had my race been different”??
 - First, make sure to understand the difference between “A” and “Perception of A”: these can lead to conceptually different interpretations, even if the model stays the same.
 - Without going in details, if those counterfactuals make you feel uneasy, just interpret them as comparing two different people who happen to match on the “other things being equal” factors.
 - This is also related to *fairness through awareness* (Dwork et al., 2011, <https://arxiv.org/abs/1104.3913>)

Non-Counterfactual Causal Models

- Contrary to folk knowledge, causality does not require counterfactuals: the “do” operator is an way of comparing treatments without comparing individuals.
- However, if features X are affected by A , then in general there is no individual where
$$P(Y = y \mid X = x, \text{do}(A = a)) = P(Y = y \mid X = x, \text{do}(A = a'))$$
- If features X are not affected by A , then we can show we do not need to explicitly model structural equations anyway!

The Upside

- Because structural causal models rely on unobserved variables, at least they can be partially falsified by eventually measuring some of those variables.
- Just keep in mind:
 - while it is preposterous to say you have “the” causal model of a social process, you should (must?) be able to explain your assumptions to a regulator or a customer.
 - Having passed testable implications, the remaining components of a counterfactual model should be understood as conjectures formulated according to the best of our knowledge. *Such models should always be deemed provisional and prone to modifications.*

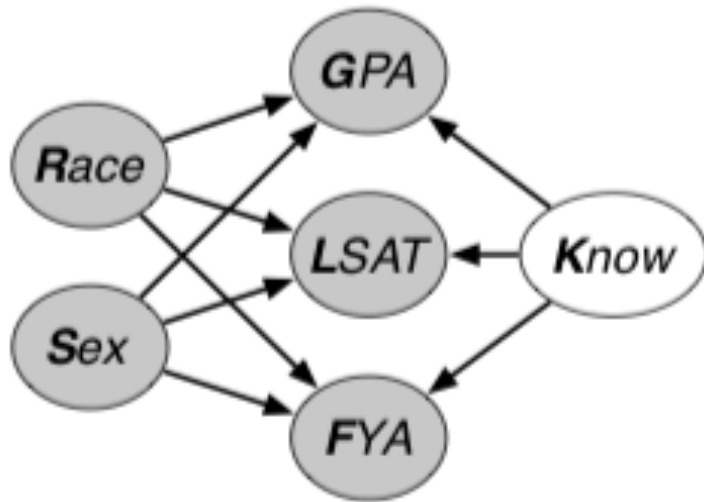
Illustration

- The Law School Admission Council conducted a survey across 163 law schools in the United States
 - It contains information on 21,790 law students such as their entrance exam scores (LSAT), their grade-point average (GPA) collected prior to law school, and their first year average grade (FYA).
- Task: predict if an applicant will have a high FYA
 - Example of decision problem: make an offer

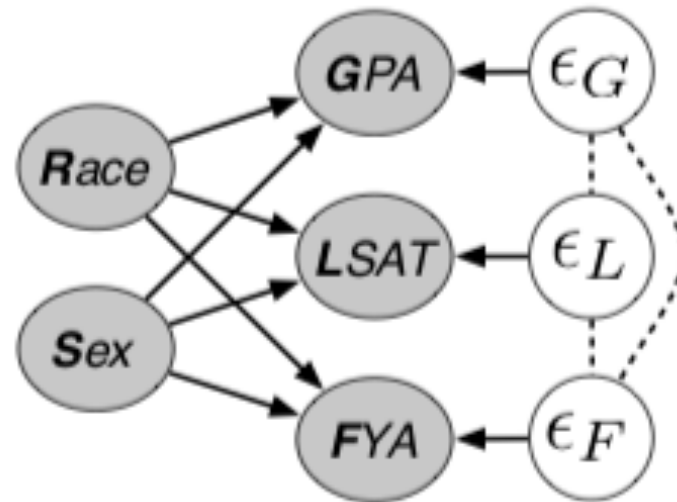
Setup

- I will present some simple causal models for this domain, which by no means I intend to sell as well-thought models. Their purpose is for illustration.
- We will fit real data to a model, then generate synthetic counterfactuals out of it. The point is to quantify to what extent a causally-oblivious method violates counterfactual fairness.

Two Models



"Fair *K*"



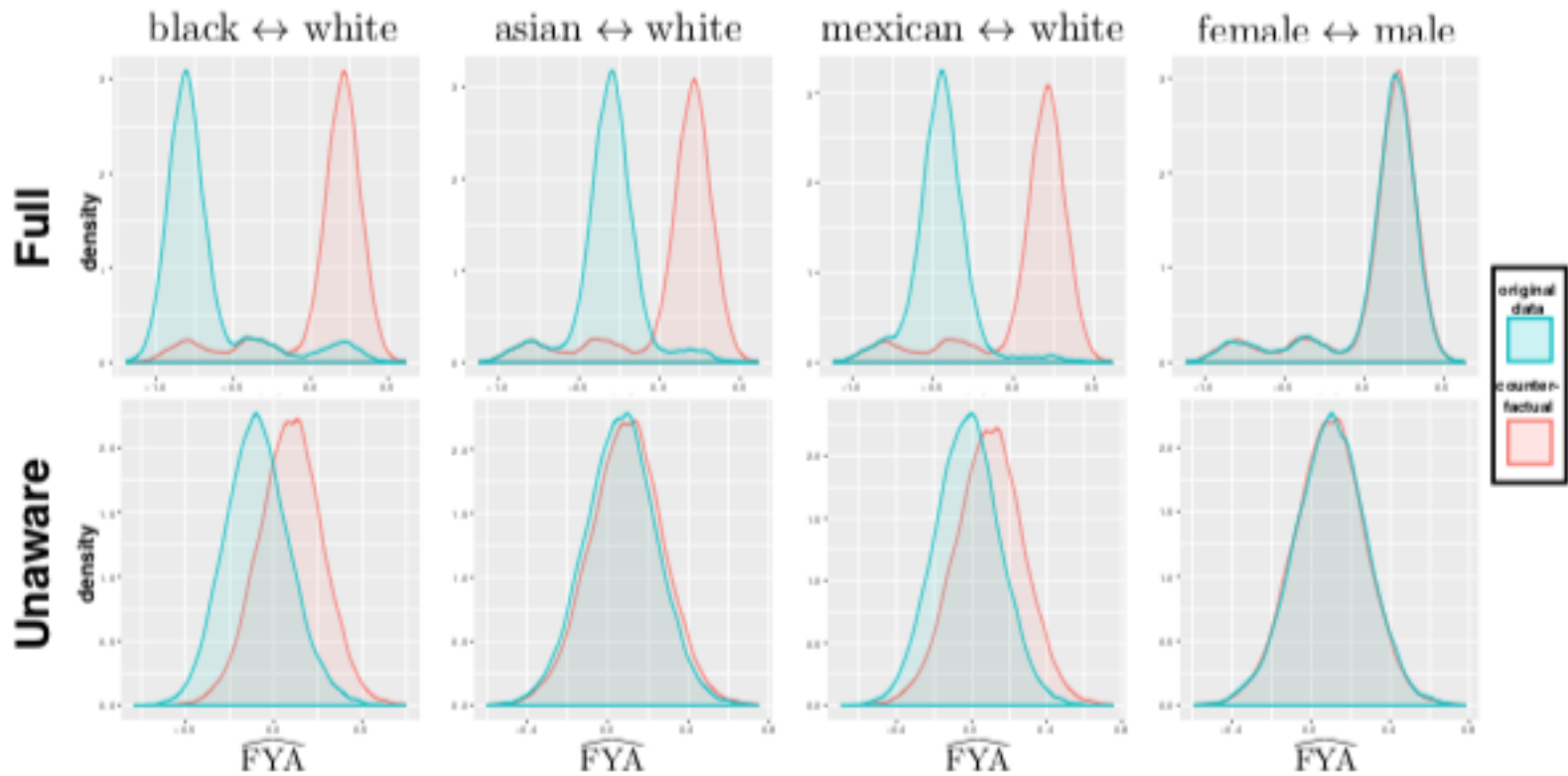
"Fair *Add*"

Predictive Error (Real Data)

- Comparison against “Full” (linear model with all variables) and “Unaware” (linear model without race and gender, but the other two predictors)
 - Evaluation by root mean squared error

	Full	Unaware	Fair K	Fair Add
RMSE	0.873	0.894	0.929	0.918

Fairness Violations (Simulated Counterfactuals)



Extension: Using Multiple Models

- We just saw two different counterfactual models that give different predictions despite being undistinguishable given the same data.
- This is OK assuming little difference between models, but we may have competing theories with some sizeable difference. We would like to be “approximately counterfactually fair” to all of them.

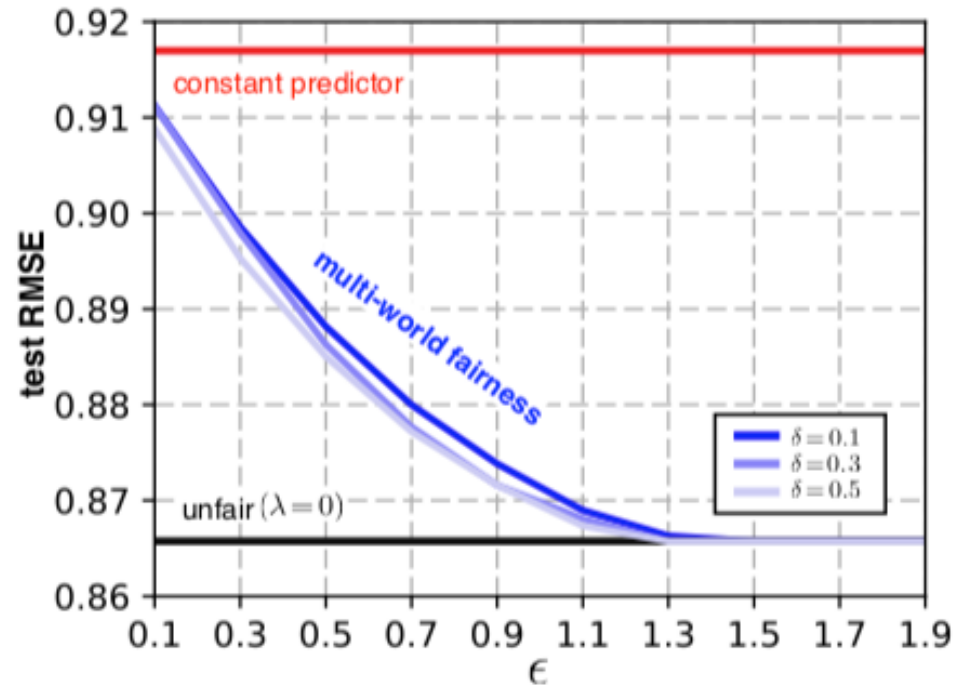
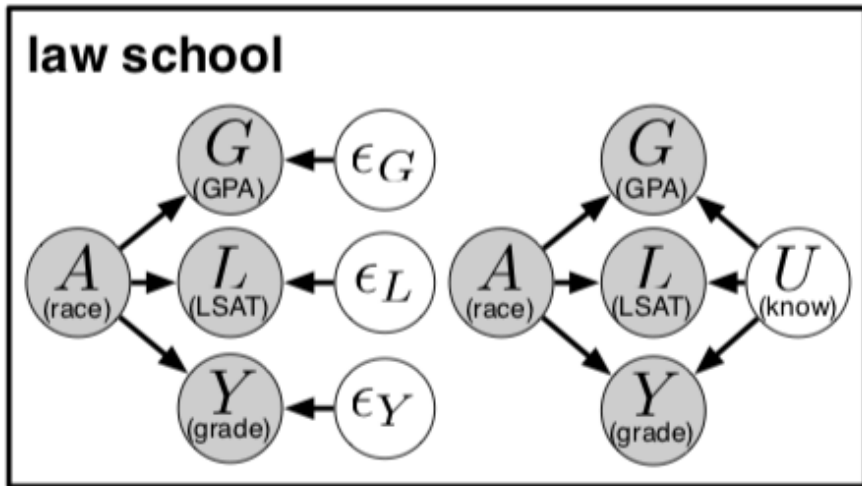
(ϵ, δ) - Counterfactual Fairness

- The following constraint provides a relaxation of counterfactual fairness:

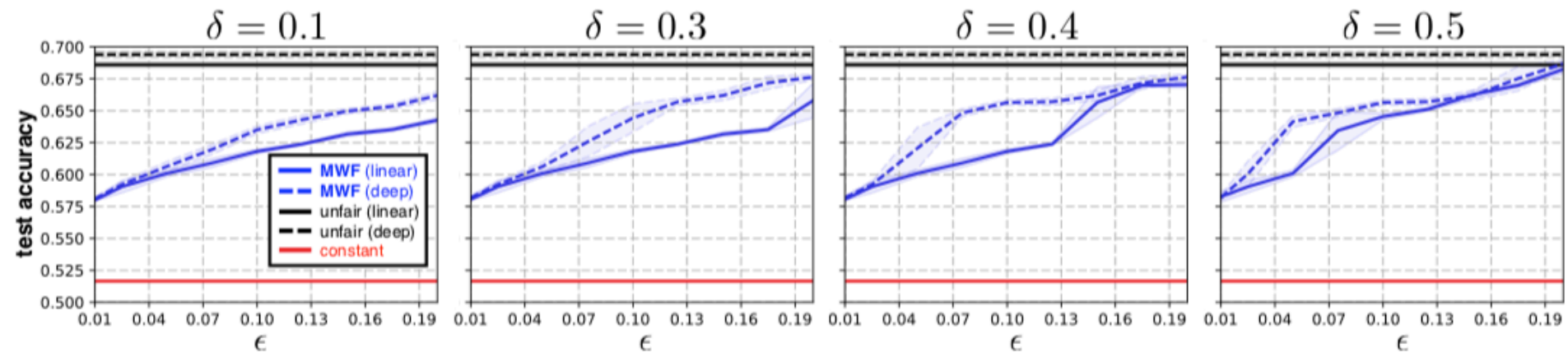
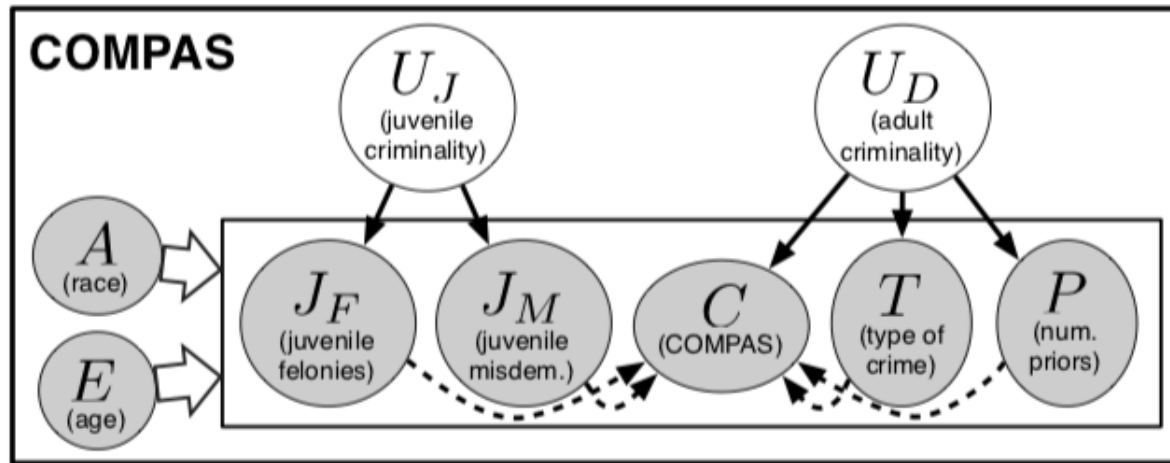
$$P(|\hat{Y}(a) - \hat{Y}(a')| \leq \epsilon \mid A = a, X = a) \geq 1 - \delta$$

- The idea is to simultaneously satisfy such constraints according to different counterfactual models.
 - It is not hard to show that this problem in general has no solution if $\epsilon = 0$, hence the need for an approximate version.

Law School Revisited



COMPAS



A Different Direction: Interventions

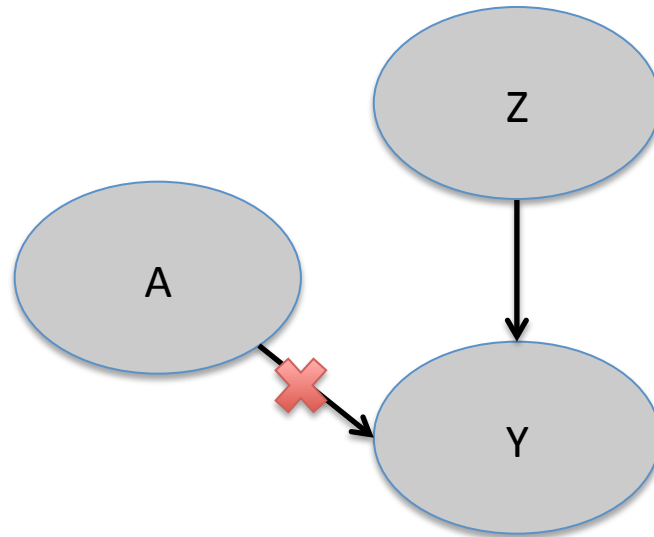
- So far, we have solely discussed the creation of predictors.
- Ideally, we would like to destroy the pathways between A and Y , the outcome of interest.
- This is in general not possible. But let's assume we have an intervention with the ability of changing the contribution of A to Y . How is related to counterfactual fairness?

Imperfect Interventions and Interference

- We will assume two generalizations of the concept of intervention used so far.
- An intervention is represented generically as a set of (action) variables, which here I will denote as Z .
 - We can define $Z = 0$ as the “no action” choice!
 - $Z \neq 0$ just means that one or more structural equations will change, not necessarily to a constant (“imperfect”, or “soft” intervention).

Ideally

- Having available some “ $Z = z$ ” which completely overrides the structural equation for Y to not depend on anything that starts on A .



In Reality

- No such an intervention is typically available.
- And this is not a prediction problem anymore.
What happens to Y ?
- Setup:
 - Assume Y is encoded so that high values are good.
 - Model allows for *interference*: that is, treatment Z_i given to person i might affect person j .
 - How is this related to counterfactual fairness?

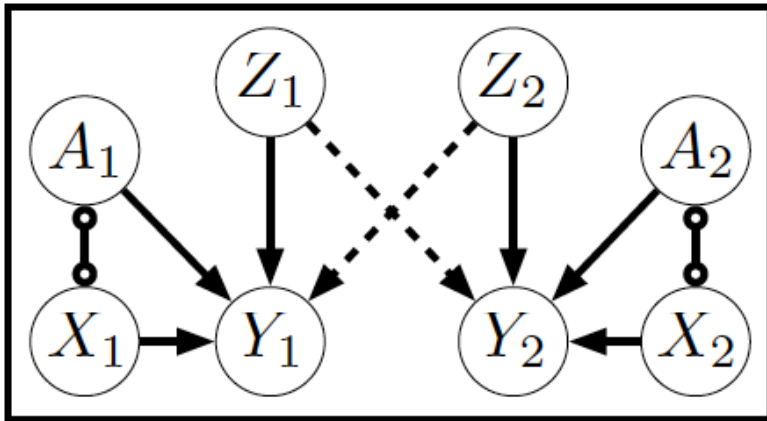
Optimization Problem and Constraints

Main family of constraints:

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

(as opposed to

$$|\mathbb{E}[Y_i(a_i, \mathbf{z}) \mid A_i = a_i, X_i = x_i] - \mathbb{E}[Y_i(a', \mathbf{z}) \mid A_i = a_i, X_i = x_i]| < \tau)$$



$$\max_{z_1, \dots, z_n} \sum_{i=1}^n \mathbb{E}[Y_i(\mathbf{z}) \mid A_i = a_i, X_i = x_i]$$

$$s.t., \sum_{i=1}^n z_i \leq B$$

$$G_{ia'} \leq \tau \quad \forall a' \in \mathcal{A}, i \in 1, \dots, n,$$

Intuitive Toy Example

- Protected attribute A is such that $A \in \{b, w\}$, X is some quantitative measure of professional competence, and Y is a measure of wealth in 5 years' time.
- $Z_i = 1$ means individual i gets a subsidy to move to a neighborhood with better transport links.

Intuitive Toy Example

- Suppose structural equation is

$$Y_i = X_i + 100Z_i + 50Z_i \times \mathbb{I}(A_i = w) + U_i$$

- So if there are two individuals, one of type w and one of type b , and $Z_1 + Z_2 = 1$.
- Without the fairness constraint, type w gets the subsidy even if type b has up to 50 more units of professional ability!

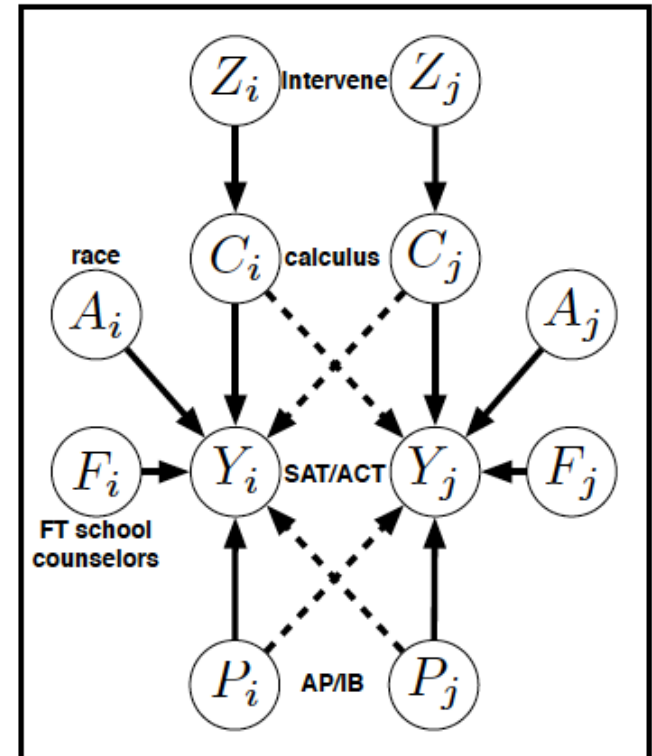
Considerations

- There might be no feasible solution if τ is small enough.
- It might be the case that the “counterfactual gap” in each constraint remains constant regardless of Z .
- These are features of the intervention, not of the fairness framework. Again, a good intervention is a matter of real world design, not of algorithm design!

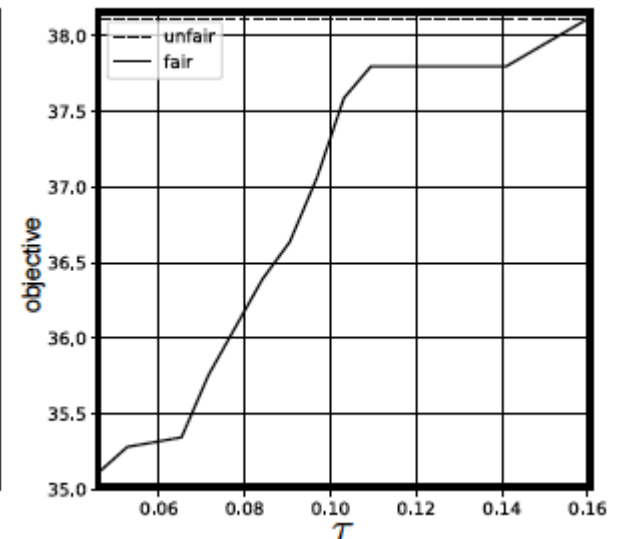
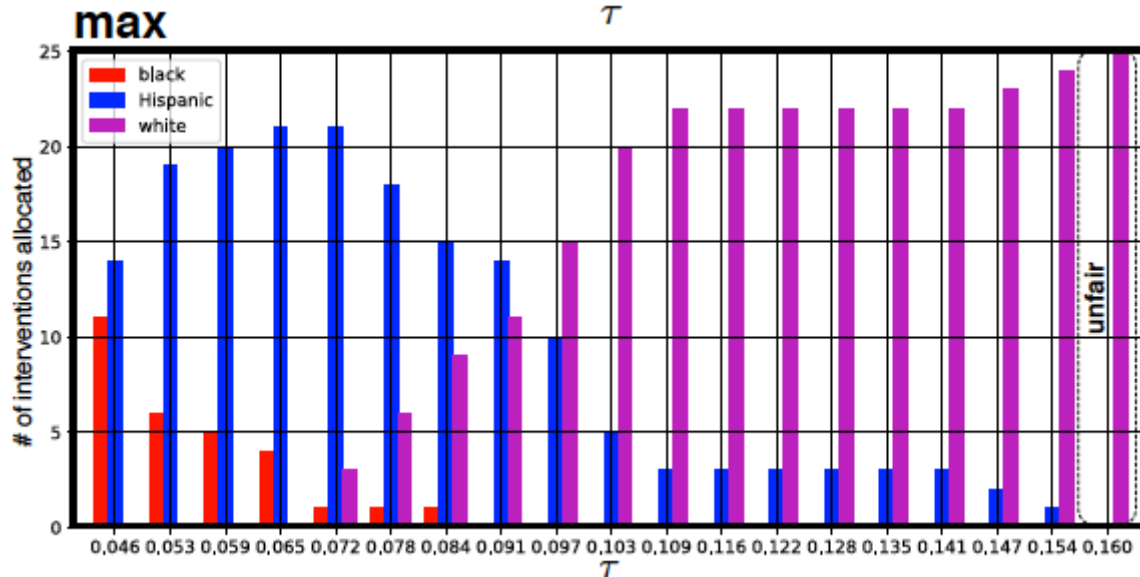
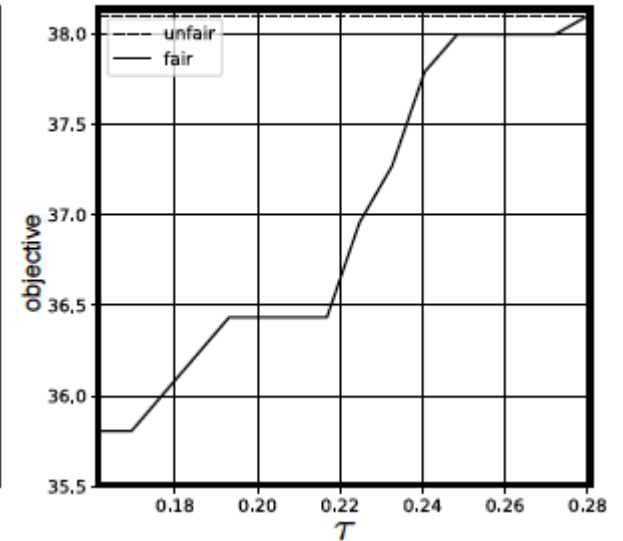
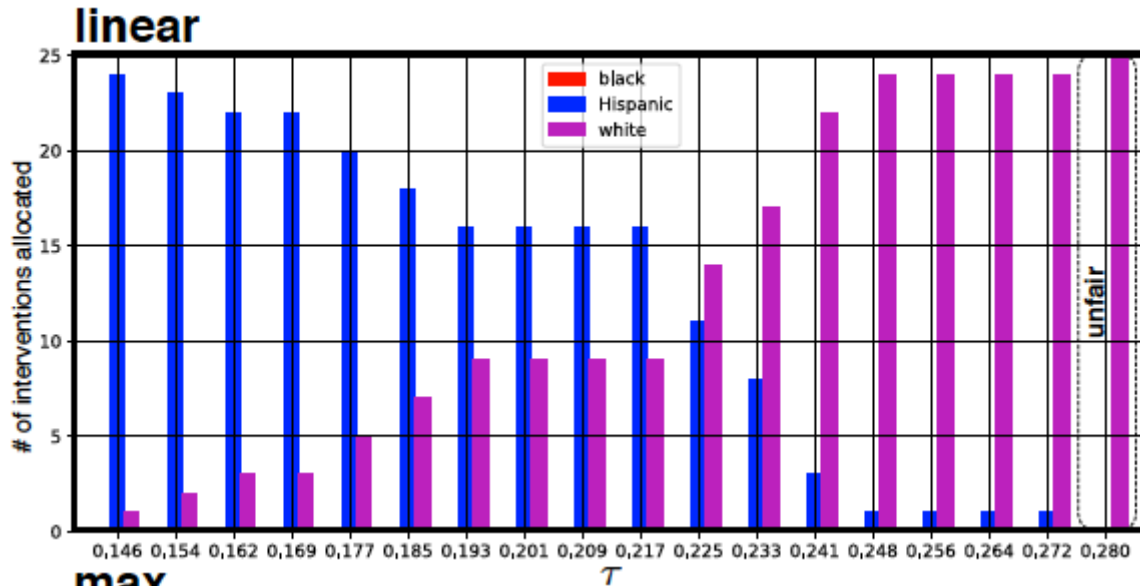
Illustration

(Partially Synthetic Data)

- NYC Public Schools: intervention Z is to provide calculus classes in schools.
- Attribute A is whether school has a white majority.
- Outcome Y is proportion of students taking the SAT/ACT.
- Geographical interference is assumed.



Results



Conclusion

- I propose that causal modeling should be a key component of fairness considerations.
- Fairness has multiple facets. Here we considered prediction and policy-making under interference.
- Much more is relevant: selection bias, dynamic prediction/treatments etc.
- Good software design could help building massive experiments in the internet, for instance.

References

- M. Kusner, C. Russell, J. Loftus and R. Silva (2017). “Counterfactual Fairness”. NIPS 2017.
<https://papers.nips.cc/paper/6995-counterfactual-fairness>
- C. Russell, M. Kusner, J. Loftus and R. Silva (2017). “When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness”. NIPS 2017.
<https://papers.nips.cc/paper/7220-when-worlds-collide-integrating-different-counterfactual-assumptions-in-fairness>
- J. Loftus, C. Russell, M. Kusner and R. Silva (2018). “Causal Reasoning for Algorithmic Fairness”.
<https://arxiv.org/abs/1805.05859>
- M. Kusner, C. Russell, J. Loftus and R. Silva (2018). “Causal Interventions for Fairness”. To be in arXiv at any moment.

Thank You