ISYE 8813: Algorithmic Foundations of Ethical ML **Fall 2021**

Lectures 20: Fairness and Causality

Lecturer: Juba Ziani

All of the fairness criteria we have studied so far are observational, and how the features X, score R, sensitive attribute A, and target variable Y statistically relate to each other. Most of these definitions have a few things going for them, in particular: i) they are easy to formalize and can be stated simply in terms of X, Y, R, A, ii) in principle if you know the joint distribution of these attributes (or have enough samples), you can check whether they hold. However, these definitions have some limitations: observational definitions are often unable to distinguish between two worlds that are structurally different (i.e., the causal structure between the different parameters of the problem is different across both worlds), but still map to the same joint distribution.

1 Some limitations of observational definitions through examples

These examples are taken from the book [BHN17], and from Boaz Barak's lecture notes at `https://windowsontheory.org/2021/06/11/causality-and-fairness/`.

1.1 Berkeley admission data

UC Berkeley admissions data from 1973.

	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	**82**
B	520	60	25	**68**
C	325	**37**	593	34
D	417	33	375	**35**
E	191	**28**	393	24
F	373	6	341	**7**

The data above comes from study [BHO75] by Bickel et al., and looks at Berkeley admissions in 1973 across a few departments. In 1973, among the applicants to Berkeley university,

among 35 percent of women were admitted, versus 44 percent of men. This seems to suggest that there was some unfairness across gender in admissions in the form of different acceptance rates for males and females, in favor of males. Looking at the 6 largest departments at Berkeley, those insights still hold (44 percent admission rate for men versus only 30 for women).

At the level of individual departments, however, the picture looks different. In 4 of the 6 biggest departments in Berkeley, the data shows that women were more likely to be admitted, and the remaining 2 departments, the numbers are fairly close! So, what happened here? The distributions of department that men and women applied to is very different: for example, very few women applied to department A which has a high admission rate, but many more women than men applied to departments C and E which have much lower admission rates!

Observing this, this is what the original study concluded: *"The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects." I.e., there was no wrongdoing by the departments, it's rather a pipeline problem in which women are socialized to choose different careers than men.*

But this is not clear-cut inference, there could be many other explanations. Why did women apply to more competitive departments in the first place?

- Maybe because the less competitive departments were unwelcoming of women at the time.
- Maybe some departments were known for poorly treating women, and they decided not to apply.
- Maybe the department did not advertise to women in the first place and discouraged them from applying.
- etc.

There is no way, simply from the static, observational data we have, to figure out which explanation is right. This causes difficulties when trying to come up with interventions for fairness, as different explanations have different interventions! The first one requires changes earlier in the education pipeline; the next two requires improving the treatment of women with some of these departments; the last one requires targeting and reaching out to women better.

I.e., one of the major shortcoming of statistical and observational data is that *completely different states of the world that requires different interventions for fairness may be statistically and observation-ally indistinguishable from each other, given only access to static data.* How to deal with these issues? We can i) make assumptions about what the state of the world is, or ii) you can try to dynamically implement interventions to learn if and how they

affect different groups and populations, and in turn which explanation of the world is more likely. This ties back nicely to one of the first point we made in the fairness part of the class: guaranteeing fairness depends on our view of the world and the assumptions we make about the state of the world.

1.2 A more formal example of indistinguishability: Exercise, weight, and heart disease

Imagine I want to understand the relationship between exercise, weight, and heart disease. To do so, let us write X to variable that corresponds to exercising regularly, W to weight, and H to heart disease. I could imagine there are several causal relationships between these three attributes. Let us focus on a few possible causal relationships for now:

- If I exercise, I am less likely to be overweight and to have heart disease. But the fact that I am overweight itself does not cause heart disease, both are symptoms of my lack of exercise. This is the left example of Figure 1.

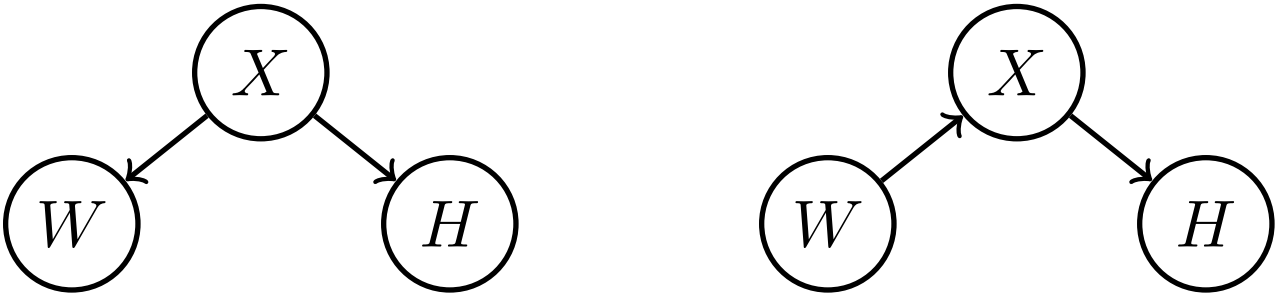


Figure 1: Different possible causal relationships between exercise, weight, and heart disease.

- The fact that I am overweight prevents me from exercising. Then, because I do not exercise, I am more likely to develop a heart disease. This is the right example of Figure 1.
- Because I do not exercise, I am overweight. Then, being overweight may lead to heart disease.
- etc. (those are non-exhaustive examples)

In these figures, think of an arrow $Z_1 \to Z_2$ denoting that Z_1 has a *direct* effect on Z_2. What that means is that if $\text{pa}(Z)$ is the set of parents of variable Z, Z can be entirely characterized as a randomized function of $\text{pa}(Z)$, without referring to other variables. Other variables may still have an effect on and be correlated with Z, but *only indirectly, through its parents.*

Now, imagine that in scenario 1 (left), the data is generating according to the following process:

- $X \sim B(1/2)$ (binomial with probability 1/2).

- If $X = 1$, then $W = 0$; else, $W \sim B(1/2)$.
- If $X = 1$, then $H = 0$; else, $H \sim B(1/2)$.

Note the process is consistent with our causal graph. First, we draw X, then the value of X affects both W and H.

In scenario 2 (right), imagine the process is the following:

- $W \sim B(1/4)$
- If $W = 1$, then $X = 0$; else, $X \sim B(2/3)$.
- If $X = 1$, then $H = 0$; else, $H \sim B(1/2)$.

We first draw W, which then affects X, which then affects H.

Clearly, the two causal models that we are looking at here are different. What happens, however, to the joint distribution of (W, X, H)? To do so, we compute this joint distribution in the table below:

X	W	H	Scenario 1	Scenario 2
0	0	0	1/8	1/8
1	0	0	1/2	1/2
0	1	0	1/8	1/8
0	0	1	1/8	1/8
1	1	0	1/4	1/4
1	0	1	0	0
0	1	1	1/8	1/8
1	1	1	0	0

We find that in both cases, the joint distribution of variables are identical! I.e., when looking at statistical and observational definitions of fairness, the two worlds/scenarios above are completely indistinguishable! I.e., when looking at fairness from the statistical point of view, we are losing information and generality. This is a typical example of "correlation does not imply causation": the joint distribution of X, W, H tells you about how correlated they are, but is not powerful enough to recover the full causation structure here.

2 Causal graphs and effects

The approach we will take here to address and discuss some of these issues is that of *causality*. Instead of trying to understand properties of individuals or groups and trying to guarantee fairness through only static, observational or distributional information, we will instead work with *causal graphs* that describe of different parameters of the problem affect each other.

2.1 Defining causal graphs

We have seen simple examples of such graphs in Section 1 of this lecture, but now let us define them formally.

Definition 1 (Structural causal model)**.** *A structural causal model M is given by a set of variables $X_1, \ldots, X_d$, and corresponding assignments of the form*

$$X_i := f_i(P_i, U_i),\ i = 1, \ldots, d.$$

*Here, $P_i \subset \{X_1, \ldots, X_d\}$ is a subset of variables that we call the **parents** of X_i. The random variables $U_1, \ldots, U_d$ are called noise variables and are mutually independent, as well as independent from the X_i's.*

Note that this kind of model aims to represent and formalize assumptions about the causal structure between the different variables $X_1, \ldots, X_d$ that we are interested in, and how they affect each other:

- For each i, since X_i can be written entirely determined as a function of its parents in P_i, these parents are the features that have a *direct* causal effect on X_i: changing some of the parents changes X_i directly.
- Changing other features in the model may lead to a change in X_i, but only *indirectly*, through, changing its parents P_i.

The random variable $U_1, \ldots, U_d$ are here to represent that X_i can be a *randomized* function of its parents, as in the example of Section 1.2. They can be seen as external/exogenous factors that influence the system.

A structural causal model can be represented by a causal graph:

Definition 2 (Causal graph)**.** *The graph corresponding to the causal structural model $\{(X_i, f_i)\}$ is a **directed** graph that has a node for each variable X_i, and each node X_i has an incoming edge from all of its parents P_i. We call this graph the causal graph corresponding to the structural causal model.*

For example, we have seen such simple causal graphs in Section 1.2., Figure 1.

2.2 Some common graph structures

Here, we will look at a few common graph structures. These graph structures are going to be particularly important in terms of understand the causal relationships between different variables, how they can affect each other directly or indirectly, and how confounding effects can arise. The reason we want to do so, in the space of fairness, is that we want to understand how bias or unfairness in one variable is going to affect other variables in our problem, and how fair our decisions are going to be.

Forks A fork is a node Z that has outgoing edges to two other variables X and Y: i.e., Z is a common cause of X and Y, as seen in Figure 2. The reason forks are an important special structure is because they have a *confounding* effect (in which there can be correlation without causation) on X and Y. Note here that X does not have a causal effect on Y, whether it be direct of indirect. However, X and Y are correlated with each other, simply because they are both consequences of/informative about Z! Conditioning on different values of X without conditioning on Z will lead to different distributions for Y.

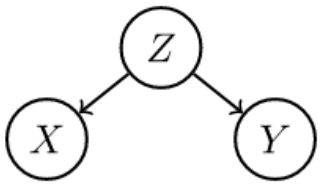


Figure 2: A fork

This will be important later on, for the following reason: confounding leads to disagreements between i) conditioning on a random variable and ii) intervening on that random variable and changing it. Here is the reason, in the example above:

- Conditioning on X *affects* the joint distribution of Z, Y. This is because conditioning on X gives me information about what Z is (since Z causes X), and in turn on what Y is.
- However, intervening to change the value of X does not change the value of Z, as Z is not a function of X! In turn, the value of Y is also unchanged.

This distinction will be important later on when trying to understand the effect of a given variable on another.

Mediators Mediators are simpler situations in which such confounding effects may not arise. A mediator is a node Z such that Z lies on a directed path from X to Y; see Figure 3 for an example of a situation in which Z is a mediator.

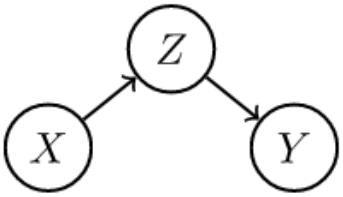


Figure 3: A mediator

In this case, we have the causal path $X \to Z \to Y$. There, as before, X and Y are correlated though Z. However, it is the case that X is a cause of Y, because X now causes Y through Z! So, here, Z does not have a counfounding effect across X and Y.

Colliders In a collider, the collider node Z is caused by both X and Y, as seen in Figure 4. Here, there is no confounding effects between X and Y, once again: X and Y are not correlated which each other, nor to they have a causal relationship. If I condition on X, I learn nothing more about Y than what I already knew through Z, and vice-versa.

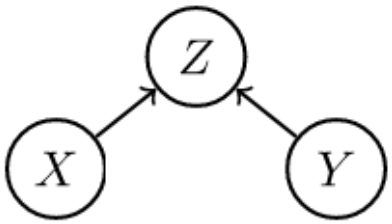


Figure 4: A collider

However, colliders have an interesting effect: if I condition on the *collider* Z itself, I am now introducing correlation between X and Y! Imagine $Z = 1$ is being admitted to the hospital, and $X = 1$ is having meningitis, and $Y = 1$ is having a broken bone. Conditional on $Z = 1$, I now learn that observing that one agent has meningitis $X = 1$ makes the likelihood that he has a bone fracture $Y = 1$ much smaller, as it is unlikely someone was admitted for both having meningitis and a broken bone at the same time.

3 Interventions and Causal Effects

The "do" operator Given a causal model M, we can take any assignment of the form $X := F(P, U)$, and replace it by another assignment. For example, we could assign X a constant value x, i.e. do $X := x$. Let us denote the resulting model $M' = M(X := x)$. This assignment operator is called to do-operator, and another notation for it is to write $\text{do}(X := x)$ for the operation of assigning x to X.

Graphically, this operation corresponds to eliminating all incoming edges to the node X, and fixing the value of that node, as seen in Figure 5. Note that this can be *different* from conditioning on $X = x$, when *counfounding effects are present.*

Causal Effect The causal effect of an action $X := x$ on a variable Y refers to all the ways in which setting X to any possible value x affects the distribution of Y. Here, let us look at this in a simple binary setting in which $X \in \{0, 1\}$. Then we are interested in a quantity of the form

$$\mathbb{E}_{M[X:=1]}[Y] - \mathbb{E}_{M[X:=0]}[Y],$$

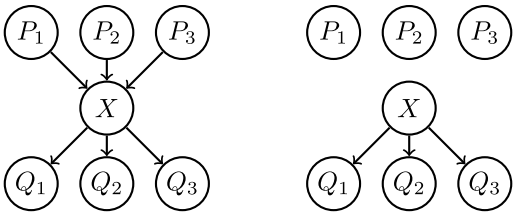


Figure 5: The do operator

that we call the *average treatment effect*. This is a measure of causality in the sense that it tells us how much changing X from 0 to 1 affects Y.

The issue here comes again from confounding. If we are trying to compute such effects without interventions, just by looking at conditional probabilities and conditioning on someone's attributes to see their effects, you might detect correlation that does not come from causation.

Avoiding confounding effects There are several possible ways to do so:

- One can run randomized trials with an control and a treatment group. We can then directly implement the do operator/an intervention by changing the value of the variable of interest in the treatment group, compared to the control group. But this has limitations: once again, you cannot change someone's race, so not all variables are ones you can even physically intervene. There may also be ethical or legal reasons why some interventions are in possible. So, sometimes, we are constrained to use conditional statements rather than actual interventions/do operators.

- Let's say we are trying to see the effect of X on Y, and see if one of the paths from X to Y introduces confounding effects. While this looks like this may be complicated, it actually has been show in the causality literature [Pea09] that there is a simple criterion for confounding effects to arise, the *backdoor criterion*. I am not going to get into the details of this criterion here, and suggest that you read [Pea09] more carefully if you are interested! Basically this characterizes, when we are conditioning on a (Set of) random variables, whenever these random variables create a fork or collider situation with respect to X and Y.

  This is more of away to detect confounding rather than correct for it.

- Another approach is to note that, conditional on its parents nodes $\text{pa}(X_i)$, X_i is independent of all other nodes. This is because we can write

$$X_i = f(\text{pa}(X_i), \{U_i\}),$$

  implying that the distribution of X_i conditional on $\text{pa}(X_i) = \vec{p}$ is given by

$$\Pr\left[X_i = x_i \mid \text{pa}(X_i) = \vec{p}\right] = \Pr\left[f(\vec{p}, \{U_i\}) = x_i\right];$$

conditional independence then follows from the fact that the U_i's are independent of the X_i's. So, we can write the following *adjustment formula*:

$$\Pr\left[Y = y | \text{do}(X := x)\right] = \sum_z \Pr\left[Y = y | X = x, \text{pa}(X) = \vec{p}\right] \Pr\left[\text{pa}(X) = \vec{p}\right].$$

This is exactly the idea of *controlling* for a variable; we estimate the effect of X on Y by controlling for/taking into account all of the possible values $\vec{p}$ of X's parents. This guarantees then that the effect we observe comes from X itself, and not from variables that are confounded with X through its parents.

One weakness of this approach is that it breaks if there is unobserved confounding, due to the fact that some unobserved variables that have not been taken into account in the model introduce confounding.

4 Using Causal Graphs for Discrimination Analysis

With causal graphs, we have a way to formalize the effect of conditioning or intervening on a variable on other variables in the model. This will be useful with respect to fairness: we can see the effect that a sensitive attribute such as gender or race has on the other variables of the model, and on our decision rule.

Let us go back to the Berkeley example from last lecture. Let A be the sensitive attribute, gender; Z the department choice; and Y, the admission decision/classifier. Let us imagine that we have the following causal relationship, for the sake of example:

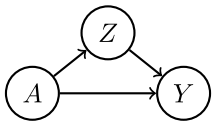


Figure 6: A possible causal graph for Berkeley admissions

So here, we assume that someone's gender A could possibly have a direct effect on admission, and there is an arrow from A to Y. But we also consider's the argument from Bickel in his study here: A might affect Y through the choice of department, in the sense that different departments are differently selective, and that gender affects the choice of department.

We can use this model to test out and put pressure on Bickel's argument, that there is no *direct* effect of sex A on admissions decisions Y, and that the reason for the data is the *indirect* effect of department choice, which should not be counted as discrimination.

4.1 Direct Effects

Here, we are interested in the direct effect of A on Y. Ideally, we want to hold Z the department choice constant to see the effect of A on Y without taking the indirect effect through Z into account.

What we are working with is observational data that tells us what someone's department was, their gender, and the admission decision. We cannot implement a do intervention here, so we have to work with conditioning on the variables instead. So we want to see the effect of A on Y *conditional on* $Z = z$ *for some* z.

In the current graph, this is fine. Z act as a mediator, and conditioning on Z does not introduce correlation between A and Y. But imagine the alternative situation described in the following figure:

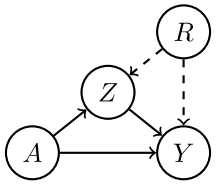


Figure 7: An alternative causal graph for Berkeley admissions

Here, we added a new variable R (say, the state of residence) that has an effect on the department you apply to and admission decisions (the school may favor candidates from within the state for example). Now Z becomes a collider between A and R, and conditioning on such a collider means that this may not be the same as holding the collider fix/performing $\text{do}(Z := z)$. Note that this would then break Bickel's argument: it may be now that the attribute A, conditional on Z, also encodes information about R, in which case we are not properly separating the effect of A from that of R. Now, it may be that there is actually a direct impact of A on Y, but we are not observing it because of the correlation with R.

Regardless, direct effect is generally a restrictive and imperfect measure of discrimination on its own, because it does not detect any form of discrimination by *proxy*. For example, I could have an alternative scenario in which Z is someone's research statement. By holding Z fixed, I am looking at the direct effect of A on Y independently of the research statement. It might very well be that the direct effect here is independent of gender, in which there is no direct discrimination; yet, I could still introduce discrimination by using someone's research statement Z to predict their gender! This is why we also need to look at indirect effects.

4.2 Indirect Effects

It could be that here the picture is a bit more complex. In fact, we could have the following: Here, we let F could denote a few different things, but they could be: how much the

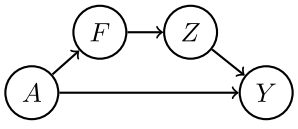


Figure 8: A more detailed causal graph for Berkeley admissions

department encourages different genders to apply; how much of a track record of hostility against women the department has; etc.

So now, if we accept Bickel's argument that this path is non-discriminatory, this means that we are ruling out all of these scenarios and deeming them impossible. Bickel claims that this path was non-discriminatory because it was indirect, but in the two new scenarios above that include F, they clearly are!

This tells you that solving fairness, even with causal models, is not purely a technical matter. This cannot be resolved without some form of knowledge of the specific matter at hand and some subjectivity. The same path could either correspond to discrimination, or the absence of discrimination. Causal modelling can expose such issues, but not always necessarily solve them.

Estimating indirect effects From a technical point of view, estimating indirect effects is not always possible given our current language and operations. In the current example, it would require to somehow disable direct influence: here, that would be disabling the edge between A and Y without disabling the indirect path. There is no way to do this with just our do-operation: we can neither hold A nor Y fixed to estimate the effect of A and Y, but we would need to do so to deactivate the direct path between A and Y.

Why do we care about understanding path-specific effects Let me look at university admissions, A is a sensitive attribute like socio-economic status, and Y is the admission decision. Now, I observe that changing A changes Y. This could mean several things:

- A influenced Y through a direct path $A \to Y$. This means that we make different admissions decisions for different people *solely* based on their socio-economic status (and not how the socio-economic status affects other attributes that are actually relevant to admission decisions).
- A influenced Y though some path $A \to X \to Y$, where X is a student's actual qualification level. Then we may argue that such a path is non-discriminatory: we hire students based on their qualification level, and it may very well be that we treat different socio-economic status differently but this is because of the fact that they have different qualifications, and we are fair in the sense that we are admitting qualified students.
- A influenced Y through some path $A \to X \to Y$, where X is whether a student can afford to take the SAT repeatedly. Then this can be seen as discriminatory: we are saying that students of a lower socio-economic status are admitted less often because they don't get to "inflate" their SAT scores (by taking the test several times until they get their desired score) as much as richer students, at everything else (and in particular maybe qualification level) equal.

(Note that I am not trying to claim that any of these reasons are the reason we see inequalities in college admission in real life and am making no personal judgment here. Those are just examples of why different paths may be seen as more or less discriminatory)

5 Counterfactual fairness

We will see how counterfactuals help us estimate path-specific effects, in a way our causal framework from before cannot. But before this, let us try to understand the following question: what is a counterfactual exactly?

Informally, the idea of counterfactuals is to answer the following question:"what would happened if I changed my model in this way? Intervened in this way? Take this action" The reason we want to run such counterfactuals here is that they can be important to understand fairness. I.e., we want to understand questions of the form "Had I changed my gender, would I have had been more likely to be admitted to college?"

Naively, it seems like we already know how to solve this from the previous section, and counterfactuals bring nothing new to the table. It seems that all I have to do is run the operation $\text{do}(X := x)$ to see how changing a given variable X changes the outcome of interest in the counterfactual, right? It turns out that this is not true; understanding counterfactuals is often not as simple as doing a single substitution in our causal model.

5.1 Example

First, we start with an example of why doing a do operation may not be enough to run a counterfactual. This example is once again taken from [BHN17].

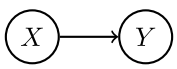
Every day, I have to decide between two routes to work, $X = 0$ and $X = 1$. On bad traffics days, indicated by the exogenous noise variable U being equal to 1, both routes are bad. On a good day however ($U = 0$), the traffic on either route is good, unless there was an accident. Let's imagine here $U \sim B(1/2)$ be the distribution of good versus bad days. Let $U_0, U_1 \sim B(1/2)$ whether an accident occurs on the route 0, respectively 1.

Now, imagine our strategy right now is to take either route at random (ignoring traffic information provided by your GPS, for example. This is 1990 and you still have to use maps to find your way). So, $X \sim B(1/2)$. Now let Y be the variable that tells us whether the traffic on the route that I chose is good or bad; i.e.,

$$Y = X \cdot \max U, U_1 + (1 - X) \cdot \max U, U_0.$$

I.e., if I choose route $X = 1$, the traffic is bad iff either $U = 1$ or $U_1 = 1$ hence $\max U, U_1 = 1$. In that case, $Y = 1 \times 1 + (1-1) \times \max U, U_0 = 1$. If I choose route $X = 0$, the traffic is bad iff either $U = 1$ or $U_0 = 1$, hence $\max U, U_0 = 1$. In that case, $Y = 0 \times maxU, U_1 + (1-0) \times 1 = 1$.

Anyway, note that since the U's are exogenous noise variables, our causal graph here is given by:



Now, one morning, I have $X = 1$ and I observe bad traffic $Y = 1$. I want to run the following counterfactual: "had I taken the other route today, would I have been better off?".

Suppose I want to answer this question by doing the most naive thing I can think of. Well, then, maybe I just need to compute the likelihood of $Y = 0$ after performing the do-operation $\text{do}(X := 0)$, i.e. $\Pr_{M[X:=0]}(Y = 0)$.

$$\Pr_{M[X:=0]}(Y = 0) = \Pr[U = 0 \cap U_0 = 0] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4},$$

where the second-to-last step follows from the fact that U and U_0 are independent. All right, we are done, right?

Well, in fact, we are not. The calculation above ignores the fact that we observed and already known that $X = 1$ and $Y = 1$. But, this restricts the possible distributions of U and U_1! In particular, it could not have been the case that both $U = 0$ and $U_1 = 0$, otherwise I would have had $Y = 0$! So, only the following settings are still possible for the noise on route 1: $U = 0, U_1 = 1$, $U = 1, U_1 = 0$, and $U = 1, U_1 = 1$. Since each of these three cases is equally likely (remember each of those have probability $1/4$), I know that *conditional on* $X = 1$ *and* $Y = 1$, they are now equally likely but with probability $1/3$. In turn, $U' = (U|(X = Y = 1) = 1$ now with probability $2/3$ instead of $1/2$!

Now we can re-do the calculations with the proper probability distribution over U; i.e., the one that is conditional on the outcome $X = Y = 1$ that I observed today. The calculation now becomes

$$\Pr_{M[X:=0]}(Y = 0) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}.$$

To summarize, taking into account what we learned from our observation (X, Y) on the state of the noise variables *decreased the probability of no traffic* if I were to choose the other route. Here is the intuitive reason: if the traffic was bad today, I am more likely to believe that this was because of too many cars on the road (the event $U = 0$ is ruled out part of the time and becomes less likely that $1/2$), in which case the other road would also be affected.

The result we just calculated is the actual *counterfactual* that we want to compute, which *takes into account what I have observed about the hidden noise in the state of the world*. The distinction with the first case is what this first case is not a counterfactual, but rather tells us "what would happen *tomorrow* if I chose the other road instead?" – in that case, you have no information about the noise variables, as their realizations change everyday. The counterfactual looks at what would have happened *today*, given the additional information that we obtained *after the fact*.

5.2 The general recipe

So, now we have to be careful about how our observations about the state of the variables that explicitly appear in our causal graph affect the unseen noise variables. We have seen above how to do so in a single example, but let us generalize this approach. We only need to implement the three following steps:

Definition 3 (Counterfactuals in causal models)**.** *Given a structural causal model M, an observed event E, an action $X := x$, and target variable Y, we define the counterfactual $Y_{X:=x}(E)$ by the following three step procedure:*

1. ***Abduction**: Adjust noise variables to be consistent with the observed events. Formally, compute the distribution U' that results from conditioning U on the observed event E.*

2. ***Action:** Perform a do-intervention on X in model M, resulting into new model $M' = M[X := x]$.*

3. ***Prediction:** Compute the target counterfactual $Y_{X:=x}(E)$ by working on new model $M' = M[X := x]$, with distribution of noise U' obtained from taking the observation E into account.*

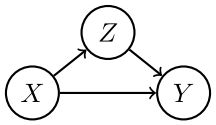
This defines what a counterfactual is in causal models. Here, note that $Y_{X:=x}(E)$ corresponds to the following counterfactual question: "what would Y be if I force $X = x$, given that I have observed and know we are in event/case E?"

A last note on counterfactuals: running a counterfactual depends heavily on the distribution of noise variables U's. These variables, however, do not appear directly in the causal graph. This means that one can construct two different causal *models* with different counterfactuals, but that correspond to the same causal *graph*. Here graph $\neq$ model.

5.3 Counterfactuals and discrimination analysis

Remember that earlier in this lecture, we talked about trying to understand how to decompose a causal effect into path-specific effects. A reason to do so is that it may be useful to determine whether discrimination occurred: some paths may be seen as discriminatory while others may not, and which paths specifically affect the outcome when changing a sensitive attribute can give you some insights as to whether discrimination occurred.

Imagine we have the following figure, and we want to understand the effect of X on Y: Here, there are two paths that we can look at: the direct path $X \to Y$, and the indirect



path $X \to Z \to Y$. Remember that earlier in this lecture, we already saw a simple way to look at the direct effect of X on Y; one way to do so is just change the value of X while holding $Z = z$ constant. I.e., we can write the direct effect as:

$$\mathbb{E}\left[Y|\text{do}(X := 1, Z := z)\right] - \mathbb{E}\left[Y|\text{do}(X := 0, Z := z)\right].$$

We can also think of this direct effect in terms of counterfactuals, i.e. as:

$$\mathbb{E}\left[Y_{X:=1,Z:=z} - Y_{X:=0,Z:=z}\right],$$

where the expectation is taken with respect to the noise variables U. This is what we call the *controlled direct effect*, since we set the mediating variable Z at a fixed value and we are controlling for Z. We can also look at the *natural direct effect*, if instead of letting $Z = z$, we

let Z follow the distribution that it was following before I changed X, i.e. the distribution it is following in the current state of the world:

$$\mathbb{E}\left[Y_{X:=1,Z:=Z_{X:=0}} - Y_{X:=0,Z:=Z_{X:=0}}\right].$$

This is nice, because a natural extension of this will allow us to now talk about natural *indirect* effects:

$$\mathbb{E}\left[Y_{X:=0,Z:=Z_{X:=1}} - Y_{X:=0,Z:=Z_{X:=0}}\right].$$

What is going on here? Well:

1. We are not changing the value of X. We are fixing X to always stay at 0. So we are not taking into account the direct effect of X on Y, since here X is held constant and so the path $X \to Y$ is held constant.

2. However, we change the distribution of Z *as if we had changed* X. Now, the edge $Z \to Y$ is acting exactly as if we had changed X and we were looking at $X \to Z \to Y$. So, we take the effect of that indirect path into account.

But so here, exactly, this means that we are *only* taking indirect effects into account, which is what we wanted!

Remark 1. *The reason we can do this now, but could not do this from our simple causal language from Section 4, is that running these indirect effects require being able to manipulate the* ***distribution*** *of the random variables. This means that we need to explicitly take the effect of our noise variables U into account, which is why we introduced the counterfactuals in the first place.*

Remark 2. *Here, we are looking at simple, toy example. But in principle, such an approach can be extended to all sort of path-specific effects in all sorts of causal graphs.*

Counterfactual discrimination criteria Now, we can start defining some of our statistical notions of fairness and discrimination, but with the additional help of the language of causal graphs and counterfactuals. Beyond path analysis and detecting discrimination, we are now explicitly defining fairness criteria on causal models.

Let us go back to the formal setup of our statistical fairness lectures. We have features X, a sensitive attribute A, an outcome variable Y, and a decision or scoring rule R. One natural criterion to look at is the following: for every possible demographic described by the event $\{X := x, A := a\}$, and every possible alternative way to set the sensitive attribute $A = a'$, we want

$$R_{A:=a}(X := x, A := a) = R_{A:=a'}(X := x, A := a),$$

i.e. they follow the same distribution. We can call this *counterfactual demographic parity*. Here, what we are saying is that given that I have observed today that your sensitive attribute is a and you features are x, I guarantee you that if you had instead had sensitive attribute a', you would have faced the same outcomes! So, we are not discriminating based on this attribute.

Remark 3. *For simplicity here, I am just looking at demographic parity. But we can similarly define similar analogues for other statistical fairness criteria.*

References

[BHN17] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.

[BHO75] 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.

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.