5

Causality

Our starting point is the difference between an observation and an action. What we see in passive observation is how individuals follow their routine behavior, habits, and natural inclination. Passive observation reflects the state of the world projected to a set of features we chose to highlight. Data that we collect from passive observation show a snapshot of our world as it is.

There are many questions we can answer from passive observation alone: Do 16 year-old drivers have a higher incidence rate of traffic accidents than 18 year-old drivers? Formally, the answer corresponds to a difference of conditional probabilities assuming we model the population as a distribution as we did in the last chapter. We can calculate the conditional probability of a traffic accident given that the driver's age is 16 years and subtract from it the conditional probability of a traffic accident given the age is 18 years. Both conditional probabilities can be estimated from a large enough sample drawn from the distribution, assuming that there are both 16 year old and 18 year old drivers. The answer to the question we asked is solidly in the realm of observational statistics.

But important questions often are not observational in nature. Would traffic fatalities decrease if we raised the legal driving age by two years? Although the question seems similar on the surface, we quickly realize that it asks for a fundamentally different insight. Rather than asking for the frequency of an event in our manifested world, this question asks for the effect of a hypothetical action.

As a result, the answer is not so simple. Even if older drivers have a lower incidence rate of traffic accidents, this might simply be a consequence of additional driving experience. There is no obvious reason why an 18 year old with two months on the road would be any less likely to be involved in an accident than, say, a 16 year-old with the same experience. We can try to address this problem by holding the number of months of driving experience fixed, while comparing individuals of different ages. But we quickly run into subtleties. What if 18 year-olds with two months of driving experience correspond to individuals who are exceptionally cautious and hence—by their natural inclination—not only drive less, but also more cautiously? What if such individuals predominantly live in regions where traffic conditions differ significantly from those in areas where people feel a greater need to drive at a younger age?

We can think of numerous other strategies to answer the original question of whether raising the legal driving age reduces traffic accidents. We could compare countries with different legal driving ages, say, the United States and Germany. But again, these countries differ in many other possibly relevant ways, such as, the legal drinking age.

At the outset, causal reasoning is a conceptual and technical framework for addressing questions about the effect of hypothetical actions or *interventions*. Once we understand what the effect of an action is, we can turn the question around and ask what action plausibly *caused* an event. This gives us a formal language to talk about cause and effect.

Not every question about cause is equally easy to address. Some questions are overly broad, such as, "What is the cause of success?" Other questions are too specific: "What caused your interest in 19th century German philosophy?" Neither question might have a clear answer. Causal inference gives us a formal language to ask these questions, in principle, but it does not make it easy to choose the right questions. Nor does it trivialize the task of finding and interpreting the answer to a question. Especially in the context of fairness, the difficulty is often in deciding what the question is that causal inference is the answer to.

In this chapter, we will develop sufficient technical understanding of causality to support at least three different purposes. The first is to conceptualize and address some limitations of the observational techniques we saw in Chapter 3. The second is to provide tools that help in the design of interventions that reliably achieve a desired effect. The third is to engage with the important normative debate about when and to which extent reasoning about discrimination and fairness requires causal understanding.

The limitations of observation

Before we develop any new formalism, it is important to understand why we need it in the first place. To see why we turn to the venerable example of graduate admissions at the University of California, Berkeley in 1973.[1] Historical data show that 12763 applicants were considered for admission to one of 101 departments and inter-departmental majors. Of the 4321 women who applied roughly 35 percent were admitted, while 44 percent of the 8442 men who applied were admitted. Standard statistical significance tests suggest that the observed difference would be highly unlikely to be the outcome of sample fluctuation if there were no difference in underlying acceptance rates.

A similar pattern exists if we look at the aggregate admission decisions of the six largest departments. The acceptance rate across all six departments for men is about 44%, while it is only roughly 30% for women, again, a significant difference. Recognizing that departments have autonomy over who to admit, we can look at the gender bias of each department.

Table 1: UC Berkeley admissions data from 1973.

	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	**82**
B	520	60	25	**68**

	Men		Women	
C	325	**37**	593	34
D	417	33	375	**35**
E	191	**28**	393	24
F	373	6	341	**7**

What we can see from the table is that four of the six largest departments show a higher acceptance ratio among women, while two show a higher acceptance rate for men. However, these two departments cannot account for the large difference in acceptance rates that we observed in aggregate. So, it appears that the higher acceptance rate for men that we observed in aggregate seems to have reversed at the department level.

Such reversals are sometimes called *Simpson's paradox*, even though mathematically they are no surprise. It's a fact of conditional probability that there can be an event Y (here, acceptance), an attribute A (here, female gender taken to be a binary variable) and a random variable Z (here, department choice) such that:

1. $\mathbb{P}\{Y \mid A\} < \mathbb{P}\{Y \mid \neg A\}$
2. $\mathbb{P}\{Y \mid A, Z = z\} > \mathbb{P}\{Y \mid \neg A, Z = z\}$ for all values z that the random variable Z assumes.

Simpson's paradox nonetheless causes discomfort to some, because intuition suggests that a trend which holds for all subpopulations should also hold at the population level.

The reason why Simpson's paradox is relevant to our discussion is that it's a consequence of how we tend to misinterpret what information conditional probabilities encode. Recall that a statement of conditional probability corresponds to passive observation. What we see here is a snapshot of the normal behavior of women and men applying to graduate school at UC Berkeley in 1973.

What is evident from the data is that gender influences department choice. Women and men appear to have different preferences for different fields of study. Moreover, different departments have different admission criteria. Some have lower acceptance rates, some higher. Therefore, one explanation for the data we see is that women *chose* to apply to more competitive departments, hence getting rejected at a higher rate than men.

Indeed, this is the conclusion the original study drew:

> *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.*[1]

In other words, the article concluded that the source of gender bias in admissions was a *pipeline problem*: Without wrongdoing by the admissions committee, women were "shunted by their socialization" that happened at an earlier stage in their lives.

It is difficult to debate this conclusion on the basis of the available data alone. The question of discrimination, however, is far from resolved. We can ask why women applied to more competitive departments in the first place. There are several possible reasons. Perhaps less competitive departments, such as engineering schools, were unwelcoming of women at the time. This may have been a general pattern at the time or specific to the university. Perhaps some departments had a track record of poor treatment of women that was known to the applicants. Perhaps the department advertised the program in a manner that discouraged women from applying.

The data we have also shows no measurement of *qualification* of an applicant. It's possible that due to self-selection women applying to engineering schools in 1973 were over-qualified relative to their peers. In this case, an equal acceptance rate between men and women might actually be a sign of discrimination.

There is no way of knowing what was the case from the data we have. There are multiple possible scenarios with different interpretations and consequences that we cannot distinguish from the data at hand. At this point, we have two choices. One is to design a new study and collect more data in a manner that might lead to a more conclusive outcome. The other is to argue over which scenario is more likely based on our beliefs and plausible assumptions about the world. Causal inference is helpful in either case. On the one hand, it can be used as a guide in the design of new studies. It can help us choose which variables to include, which to exclude, and which to hold constant. On the other hand, causal models can serve as a mechanism to incorporate scientific domain knowledge and exchange plausible assumptions for plausible conclusions.

Causal models

We will develop just enough formal concepts to engage with the technical and normative debate around causality and discrimination. The topic is much deeper than what we can explore in this chapter.

We choose *structural causal models* as the basis of our formal discussion as they have the advantage of giving a sound foundation for various causal notions we will encounter. The easiest way to conceptualize a structural causal model is as a program for generating a distribution from independent noise variables through a sequence of formal instructions. Let's unpack this statement. Imagine instead of samples from a distribution, somebody gave you a step-by-step computer program to generate samples on your own starting from a random seed. The process is not unlike how you would write code. You start from a simple random seed and build up increasingly more complex constructs. That is basically what a structural causal model is, except that each assignment uses the language of mathematics rather

than any concrete programming syntax.

A first example

Let's start with a toy example not intended to capture the real world. Imagine a hypothetical population in which an individual exercises regularly with probability $1/2$. With probability $1/3$, the individual has a latent disposition to develop overweight that manifests in the absence of regular exercise. Similarly, in the absence of exercise, heart disease occurs with probability $1/3$. Denote by X the indicator variable of regular exercise, by W that of excessive weight, and by H the indicator of heart disease. Below is a structural causal model to generate samples from this hypothetical population. To ease the description, we let $\mathrm{B}(p)$ denote a Bernoulli random variable with bias p, i.e., a biased coin toss that assumes value 1 with probability p and value 0 with probability $1 - p$.

1. Sample independent Bernoulli random variables $U_1 \sim \mathrm{B}(1/2), U_2 \sim \mathrm{B}(1/3), U_3 \sim \mathrm{B}(1/3)$.
2. $X := U_1$
3. $W :=$ if $X = 1$ then 0 else U_2
4. $H :=$ if $X = 1$ then 0 else U_3

Contrast this generative description of the population with a random sample drawn from the population. From the program description, we can immediately see that in our hypothetical population *exercise* averts both *overweight* and *heart disease*, but in the absence of exercise the two are independent. At the outset, our program generates a joint distribution over the random variables (X, W, H). We can calculate probabilities under this distribution. For example, the probability of heart disease under the distribution specified by our model is $1/2 \cdot 1/3 = 1/6$. We can also calculate the conditional probability of heart diseases given overweight. From the event $W = 1$ we can infer that the individual does not exercise so that the probability of heart disease given overweight increases to $1/3$ compared with the baseline of $1/6$.

Does this mean that overweight causes heart disease in our model? The answer is *no* as is intuitive given the program to generate the distribution. But let's see how we would go about arguing this point formally. Having a program to generate a distribution is substantially more powerful than just having sampling access. One reason is that we can manipulate the program in whichever way we want, assuming we still end up with a valid program. We could, for example, set $W := 1$, resulting in a new distribution. The resulting program looks like this:

2. $X := U_1$
3. $W := 1$
4. $H :=$ if $X = 1$ then 0 else U_3

This new program specifies a new distribution. We can again calculate the probability of heart disease under this new distribution. We still get $1/6$. This

simple calculation reveals a significant insight. The substitution $W := 1$ does not correspond to a conditioning on $W = 1$. One is an action, albeit inconsequential in this case. The other is an observation from which we can draw inferences. If we observe that an individual is overweight, we can infer that they have a higher risk of heart disease (in our toy example). However, this does not mean that lowering body weight would avoid heart disease. It wouldn't in our example. The active substitution $W := 1$ in contrast creates a new hypothetical population in which all individuals are overweight with all that it entails in our model.

Let us belabor this point a bit more by considering another hypothetical population, specified by the equations:

2. $W := U_2$
3. $X :=$ if $W = 0$ then 0 else U_1
4. $H :=$ if $X = 1$ then 0 else U_3

In this population exercise habits are driven by body weight. Overweight individuals choose to exercise with some probability, but that's the only reason anyone would exercise. Heart disease develops in the absence of exercise. The substitution $W := 1$ in this model leads to an increased probability of exercise, hence lowering the probability of heart disease. In this case, the conditioning on $W = 1$ has the same affect. Both lead to a probability of $1/6$.

What we see is that fixing a variable by substitution may or may not correspond to a conditional probability. This is a formal rendering of our earlier point that observation isn't action. A substitution corresponds to an action we perform. By substituting a value we break the natural course of action our model captures. This is the reason why the substitution operation is sometimes called the *do-operator*, written as $\text{do}(W := 1)$.

Structural causal models give us a formal calculus to reason about the effect of hypothetical actions. We will see how this creates a formal basis for all the different causal notions that we will encounter in this chapter.

Structural causal models, more formally

Formally, a structural causal model is a sequence of assignments for generating a joint distribution starting from independent noise variables. By executing the sequence of assignments we incrementally build a set of jointly distributed random variables. A structural causal model therefore not only provides a joint distribution, but also a description of how the joint distribution can be generated from elementary noise variables. The formal definition is a bit cumbersome compared with the intuitive notion.

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

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

Here, $P_i \subseteq \{X_1, ..., X_d\}$ is a subset of the variables that we call the parents *of X_i. The random variables $U_1, ..., U_d$ are called* noise variables*, which we require to be jointly independent. The* causal graph *corresponding to the structural causal model is the directed graph that has one node for each variable X_i with incoming edges from all the parents P_i.*

Let's walk through the formal concepts introduced in this definition in a bit more detail. The noise variables that appear in the definition model *exogenous factors* that influence the system. Consider, for example, how the weather influences the delay on a traffic route you choose. Due to the difficulty of modeling the influence of weather more precisely, we could take the weather induced delay to be an exogenous factor that enters the model as a noise variable. The choice of exogenous variables and their distribution can have important consequences for what conclusions we draw from a model.

The parent nodes P_i of node i in a structural causal model are often called the *direct causes* of X_i. Similarly, we call X_i the direct effect of its direct causes P_i. Recall our hypothetical population in which weight gain was determined by lack of exercise via the assignment $W := \min\{U_1, 1 - X\}$. Here we would say that exercise (or lack thereof) is a direct cause of weight gain.

Structural causal model are a collection of formal *assumptions* about how certain variables interact. Each assignment specifies a *response function*. We can think of nodes as receiving messages from their parents and acting according to these messages as well as the influence of an exogenous noise variable.

To which extent a structural causal model conforms to reality is a separate and difficult question that we will return to in more detail later. For now, think of a structural causal model as formalizing and exposing a set of assumptions about a data generating process. As such different models can expose different hypothetical scenarios and serve as a basis for discussion. When we make statements about cause and effect in reference to a model, we don't mean to suggest that these relationship necessarily hold in the real world. Whether they do depends on the scope, purpose, and validity of our model, which may be difficult to substantiate.

It's not hard to show that a structural causal model defines a unique joint distribution over the variables $(X_1, ..., X_d)$ such that $X_i = f_i(P_i, U_i)$. It's convenient to introduce a notion for probabilities under this distribution. When M denotes a structural causal model, we will write the probability of an event E under the entailed joint distribution as $\mathbb{P}_M\{E\}$. To gain familiarity with the notation, let M denote the structural causal model for the hypothetical population in which both weight gain and heart disease are directly caused by an absence of exercise. We calculated earlier that the probability of heart disease in this model is $\mathbb{P}_M\{H\} = 1/6$.

In what follows we will derive from this single definition of a structural causal model all the different notions and terminology that we'll need in this chapter. Throughout, we restrict our attention to acyclic assignments. Many real-world systems are naturally described as stateful dynamical system with closed feedback loops. There are some ways of dealing with such closed loop systems. For example, often cycles can be broken up by introducing time dependent variables, such as, investments at time 0 grow the economy at time 1 which in turn grows investments

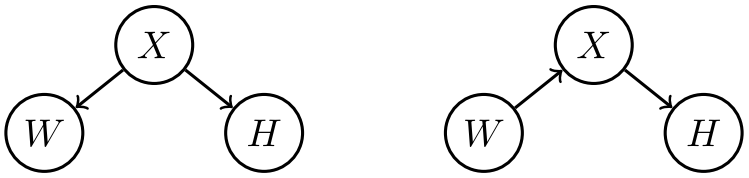

Figure 1: Causal diagrams for the heart disease examples.

at time 2, continuing so forth until some chosen time horizon t. This processing is called *unrolling* a dynamical system.

Causal graphs

We saw how structural causal models naturally give rise to *causal graphs* that represent the assignment structure of the model graphically. We can go the other way as well by simply looking at directed graphs as placeholders for an unspecified structural causal model which has the assignment structure given by the graph. Causal graphs are often called *causal diagrams*. We'll use these terms interchangeably.

The causal graphs for the two hypothetical populations from our heart disease example each have two edges and the same three nodes. They agree on the link between exercise and heart disease, but they differ in the direction of the link between exercise and weight gain.

Causal graphs are convenient when the exact assignments in a structural causal models are of secondary importance, but what matters are the paths present and absent in the graph. Graphs also let us import the established language of graph theory to discuss causal notions. We can say, for example, that an *indirect cause* of a node is any ancestor of the node in a given causal graph. In particular, causal graphs allow us to distinguish cause and effect based on whether a node is an ancestor or descendant of another node.

Let's take a first glimpse at a few important graph structures.

Forks

A *fork* is a node Z in a graph that has outgoing edges to two other variables X and Y. Put differently, the node Z is a common cause of X and Y. We already saw an example of a fork in our weight and exercise example: $W \leftarrow X \rightarrow H$. Here, exercise X influences both weight and heart disease. We also learned from the example that Z has a *confounding* effect: Ignoring exercise X, we saw that W and H appear to be positively correlated. However, the correlation is a mere result of confounding. Once we hold exercise levels constant (via the do-operation), weight has no effect on heart disease in our example.

Confounding leads to a disagreement between the calculus of conditional probabilities (observation) and do-interventions (actions). Real-world examples of confounding are a common threat to the validity of conclusions drawn from

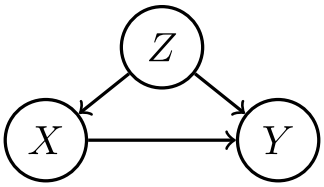

Figure 2: Example of a fork (confounder).

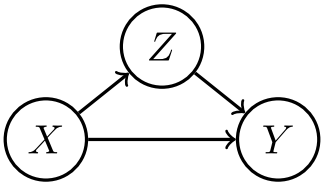

Figure 3: Example of a chain (mediator).

data. For example, in a well known medical study a suspected beneficial effect of *hormone replacement therapy* in reducing cardiovascular disease disappeared after identifying *socioeconomic status* as a confounding variable.[2]

Mediators

The case of a fork is quite different from the situation where Z lies on a directed path from X to Y. In this case, the path $X \to Z \to Y$ contributes to the total effect of X on Y. It's a causal path and thus one of the ways in which X causally influences Y. That's why Z is not a confounder. We call Z a *mediator* instead.

We saw a plausible example of a mediator in our UC Berkeley admissions example. In one plausible causal graph, department choice mediates the influences of gender on the admissions decision. The notion of a mediator is particularly relevant to the topic of discrimination analysis, since mediators can be interpreted as the mechanism behind a causal link.

Colliders

Finally, let's consider another common situation: the case of a *collider*. Colliders aren't confounders. In fact, in the above graph, X and Y are unconfounded, meaning that we can replace do-statements by conditional probabilities. However, something interesting happens when we condition on a collider. The conditioning step can create correlation between X and Y, a phenomenon called *explaining away*. A good example of the explaining away effect, or *collider bias*, is due to Berkson. Two independent diseases can become negatively correlated when analyzing hospitalized patients. The reason is that when either disease (X or Y) is sufficient for admission to the hospital (indicated by variable Z), observing that a patient has one disease makes the other statistically less likely.[3]

Berkson's law is a cautionary tale for statistical analysis when we're studying a

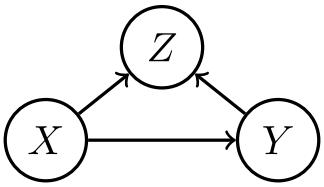

Figure 4: Example of a collider.

cohort that has been subjected to a selection rule. For example, there's an ongoing debate about the effectiveness of GRE scores in higher education. Some studies[4,5] argue that GRE scores are not predictive of various success outcomes in a graduate student population. However, care must be taken when studying the effectiveness of educational tests, such as the GRE, by examining a sample of admitted students. After all, students were in part admitted on the basis of the test score. It's the selection rule that introduces the potential for collider bias.

Interventions and causal effects

Structural causal models give us a way to formalize the effect of hypothetical actions or interventions on the population within the assumptions of our model. As we saw earlier all we needed was the ability to do substitutions.

Substitutions and the do-operator

Given a structural causal model M we can take any assignment of the form

$$X := f(P, U)$$

and replace it by another assignment. The most common substitution is to assign X a constant value x:

$$X := x$$

We will denote the resulting model by $M' = M[X := x]$ to indicate the surgery we performed on the original model M. Under this assignment we hold X constant by removing the influence of its parent nodes and thereby any other variables in the model.

Graphically, the operation corresponds to eliminating all incoming edges to the node X. The children of X in the graph now receive a fixed message x from X when they query the node's value. The assignment operator is also called the *do-operator* to emphasize that it corresponds to performing an action or intervention. We already have notation to compute probabilities after applying the do-operator, namely, $\mathbb{P}_{M[X:=x]}(E)$. Another notation is popular and common:

$$\mathbb{P}\{E \mid \text{do}(X := x)\} = \mathbb{P}_{M[X:=x]}(E)$$

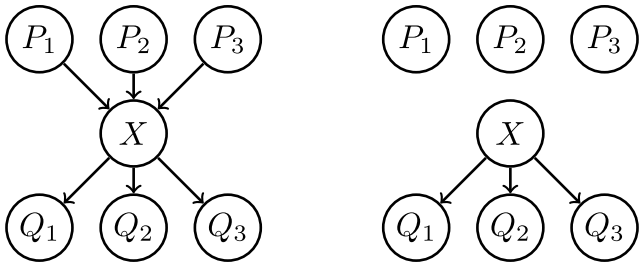

Figure 5: Graph before and after substitution.

This notation analogizes the do-operation with the usual notation for conditional probabilities, and is often convenient when doing calculations involving the do-operator. Keep in mind, however, that the do-operator (action) is fundamentally different from the conditioning operator (observation).

Causal effects

The *causal effect* of an action $X := x$ on a variable Y refers to the distribution of the variable Y in the model $M[X := x]$. When we speak of the causal effect of a variable X on another variable Y we refer to all the ways in which setting X to any possible value x affects the distribution of Y.

Often we think of X as a binary treatment variable and are interested in a quantity such as

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

This quantity is called the *average treatment effect*. It tells us how much treatment (action $X := 1$) increases the expectation of Y relative to no treatment (action $X := 0$). Causal effects are population quantities. They refer to effects averaged over the whole population. Often the effect of treatment varies greatly from one individual or group of individuals to another. Such treatment effects are called *heterogeneous*.

Confounding

Important questions in causality relate to when we can rewrite a do-operation in terms of conditional probabilities. When this is possible, we can estimate the effect of the do-operation from conventional conditional probabilities that we can estimate from data.

The simplest question of this kind asks when a causal effect $\mathbb{P}\{Y = y \mid \mathrm{do}(X := x)\}$ coincides with the condition probability $\mathbb{P}\{Y = y \mid X = x\}$. In general, this is not true. After all, the difference between observation (conditional probability) and action (interventional calculus) is what motivated the development of causality.

The disagreement between interventional statements and conditional statements is so important that it has a well-known name: *confounding*. We say that X and Y

are confounded when the causal effect of action $X := x$ on Y does not coincide with the corresponding conditional probability.

When X and Y are confounded, we can ask if there is some combination of conditional probability statements that give us the desired effect of a do-intervention. This is generally possible given a causal graph by conditioning on the parent nodes PA of the node X:

$$\mathbb{P}\{Y = y \mid \text{do}(X := x)\} = \sum_z \mathbb{P}\{Y = y \mid X = x, PA = z\}\mathbb{P}\{PA = z\}$$

This formula is called the *adjustment formula*. It gives us one way of estimating the effect of a do-intervention in terms of conditional probabilities.

The adjustment formula is one example of what is often called *controlling for* a set of variables: We estimate the effect of X on Y separately in every slice of the population defined by a condition $Z = z$ for every possible value of z. We then average these estimated sub-population effects weighted by the probability of $Z = z$ in the population. To give an example, when we control for age, we mean that we estimate an effect separately in each possible age group and then average out the results so that each age group is weighted by the fraction of the population that falls into the age group.

Controlling for more variables in a study isn't always the right choice. It depends on the graph structure. Let's consider what happens when we control for the variable Z in the three causal graphs we discussed above.

- Controlling for a confounding variable Z in a fork $X \leftarrow Z \rightarrow Y$ will deconfound the effect of X on Y.
- Controlling for a mediator Z on a chain $X \rightarrow Z \rightarrow Y$ will eliminate some of the causal influence of X on Y.
- Controlling for a collider will create correlation between X and Y. That is the opposite of what controlling for Z accomplishes in the case of a fork. The same is true if we control for a descendant of a collider.

The backdoor criterion

At this point, we might worry that things get increasingly complicated. As we introduce more nodes in our graph, we might fear a combinatorial explosion of possible scenarios to discuss. Fortunately, there are simple sufficient criteria for choosing a set of deconfounding variables that is safe to control for.

A well known graph-theoretic notion is the *backdoor* criterion.[6] Two variables are confounded if there is a so-called *backdoor* path between them. A *backdoor path* from X to Y is any path starting at X with a backward edge "$\leftarrow$" into X such as:

$$X \leftarrow A \rightarrow B \leftarrow C \rightarrow Y$$

Intuitively, backdoor paths allow information flow from X to Y in a way that is not causal. To deconfound a pair of variables we need to select a *backdoor set* of variables that "blocks" all backdoor paths between the two nodes. A backdoor path

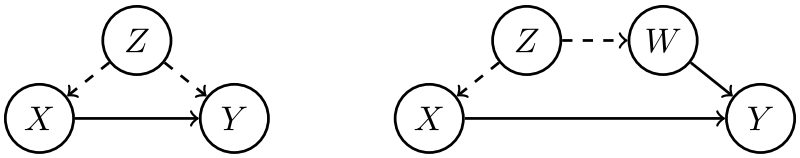

Figure 6: Two cases of unobserved confounding.

involving a chain $A \to B \to C$ can be blocked by controlling for B. Information by default cannot flow through a collider $A \to B \leftarrow C$. So we only have to be careful not to open information flow through a collider by conditioning on the collider, or descendant of a collider.

Unobserved confounding

The adjustment formula might suggest that we can always eliminate confounding bias by conditioning on the parent nodes. However, this is only true in the absence of *unobserved confounding*. In practice often there are variables that are hard to measure, or were simply left unrecorded. We can still include such unobserved nodes in a graph, typically denoting their influence with dashed lines, instead of solid lines.

The above figure shows two cases of unobserved confounding. In the first example, the causal effect of X on Y is unidentifiable. In the second case, we can block the confounding backdoor path $X \leftarrow Z \to W \to Y$ by controlling for W even though Z is not observed. The backdoor criterion lets us work around unobserved confounders in some cases where the adjustment formula alone wouldn't suffice.

Unobserved confounding nonetheless remains a major obstacle in practice. The issue is not just lack of measurement, but often lack of anticipation or awareness of a counfounding variable. We can try to combat unobserved confounding by increasing the number of variables under consideration. But as we introduce more variables into our study, we also increase the burden of coming up with a valid causal model for all variables under consideration. In practice, it is not uncommon to control for as many variables as possible in a hope to disable confounding bias. However, as we saw, controlling for mediators or colliders can be harmful.

Randomization

The backdoor criterion gives a non-experimental way of eliminating confounding bias given a causal model and a sufficient amount of observational data from the joint distribution of the variables. An alternative experimental method of eliminating confounding bias is the well-known *randomized controlled trial*.

In a *randomized controlled trial* a group of subjects is randomly partitioned into a *control group* and a *treatment group*. Participants do not know which group they were assigned to and neither do the staff administering the trial. The treatment group receives an actual treatment, such as a drug that is being tested for efficacy,

while the control group receives a placebo identical in appearance. An outcome variable is measured for all subjects.

The goal of a randomized controlled trial is to break natural inclination. Rather than observing who chose to be treated on their own, we assign treatment randomly. Thinking in terms of causal models, what this means is that we eliminate all incoming edges into the treatment variable. In particular, this closes all backdoor paths and hence avoids confounding bias.

There are many reasons why often randomized controlled trials are difficult or impossible to administer. Treatment might be physically or legally impossible, too costly, or too dangerous. As we saw, randomized controlled trials are not always necessary for avoiding confounding bias and for reasoning about cause and effect. Nor are they free of issues and pitfalls.[7]

Graphical discrimination analysis

We now explore how we can bring causal graphs to bear on discussions of discrimination. We return to the example of graduate admissions at Berkeley and develop a causal perspective on the earlier analysis.

The first step is to come up with a plausible causal graph consistent with the data that we saw earlier. The data contained only three variables, sex A, department choice Z, and admission decision Y. It makes sense to draw two arrows $A \to Y$ and $Z \to Y$, because both features A and Z are available to the institution when making the admissions decision. We'll draw one more arrow, for now, simply because we have to. If we only included the two arrows $A \to Y$ and $Z \to Y$, our graph would claim that A and Z are statistically independent. However, this claim is inconsistent with the data. We can see from the table that several departments have a statistically significant gender bias among applicants. This means we need to include either the arrow $A \to Z$ or $Z \to A$. Deciding between the two isn't as straightforward as it might first appear.

If we interpreted A in the narrowest possible sense as the applicant's *reported sex*, i.e., literally which box they checked on the application form, we could imagine a scenario where some applicants choose to (mis-)report their sex in a certain way that depends in part on their department choice. Even if we assume no misreporting occurs, it's hard to substantiate *reported sex* as a plausible cause of department choice. The fact that an applicant checked a box labeled *male* certainly isn't the cause for their interest in engineering.

The proposed causal story in the study is a different one. It alludes to a socialization and preference formation process that took place in the applicant's life before they applied which. It is this process that, at least in part, depended on the applicant's sex. To align this story with our causal graph, we need the variable A to reference whatever ontological entity it is that through this "socialization process" influences intellectual and professional preferences, and hence, department choice. It is difficult to maintain that this ontological entity coincides with sex as a biological trait. There is no scientific basis to support that the biological trait *sex* is what

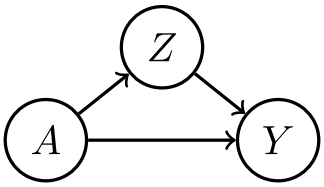

Figure 7: Possible causal graph for the UC Berkeley graduate admissions scenario.

determines our intellectual preferences. Few scholars (if any) would currently attempt to maintain a claim such as *two X chromosomes cause an interest in English literature*.

The truth is that we don't know the exact mechanism by which the thing referenced by A influences department choice. In drawing the arrow A to Z we assert—perhaps with some naivety or ignorance—that there exists such a mechanism. We will discuss the important difficulty we encountered here in depth later on. For now, we commit to this modeling choice and thus arrive at the following graph.

In this graph, department choice mediates the influence of gender on admissions. There's a direct path from A to Y and an indirect path that goes through Z. We will use this model to put pressure on the claim that *there is no evidence of sex discrimination*. In causal language, the argument had two components:

1. There appears to be no direct effect of sex A on the admissions decision Y that favors men.
2. The indirect effect of A on Y that is mediated by department choice should not be counted as evidence of discrimination.

We will discuss both arguments in turn.

Direct effects

To obtain the direct effect of A on Y we need to disable all paths between A and Y except for the direct link. In our model, we can accomplish this by holding department choice Z constant and evaluating the conditional distribution of Y given A. Recall that holding a variable constant is generally not the same as conditioning on the variable. Specifically, a problem would arise if department choice and admissions outcome were confounded by another variable, such as, state of residence R

Department choice is now a collider between A and R. Conditioning on a collider opens the backdoor path $A \to Z \leftarrow R \to Y$. In this graph, conditioning on department choice does *not* give us the desired direct effect. The real possibility that state of residence confounds department choice and decision was the subject of an exchange between Bickel and Kruskal.[8]

If we assume, however, that department choice and admissions decisions are unconfounded, then the approach Bickel, Hammel, and O'Connell took indeed

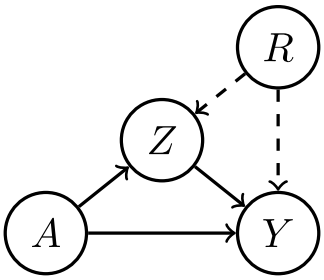

Figure 8: Alternative causal graph for the UC Berkeley graduate admissions scenario showing influence of residence.

supports the first claim. Unfortunately, the direct effect of a protected variable on a decision is a poor measure of discrimination on its own. At a technical level, it is rather brittle as it cannot detect any form of *proxy discrimination*. The department could, for example, use the applicant's personal statement to make inferences about their gender, which are then used to discriminate.

We can think of the direct effect as corresponding to the explicit *use* of the attribute in the decision rule. The absence of a direct effect loosely corresponds to the somewhat troubled notion of a *blind* decision rule that doesn't have explicit access to the sensitive attribute. As we argued in preceding chapters, blind decision rules can still be the basis of discriminatory practices.

Indirect paths

Let's turn to the indirect effect of sex on admission that goes through department choice. It's tempting to think of the the node Z as referencing the applicant's inherent department preferences. In this view, the department is not responsible for the applicant's preferences. Therefore the mediating influence of department preferences is not interpreted as a sign of discrimination. This, however, is a substantive judgment that may not be a fact. There are other plausible scenarios consistent with both the data and our causal model, in which the indirect path encodes a pattern of discrimination.

For example, the admissions committee may have advertised the program in a manner that strongly discouraged women from applying. In this case, department preference in part measures exposure to this hostile advertising campaign. Alternatively, the department could have a track record of hostile behavior against women and it is awareness of such that shapes preferences in an applicant. Finally, even blatant discriminatory practices, such as compensating women at a lower rate than equally qualified male graduate students, correspond to an indirect effect mediated by department choice.

Accepting the indirect path as *non-discriminatory* is to assert that all these scenarios we described are deemed implausible. Fundamentally, we are confronted with a substantive question. The path $A \to Z \to Y$ could either be where discrimination occurs or what explains the absence thereof. Which case we're in isn't a purely

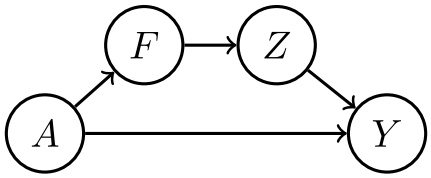

Figure 9: Alternative causal graph for the UC Berkeley graduate admissions scenario where department preferences are shaped by fear of discrimination.

technical matter and cannot be resolved without subject matter knowledge. Causal modeling gives us a framework for exposing these questions, but not necessarily one to resolve them.

Path inspection

To summarize, discrimination may not only occur on the direct pathway from the sensitive category to the outcome. Seemingly innocuous mediating paths can hide discriminatory practices. We have to carefully discuss what pathways we consider evidence for or against discrimination.

To appreciate this point, contrast our Berkeley scenario with the important legal case *Griggs v. Duke Power Co.* that was argued before the U.S. Supreme Court in 1970. Duke Power Company had introduced the requirement of a high school diploma for certain higher paying jobs. We could draw a causal graph for this scenario not unlike the one for the Berkeley case. There's a mediating variable (here, level of education), a sensitive category (here, race) and an employment outcome (here, employment in a higher paying job). The company didn't directly make employment decisions based on race, but rather used the mediating variable. The court ruled that the requirement of a high school diploma was not justified by business necessity, but rather had adverse impact on ethnic minority groups where the prevalence of high school diplomas is lower. Put differently, the court decided that the use of this mediating variable was not an argument against, but rather for discrimination.

Glymour[9] makes another related and important point about the moral character of mediation analysis:

> Implicitly, the question of what mediates observed social effects informs our view of which types of inequalities are socially acceptable and which types require remediation by social policies. For example, a conclusion that women are "biologically programmed" to be depressed more than men may ameliorate the social obligation to try to reduce gender inequalities in depression. Yet if people get depressed whenever they are, say, sexually harassed—and women are more frequently sexually harassed than men—this suggests a very strong social obligation to reduce the depression disparity by reducing the sexual harassment disparity.

Ending on a technical note, we currently do not have a method to estimate indirect effects. Estimating an indirect effect somehow requires us to *disable* the direct influence. There is no way of doing this with the do-operation that we've seen so far. However, we will shortly introduce *counterfactuals*, which among other applications will give us a way of estimating path-specific effects.

Structural discrimination

There's an additional problem we neglected so far. Imagine a spiteful university administration that systematically defunds graduate programs that attract more female applicants. This structural pattern of discrimination is invisible from the causal model we drew. There is a kind of type mismatch here. Our model talks about individual applicants, their department preferences, and their outcomes. Put differently, individuals are the *units* of our investigation. University policy is not one of the mechanisms that our model exposes. We cannot *featurize* university policy to make it an attribute of the individual. As a result we cannot talk about university policy as a cause of discrimination in our model.

The model we chose commits us to an individualistic perspective that frames discrimination as the consequence of how decision makers respond to information about individuals. An analogy is helpful. In epidemiology, scientists can seek the cause of health outcomes in biomedical aspects and lifestyle choices of individuals, such as whether or not an individual smokes, exercises, maintains a balanced diet etc. The growing field of social epidemiology criticizes the view of individual choices as causes of health outcomes, and instead draws attention to social and structural causes,[10] such as poverty and inequality.

Similarly, we can contrast the individualistic perspective on discrimination with structural discrimination. Causal modeling can in principle be used to study the causes of structural discrimination, as well. But it requires a different perspective than the one we chose for our Berkeley scenario.

Counterfactuals

Fully specified structural causal models allow us to ask causal questions that are more delicate than the mere effect of an action. Specifically, we can ask *counterfactual* questions such as: Would I have avoided the traffic jam had I taken a different route this morning? Counterfactual questions are common and relevant for questions of discrimination. We can computer the answer to counterfactual questions given a structural causal model. The procedure for extracting the answer from the model looks a bit subtle at first. We'll walk through the formal details starting from a simple example before returning to our discussion of discrimination.

A simple counterfactual

To understand counterfactuals, we first need to convince ourselves that they aren't quite as straightforward as a single substitution in our model.

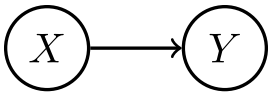

Figure 10: Causal diagram for our traffic scenario.

Assume every morning we need to decide between two routes $X = 0$ and $X = 1$. On bad traffic days, indicated by $U = 1$, both routes are bad. On good days, indicated by $U = 0$, the traffic on either route is good unless there was an accident on the route. Let's say that $U \sim B(1/2)$ follows the distribution of an unbiased coin toss. Accidents occur independently on either route with probability $1/2$. So, choose two Bernoulli random variables $U_0, U_1 \sim B(1/2)$ that tell us if there is an accident on route 0 and route 1, respectively. We reject all external route guidance and instead decide on which route to take uniformly at random. That is, $X := U_X \sim B(1/2)$ is also an unbiased coin toss.

Introduce a variable $Y \in \{0, 1\}$ that tells us whether the traffic on the chosen route is good ($Y = 0$) or bad ($Y = 1$). Reflecting our discussion above, we can express Y as

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

In words, when $X = 0$ the first term disappears and so traffic is determined by the larger of the two values U and U_0. Similarly, when $X = 1$ traffic is determined by the larger of U and U_1.

Now, suppose one morning we have $X = 1$ and we observe bad traffic $Y = 1$. Would we have been better off taking the alternative route this morning?

A natural attempt to answer this question is to compute the likelihood of $Y = 0$ after the do-operation $X := 0$, that is, $\mathbb{P}_{M[X:=0]}(Y = 0)$. A quick calculation reveals that this probability is $\frac{1}{2} \cdot \frac{1}{2} = 1/4$. Indeed, given the substitution $X := 0$ in our model, for the traffic to be good we need that $\max\{U, U_0\} = 0$. This can only happen when both $U = 0$ (probability $1/2$) and $U_0 = 0$ (probability $1/2$).

But this isn't the correct answer to our question. The reason is that we took route $X = 1$ and observed that $Y = 1$. From this observation, we can deduce that certain background conditions did not manifest for they are inconsistent with the observed outcome. Formally, this means that certain settings of the noise variables (U, U_0, U_1) are no longer feasible given the observed event $\{Y = 1, X = 1\}$. Specifically, if U and U_1 had both been zero, we would have seen no bad traffic on route $X = 1$, but this is contrary to our observation. In fact, the available evidence $\{Y = 1, X = 1\}$ leaves only the following settings for U and U_1:

Table 2: Possible noise settings after observing evidence

U	U_1
0	1
1	1
1	0

U	U_1

We leave out U_0 from the table, since its distribution is unaffected by our observation. Each of the remaining three cases is equally likely, which in particular means that the event $U = 1$ now has probability $2/3$. In the absence of any additional evidence, recall, $U = 1$ had probability $1/2$. What this means is that the observed evidence $\{Y = 1, X = 1\}$ has biased the distribution of the noise variable U toward 1. Let's use the letter U' to refer to this biased version of U. Formally, U' is distributed according to the distribution of U conditional on the event $\{Y = 1, X = 1\}$.

Working with this biased noise variable, we can again entertain the effect of the action $X := 0$ on the outcome Y. For $Y = 0$ we need that $\max\{U', U_0\} = 0$. This means that $U' = 0$, an event that now has probability $1/3$, and $U_0 = 0$ (probability $1/2$ as before). Hence, we get the probability $1/6 = 1/2 \cdot 1/3$ for the event that $Y = 0$ under our do-operation $X := 0$, and after updating the noise variables to account for the observation $\{Y = 1, X = 1\}$.

To summarize, incorporating available evidence into our calculation decreased the probability of no traffic ($Y = 0$) when choosing route 0 from $1/4$ to $1/6$. The intuitive reason is that the evidence made it more likely that it was generally a bad traffic day, and even the alternative route would've been clogged. More formally, the event that we observed biases the distribution of exogenous noise variables.

We think of the result we just calculated as the *counterfactual* of choosing the alternative route given the route we chose had bad traffic.

The general recipe

We can generalize our discussion of computing counterfactuals from the previous example to a general procedure. There were three essential steps. First, we incorporated available observational evidence by biasing the exogenous noise variables through a conditioning operation. Second, we performed a do-operation in the structural causal model after we substituted the biased noise variables. Third, we computed the distribution of a target variable. These three steps are typically called *abduction*, *action*, and *prediction*, as can be described as follows.

Definition 2. *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 event. Formally, condition the joint distribution of $U = (U_1, ..., U_d)$ on the event E. This results in a biased distribution U'.*
2. ***Action:*** *Perform do-intervention $X := x$ in the structural causal model M resulting in the model $M' = M[X := x]$.*
3. ***Prediction:*** *Compute target counterfactual $Y_{X:=x}(E)$ by using U' as the random seed in M'.*

It's important to realize that this procedure *defines* what a counterfactual is in a structural causal model. The notation $Y_{X:=x}(E)$ denotes the outcome of the procedure and is part of the definition. We haven't encountered this notation before. Put in words, we interpret the formal counterfactual $Y_{X:=x}(E)$ as the value Y would've taken had the variable X been set to value x in the circumstances described by the event E.

In general, the counterfactual $Y_{X:=x}(E)$ is a random variable that varies with U'. But counterfactuals can also be deterministic. When the event E narrows down the distribution of U to a single point mass, called *unit*, the variable U' is constant and hence the counterfactual $Y_{X:=x}(E)$ reduces to a single number. In this case, it's common to use the shorthand notation $Y_x(u) = Y_{X:=x}(\{U = u\})$, where we make the variable X implicit, and let u refer to a single unit.

The motivation for the name *unit* derives from the common situation where the structural causal model describes a population of entities that form the atomic units of our study. It's common for a unit to be an individual (or the description of a single individual). However, depending on application, the choice of units can vary. In our traffic example, the noise variables dictate which route we take and what the road conditions are.

Answers to counterfactual questions strongly depend on the specifics of the structural causal model, including the precise model of how the exogenous noise variables come into play. It's possible to construct two models that have identical graph structures, and behave identically under interventions, yet give different answers to counterfactual queries.[11]

Potential outcomes

The *potential outcomes* framework is a popular formal basis for causal inference, which goes about counterfactuals differently. Rather than deriving them from a structural causal model, we assume their existence as ordinary random variables, albeit some unobserved.

Specifically, we assume that for every unit u there exist random variables $Y_x(u)$ for every possible value of the assignment x. In the potential outcomes model, it's customary to think of a binary *treatment* variable X so that x assumes only two values, 0 for *untreated*, and 1 for *treated*. This gives us two potential outcome variables $Y_0(u)$ and $Y_1(u)$ for each unit u. There is some potential for notational confusion here. Readers familiar with the potential outcomes model may be used to the notation "$Y_i(0), Y_i(1)$" for the two potential outcomes corresponding to unit i. In our notation the unit (or, more generally, set of units) appears in the parentheses and the subscript denotes the substituted value for the variable we intervene on.

The key point about the potential outcomes model is that we only observe the potential outcome $Y_1(u)$ for units that were treated. For untreated units we observe $Y_0(u)$. In other words, we can never simultaneously observe both, although they're both assumed to exist in a formal sense. Formally, the outcome $Y(u)$ for unit u that we observe depends on the binary treatment $T(u)$ and is given by the expression:

$$Y(u) = Y_0(u) \cdot (1 - T(u)) + Y_1(u) \cdot T(u)$$

It's often convenient to omit the parentheses from our notation for counterfactuals so that this expression would read $Y = Y_0 \cdot (1 - T) + Y_1 \cdot T$.

We can revisit our traffic example in this framework. The next table summarizes what information is observable in the potential outcomes model. We think of the route we choose as the treatment variable, and the observed traffic as reflecting one of the two potential outcomes.

Table 3: Traffic example in the potential outcomes model

Route X	Outcome Y_0	Outcome Y_1	Probability
0	0	?	1/8
0	1	?	3/8
1	?	0	1/8
1	?	1	3/8

Often this information comes in the form of samples. For example, we might observe the traffic on different days. With sufficiently many samples, we can estimate the above frequencies with arbitrary accuracy.

Table 4: Traffic data in the potential outcomes model

Day	Route X	Outcome Y_0	Outcome Y_1
1	0	1	?
2	0	0	?
3	1	?	1
4	0	1	?
5	1	?	0
...	...	...	...

A typical query in the potential outcomes model is the *average treatment effect* $\mathbb{E}[Y_1 - Y_0]$. Here the expectation is taken over the properly weighted units in our study. If units correspond to equally weighted individuals, the expectation is an average over these individuals.

In our original traffic example, there were 16 units corresponding to the background conditions given by the four binary variables U, U_0, U_1, U_X. When the units in the potential outcome model agree with those of a structural causal model, then causal effects computed in the potential outcomes model agree with those computed in the structural equation model. The two formal frameworks are perfectly consistent with each other.

As is intuitive from the table above, causal inference in the potential outcomes framework can be thought of as filling in the missing entries ("?") in the table

above. This is sometimes called *missing data imputation* and there are numerous statistical methods for this task. If we could *reveal* what's behind the question marks, estimating the average treatment effect would be as easy as counting rows.

There is a set of established conditions under which causal inference becomes possible:

1. **Stable Unit Treatment Value Assumption** (SUTVA): The treatment that one unit receives does not change the effect of treatment for any other unit.
2. **Consistency**: Formally, $Y = Y_0(1 - T) + Y_1 T$. That is, $Y = Y_0$ if $T = 0$ and $Y = Y_1$ if $T = 1$. In words, the outcome Y agrees with the potential outcome corresponding to the treatment indicator.
3. **Ignorability**: The potential outcomes are independent of treatment given some deconfounding variables Z, i.e., $T \perp (Y_0, Y_1) \mid Z$. In words, the potential outcomes are conditionally independent of treatment given some set of deconfounding variables.

The first two assumptions automatically hold for counterfactual variables derived from structural causal models according to the procedure described above. This assumes that the units in the potential outcomes framework correspond to the atomic values of the background variables in the structural causal model.

The third assumption is a major one. It's easiest to think of it as aiming to formalize the guarantees of a perfectly executed randomized controlled trial. The assumption on its own cannot be verified or falsified, since we never have access to samples with both potential outcomes manifested. However, we can verify if the assumption is consistent with a given structural causal model by checking if the set Z blocks all backdoor paths from treatment T to outcome Y.

There's no tension between structural causal models and potential outcomes and there's no harm in having familiarity with both. It nonetheless makes sense to say a few words about the differences of the two approaches.

We can derive potential outcomes from a structural causal model as we did above, but we cannot derive a structural causal model from potential outcomes alone. A structural causal model in general encodes more assumptions about the relationships of the variables. This has several consequences. On the one hand, a structural causal model gives us a broader set of formal concepts (causal graphs, mediating paths, counterfactuals for every variable, and so on). On the other hand, coming up with a plausibly valid structural causal model is often a daunting task that might require knowledge that is simply not available. We will dive deeper into questions of validity below. Difficulty to come up with a plausible causal model often exposes unsettled substantive questions that require resolution first.

The potential outcomes model, in contrast, is generally easier to apply. There's a broad set of statistical estimators of causal effects that can be readily applied to observational data. But the ease of application can also lead to abuse. The assumptions underpinning the validity of such estimators are experimentally unverifiable. Frivolous application of causal effect estimators in situations where crucial assumptions do not hold can lead to false results, and consequently to ineffective or harmful interventions.

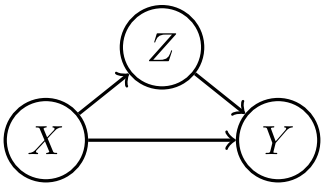

Figure 11: Causal graph with mediator Z.

Counterfactual discrimination analysis

Counterfactuals serve at least two purposes for us. On the technical side, counterfactuals give us a way to compute path-specific causal effects. This allows us to make path analysis a quantitative matter. On the conceptual side, counterfactuals let us engage with the important normative debate about whether discrimination can be captured by counterfactual criteria. We will discuss each of these in turn.

Quantitative path analysis

Mediation analysis is a venerable subject dating back decades.[12] Generally speaking, the goal of mediation analysis is to identify a mechanism through which a cause has an effect. We will review some recent developments and how they relate to questions of discrimination.

In the language of our formal framework, mediation analysis aims to decompose a total causal effect into path-specific components. We will illustrate the concepts in the basic three variable case of a mediator, although the ideas extend to more complicated structures.

There are two different paths from X to Y. A direct path and a path through the mediator Z. The conditional expectation $\mathbb{E}[Y \mid X = x]$ lumps together influence from both paths. If there were another confounding variable in our graph influencing both X and Y, then the conditional expectation would also include whatever correlation is the result of confounding. We can eliminate the confounding path by virtue of the do-operator $\mathbb{E}[Y \mid \mathrm{do}(X := x)]$. This gives us the total effect of the action $X := x$ on Y. But the total effect still conflates the two causal pathways, the direct effect and the indirect effect. We will now see how we can identify the direct and indirect effects separately.

The direct effect we already dealt with earlier as it did not require any counterfactuals. Recall, we can hold the mediator fixed at level $Z := z$ and consider the effect of treatment $X := 1$ compared with no treatment $X := 0$ as follows:

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

We can rewrite this expression in terms of counterfactuals equivalently as:

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

To be clear, the expectation is taken over the background variables in our structural causal models. In other words, the counterfactuals inside the expectation are invoked with an elementary setting u of the background variables, i.e., $Y_{X:=1,Z:=z}(u) - Y_{X:=0,Z:=x}(u)$ and the expectation averages over all possible settings.

The formula for the direct effect above is usually called *controlled direct effect*, since it requires setting the mediating variable to a specified level. Sometimes it is desirable to allow the mediating variable to vary as it would had no treatment occurred. This too is possible with counterfactuals and it leads to a notion called *natural direct effect*, defined as:

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

The counterfactual $Y_{X:=1,Z:=Z_{X:=0}}$ is the value that Y would obtain had X been set to 1 and had Z been set to the value Z would've assumed had X been set to 0.

The advantage of this slightly mind-bending construction is that it gives us an analogous notion of *natural indirect effect*:

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

Here we hold the treatment variable constant at level $X := 0$, but let the mediator variable change to the value it would've attained had treatment $X := 1$ occurred.

In our three node example, the effect of X on Y is unconfounded. In the absence of confounding, the natural indirect effect corresponds to the following statement of conditional probability (involving neither counterfactuals nor do-interventions):

$$\sum_z \mathbb{E}\left[Y \mid X = 0, Z = z\right]\left(\mathbb{P}(Z = z \mid X = 1) - \mathbb{P}(Z = z \mid X = 0)\right) .$$

In this case, we can estimate the natural direct and indirect effect from observational data.

The technical possibilities go beyond the case discussed here. In principle, counterfactuals allow us to compute all sorts of path-specific effects even in the presence of (observed) confounders. We can also design decision rules that eliminate path-specific effects we deem undesirable.

Counterfactual discrimination criteria

Beyond their application to path analysis, counterfactuals can also be used as a tool to put forward normative fairness criteria. Consider the typical setup of Chapter 3. We have features X, a sensitive attribute A, an outcome variable Y and a predictor $\widehat{Y}$.

One criterion that is technically natural would say the following: For every possible demographic described by the event $E := \{X := x, A := a\}$ and every possible setting a' of A we ask that the counterfactual $\widehat{Y}_{A:=a}(E)$ and the counterfactual $\widehat{Y}_{A:=a'}(E)$ follow the same distribution.

Introduced as *counterfactual fairness*,[13] we refer to this condition as *counterfactual demographic parity*, since it's closely related to the observational criterion *conditional demographic parity*. Recall, conditional demographic parity requires that in each demographic defined by a feature setting $X = x$, the sensitive attribute is independent of the predictor. Formally, we have the conditional independence relation $\widehat{Y} \perp A \mid X$. In the case of a binary predictor, this condition is equivalent to requiring for all feature settings x and groups a, a':

$$\mathbb{E}[\widehat{Y} \mid X = x, A = a] = \mathbb{E}[\widehat{Y} \mid X = x, A = a']$$

The easiest way to satisfy counterfactual demographic parity is for the predictor $\widehat{Y}$ to only use non-descendants of A in the causal graph. This is analogous to the statistical condition of only using features that are independent of A.

In the same way that we defined a counterfactual analog of demographic parity, we can explore causal analogs of other statistical criteria in Chapter 3. In doing so, we need to be careful in separating technical questions about the difference between observational and causal criteria from the normative content of the criterion. Just because a causal variant of a criterion might get around some statistical issues of non-causal correlations does not mean that the causal criterion resolves normative concerns or questions with its observational cousin.

Counterfactuals in the law

We'll now scratch the surface of a deep subject in legal scholarship that we return to in Chapter 6 after developing greater familiarity with the legal background. The subject is the relationship of causal counterfactual claims and legal cases of discrimination. Many technical scholars see support for a counterfactual interpretation of United States discrimination law in various rulings by judges that seemed to have invoked counterfactual language. Here's a quote from a popular textbook on causal inference:[14]

> U.S. courts have issued clear directives as to what constitutes employment discrimination. According to law makers, "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (In Carson vs Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996).)

Unfortunately, the situation is not so simple. This quote invoked here—and in several other technical papers on the topic—expresses the opinion of judges in the 7th Circuit Court at the time. This court is one of thirteen United States courts of appeals. The case has little precedential value; the quote cannot be considered a definitive statement on what employment discrimination means under either Title VII or Equal Protection law.

More significant in U.S. jurisprudence is the standard of "but-for causation" that has gained support through a 2020 U.S. Supreme Court decision relating to

sex discrimination in the case Bostock v. Clayton County. In reference to the Title VII statute about employment discrimination in the Civil Rights Act of 1964, the court argued:

> While the statute's text does not expressly discuss causation, it is suggestive. The guarantee that each person is entitled to the 'same right ... as is enjoyed by White citizens' directs our attention to the counterfactual—what would have happened if the plaintiff had been White? This focus fits naturally with the ordinary rule that a plaintiff must prove but-for causation.

Although the language of counterfactuals appears here, the notion of but-for causation may not effectively correspond to a correct causal counterfactual. Expanding on how to interpret but-for causation, the court noted:

> a but-for test directs us to change one thing at a time and see if the outcome changes. If it does, we have found a but-for cause.

Changing one attribute while holding all others fixed is not in general a correct way of computing counterfactuals in a causal graph. This important issue was central to an major discrimination lawsuit.

Harvard college admissions

In a trial dating back to 2015, the plaintiff *Students for Fair Admissions* (SFFA) allege discrimination in Harvard undergraduate admissions against Asian-Americans. Plaintiff SFFA is an offshoot of a legal defense fund which aims to end the use of race in voting, education, contracting, and employment.

The trial entailed unprecedented discovery regarding higher education admissions processes and decision-making, including statistical analyses of individual-level applicant data from the past five admissions cycles.

The plaintiff's expert report by Peter S. Arcidiacono, Professor of Economics at Duke University, claims:

> Race plays a significant role in admissions decisions. Consider the example of an Asian-American applicant who is male, is not disadvantaged, and has other characteristics that result in a 25% chance of admission. Simply changing the race of the applicant to white—and leaving all his other characteristics the same—would increase his chance of admission to 36%. Changing his race to Hispanic (and leaving all other characteristics the same) would increase his chance of admission to 77%. Changing his race to African-American (again, leaving all other characteristics the same) would increase his chance of admission to 95%.

The plaintiff's charge, summarized above, is based technically on the argument that conditional statistical parity is not satisfied by a model of Harvard's admissions

decisions. Harvard's decision process isn't codified as a formal decision rule. Hence, to talk about Harvard's decision rule formally, we first need to model Harvard's decision rule. The plaintiff's expert did so by fitting a logistic regression model against Harvard's past admissions decisions in terms of variables deemed relevant for the admission decision.

Formally, denote by $\widehat{Y}$ the model of Harvard's admissions decisions, by X a set of applicant features deemed relevant for admission, and denoting by A the applicant's reported race we have that

$$\mathbb{E}[\widehat{Y} \mid X = x, A = a] < \mathbb{E}[\widehat{Y} \mid X = x, A = a'] - \delta\,,$$

for some groups a, a' and some significant value of $\delta > 0$.

The violation of this condition certainly depends on which features we deem relevant for admissions, formally, which features X we should condition on. Indeed, this point is to a large extent the basis of the response of the defendant's expert David Card, Professor of Economics at the University of California, Berkeley. Card argues that under a different reasonable choice of X, one that includes among other features the applicant's interview performance and the year they applied in, the observed disparity disappears.

The selection and discussion of what constitute relevant features is certainly important for the interpretation of conditional statistical parity. But arguably a bigger question is whether a violation of conditional statistical parity constitutes evidence of discrimination in the first place. This isn't merely a question of having selected the right features to condition on.

What is it the plaintiff's expert report means by "changing his race"? The literal interpretation is to "flip" the race attribute in the input to the model without changing any of the other features of the input. But a formal interpretation in terms of attribute swapping is not necessarily what triggers our moral intuition. As we know now, attribute flipping generally does not produce valid counterfactuals. Indeed, if we assume a causal graph in which some of the relevant features are influenced by race, then computing counterfactuals with respect to race would require adjusting downstream features. Changing the race attribute without a change in any other attribute only corresponds to a counterfactual in the case where race does not have any descendant nodes—an implausible assumption.

Attribute flipping is often mistakenly given a counterfactual causal interpretation. Obtaining valid counterfactuals is in general substantially more involved than flipping a single attribute independently of the others. In particular, we cannot meaningfully talk about counterfactuals without bringing clarity to what exactly we refer to in our causal model and how we can produce *valid* causal models. We turn to this important topic next.

Validity of causal modeling

Consider a claim of employment discrimination of the kind: *The company's hiring practices discriminated against applicants of a certain religion.* Suppose we want to

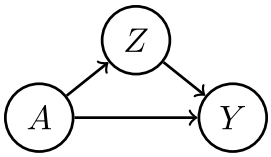

Figure 12: Religion as a root node.

interrogate this claim using the formal machinery developed in this chapter. At the outset, this requires that we formally introduce an attributed corresponding to the "religious affiliation" of an individual.

Our first attempt is to model *religious affiliation* as a personal trait or characteristic that someone either does or does not possess. This trait, call it A, may influence choices relating to one's appearance, social practices, and variables relevant to the job, such as, the person's level of education Z. So, we might like to start with a model such as the following:

Religious affiliation A is a source node in this graph, which influences the person's level of education Z. Members of certain religions may be steered away from or encouraged towards obtaining a higher level of education by their social peer group. This story is similar to how in our Berkeley admissions graph *sex* influences *department choice*.

This view of religion places burden on understanding the possible indirect pathways, such as $A \to Z \to Y$, through which religion can influence the outcome. There may be insufficient understanding of how a religious affiliation affects numerous other relevant variables throughout life. If we think of religion as a source node in a causal graph, changing it will potentially affect all downstream nodes. For each such downstream node we would need a clear understanding of the mechanisms by which religion influence the node. Where would such *scientific knowledge* of such relationships come from?

But the causal story around religion might also be different. It could be that obtaining a higher level of education causes an individual to lose their religious beliefs. In fact, this modeling choice has been put forward in technical work on this topic.[15] Empirically, data from the United States General Social Survey show that the fraction of respondents changing their reported religion at least once during a 4-year period ranged from about 20% to about 40%.[16] Identities associated with sexuality and social class were found to be even more unstable. Changing one's identity to better align with one's politics appeared to explain some of this shift. From this perspective, religious affiliation is influenced by level of education and so the graph might look like this:

This view of religion forces us to correctly identify the variables that influence religious affiliation and are also relevant to the decision. After all, these are the confounders between religion and outcome. Perhaps it is not just level of education, but also socioeconomic status and other factors that have a similar confounding influence.

What is troubling is that in our first graph education is a mediator, while in

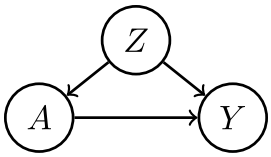

Figure 13: Religion as ancestor.

our second graph it is a confounder. The difference is important; to quote Pearl:

> As you surely know by now, mistaking a mediator for a confounder is one of the deadliest sins in causal inference and may lead to the most outrageous error. The latter invites adjustment; the former forbids it.[8]

The point is not that these are the only two possible modeling choices for how religious affiliation might interact with decision making processes. Rather, the point is that there exist multiple plausible choices. Either of our modeling choices follows a natural causal story. Identifying which one is justified is no easy task. It's also not a task that we can circumvent by appeal to some kind of pragmatism. Different modeling choices can lead to completely different claims and consequences.

In order to create a valid causal model, we need to provide clarity about what the thing is that each node references, and what relationships exist between these things. This is a problem of ontology and metaphysics. But we also need to know facts about the things we reference in causal models. This is a problem is epistemology, the theory of knowledge.

These problems might seem mundane for some objects of study. We might have strong scientifically justified beliefs on how certain mechanical parts in an airplane interact. We can use this knowledge to reliably diagnose the cause of an airplane crash. In other domains, especially ones relevant to disputes about discrimination, our subject matter knowledge is less stable and subject to debate.

Social construction of categories

The difficulties we encountered in our motivating example arise routinely when making causal statements involving human kinds and categories, such as, race, religion, or gender, and how these interact with consequential decisions.

Consider the case of *race*. The metaphysics of race is a complex subject, highly debated, featuring a range of scholarly accounts today. A book by Glasgow, Haslanger, Jeffers, and Spencer represents four contemporary philosophical views of what race is.[17] The construction of racial categories and racial classification of individuals is inextricably tied to a long history of oppression, segregation, and discriminatory practices.[18,19,20]

In the technical literature around discrimination and causality, it's common for researchers to model *race* as a source node in a causal graph, which is to say that race has no incoming arrows. As a source node it can directly and indirectly

influence an outcome variable, say, *getting a job offer*. Implicit in this modeling choice is a kind of naturalistic perspective that views race as a biologically grounded trait, similar to *sex*. The trait exists at the beginning of one's life. Other variables that come later in life, education and income, for example, thus become ancestors in the causal graph.

This view of race challenges us to identify all the possible indirect pathways through which race can influence the outcome. But it's not just this modeling challenge that we need to confront. The view of race as a biologically grounded trait stands in contrast with the *social constructivist* account of race.[21,22,23,17] In this view, roughly speaking, race has no strong biological grounding but rather is a social construct. Race stems from a particular classification of individuals by society, and the shared experiences that stem from the classification. As such, the surrounding social system of an individual influences what race is and how it is perceived. In the constructivist view, *race* is a socially constructed category that individuals are assigned to.

The challenge with adopting this view is that it is difficult to tease out a set of nodes that faithfully represent the influence that society has on race, and perceptions of race. The social constructivist perspective does not come with a simple operational guide for identifying causal structures. In particular, socially constructed categories often lack the kind of modularity that a causal diagram requires. Suppose that group membership is constructed from a set of social facts about the group and practices of individuals within the group. We might have some understanding of how these facts and practices constitutively identify group membership. But we may not have an understanding of how each factor individually interacts with each other factor, or whether such a decomposition is even possible.[24]

Ontological instability

The previous arguments notwithstanding, pragmatist might accuse our discussion of adding unnecessary complexity to what might seem like a matter of common sense to some. Surely, we could also find subtlety in other characteristics, such as, smoking habits or physical exercise. How is race different from other things we reference in causal models?

An important difference is a matter of ontological stability. When we say *rain caused the grass to be wet* we also refer to an implicit understanding of what rain is, what grass is, and what wet means. However, we find that acceptable in this instance, because all three things we refer to in our causal statement have *stable enough* ontologies. We know what we reference when we invoke them. To be sure, there could be subtleties in what we call grass. Perhaps the colloquial term *grass* does not correspond to a precise botanical category, or one that has changed over time and will again change in the future. However, by making the causal claim, we implicitly assert that these subtleties are irrelevant for the claim we made. We know that grass is a plant and that other plants would also get wet from rain. In short, we believe the ontologies we reference are *stable enough* for the claim we

make.

This is not always an easy judgment to make. There are, broadly speaking, at least two sources of ontological instability. One stems from the fact that the world changes over time. Both social progress, political events, and our own epistemic activities may obsolete theories, create new categories, or disrupt existing ones.[23] Hacking's work describes another important source of instability. Categories lead people who putatively fall into such categories to change their behavior in possibly unexpected ways. Individuals might conform or disconform to the categories they are confronted with. As a result, the responses of people, individually or collectively, invalidate the theory underlying the categorization. Hacking calls this a "looping effect".[25] As such, social categories are moving targets that need constant revision.

Certificates of ontological stability

The debate around human categories in causal models is by no means new. But it often surfaces in a seemingly unrelated, yet long-standing discussion around causation and manipulation. One school of thought in causal inference aligns with the mantra *no causation without manipulation*, a view expressed by Holland in an influential article from 1986:

> Put as bluntly and as contentiously as possible, in this article I take the position that causes are only those things that could, in principle, be treatments in experiments.[26]

Holland goes further by arguing that statements involving "attributes" are necessarily statements of association:

> The only way for an attribute to change its value is for the unit to change in some way and no longer be the same unit. Statements of "causation" that involve attributes as "causes" are always statements of association between the values of an attribute and a response variable across the units in a population.[26]

To give an example, Holland maintains that the sentence "She did well on the exam because she is a woman" means nothing but "the performance of women on the exam exceeds, in some sense, that of men."[26]

If we believed that there is no causation without manipulation, we would have to refrain from including immutable characteristics in causal models altogether. After all, there is by definition no experimental mechanism that turns immutable attributes into treatments.

Holland's view remains popular among practitioners of the potential outcomes model. The assumptions common in the potential outcomes model are easiest to conceptualize by analogy with a well-designed randomized trial. Practitioners in this framework are therefore used to conceptualizing causes as things that could, in principle, be a treatment in randomized controlled trials.

The desire or need to make causal statements involving race in one way or the other not only arises in the context of discrimination. Epidemiologists encounter the same difficulties when confronting health disparities,[27,28] as do social scientists when reasoning about inequality in poverty, crime, and education.

Practitioners facing the need of making causal statements about race often turn to a particular conceptual trick. The idea is to change object of study from the *effect of race* to the effect of *perceptions of race*.[29] What this boils down to is that we change the units of the study from individuals with a race attribute to *decision makers*. The treatment becomes *exposure to race* through some observable trait, like the name on a CV in a job application setting. The target of the study is then how decision makers respond to such *racial stimuli* in the decision-making process. The hope behind this maneuver is that exposure to race, unlike race itself, may be something that we can control, manipulate, and experiment with.

While this approach superficially avoids the difficulty of conceptualizing manipulation of immutable characteristics, it shifts the burden elsewhere. We now have to sort out all the different ways in which we think that race could possibly be perceived: through names, speech, style, and all sorts of other characteristics and combinations thereof. But not only that. To make a counterfactual statements viz-a-viz *exposure to race*, we would have to be able to create the authentic background conditions under which all these perceptible characteristics would've come out in a manner that's consistent with a different racial category. There is no way to construct such counterfactuals accurately without a clear understanding of what we mean by the category of race.[30] Just as we cannot talk about witchcraft in a valid causal model for lack of any scientific basis, we also cannot talk about perceptions of witchcraft in a valid causal model for the very same reason. Similarly, if we lack the ontological and epistemic basis for talking about race in a valid causal model, there is no easy remedy to be found in moving to perceptions of race.

In opposition to Holland's view, other scholars, including Pearl, argue that causation does not require manipulability but rather an understanding of *interactions*. We can reason about hypothetical Volcano eruptions without being able to manipulate Volcanoes. We can explain the mechanism that causes tides without being able to manipulate the moon by any feasible intervention. What is required is an understanding of the ways in which a variable interacts with other variables in the model. Structural equations in a causal model are *response functions*. We can think of a node in a causal graph as receiving messages from its parent nodes and responding to those messages. Causality is thus about who *listens* to whom. We can form a causal model once we know how the nodes in it interact.

But as we saw the conceptual shift to *interaction*—who *listens* to whom—by no means makes it straightforward to come up with valid causal models. If causal models organize available scientific or empirical information, there are inevitably limitations to what constructs we can include in a causal model without running danger of divorcing the model from reality. Especially in sociotechnical systems, scientific knowledge may not be available in terms of precise modular response functions.

We take the position that causes need not be experimentally manipulable.

However, our discussion motivates that constructs referenced in causal models need a certificate of ontological and epistemic stability. Manipulation can be interpreted as a somewhat heavy-handed approach to clarify the ontological nature of a node by specifying an explicit experimental mechanism for manipulating the node. This is one way, but not the only way, to clarify what it is that the node references.

Chapter notes

There are several introductory textbooks on the topic of causality. For a short introduction to causality turn to the primer by Pearl, Glymour, and Jewell,[14] or the more comprehensive textbook by Pearl.[6] At the technical level, Pearl's text emphasizes causal graphs and structural causal models. Our exposition of Simpson's paradox and the UC Berkeley was influenced by Pearl's discussion, updated for a new popular audience book.[8] All of these texts touch on the topic of discrimination. In these books, Pearl takes the position that discrimination corresponds to the direct effect of the sensitive category on a decision.

The technically-minded reader will enjoy complementing Pearl's book with the an open access text by Peters, Janzing, and Schölkopf[11] that is also available online. The text emphasizes two variable causal models and applications to machine learning. See Spirtes, Glymour and Scheines[31] for a general introduction based on causal graphs with an emphasis on *graph discovery*, i.e., inferring causal graphs from observational data.

Morgan and Winship[32] focus on applications in the social sciences. Imbens and Rubin[33] give a comprehensive overview of the technical repertoire of causal inference in the potential outcomes model. Angrist and Pischke[34] focus on causal inference and potential outcomes in econometrics.

Hernan and Robins[35] give another detailed introduction to causal inference that draws on the authors' experience in epidemiology.

Pearl[6] already considered the example of gender discrimination in UC Berkeley graduate admissions that we discussed at length. In his discussion, he implicitly advocates for a view of discussing discrimination based on the causal graphs by inspecting which paths in the graph go from the sensitive variable to the decision point. The UC Berkeley example has been discussed in various other writings, such as Pearl's discussion in the Book of Why.[8] However, the development in this chapter differs significantly in its arguments and conclusions.

For clarifications regarding the popular interpretation of Simpson's original article,[36] see Hernan's article[37] and Pearl's text.[6]

The topic of causal reasoning and discrimination gained significant momentum in the computer science and statistics community around 2017. Zhang, Wu, and Wu[38] previously considered discrimination analysis via path-specific causal effects. Kusner, Loftus, Russell, and Silva[13] introduced a notion of *counterfactual fairness*. The authors extend this line of thought in another work.[39] Chiappa introduces a path-specific notion of counterfactual fairness.[40] Kilbertus et al.[41] distinguish

between two graphical causal criteria, called *unresolved discrimination* and *proxy discrimination*. Both notions correspond to either allowing or disallowing paths in causal models. Razieh and Shpitser[42] conceptualize discrimination as the influence of the sensitive attribute on the outcome along certain *disallowed* causal paths. Chiappa and Isaac[43] give a tutorial on causality and fairness with an emphasis on the COMPAS debate. Kasirzadeh and Smart extend on the discussion about the difficulties with constructing causal counterfactual claims about social categories in the context of machine learning problems.[44]

There is also extensive relevant scholarship in other disciplines that we cannot fully survey here. Of relevance is the vast literature in epidemiology on health disparities. In particular, epidemiologists have grappled with race and gender in causal models. See, for example, the article by VanderWeele and Robinson,[28] as well as Krieger's comment on the article,[45] and Krieger's article on discrimination and health inequalities[46] for a starting point.

We retrieved the data about UC Berkeley admissions from `http://www.randomservices.org/random/d` on Dec 27, 2018. There is some discrepancy with the data displayed on the Wikipedia page for Simpson's paradox, which does not affect our discussion.

Bibliography

[1] Bickel, Peter J, Hammel, Eugene A, O'Connell, J William, *et al.*. 1975. "Sex bias in graduate admissions: Data from berkeley". *Science*, 187(4175):398–404.

[2] Humphrey, Linda L., Chan, Benjamin K.S., and Sox, Harold C.. 2002. "Postmenopausal Hormone Replacement Therapy and the Primary Prevention of Cardiovascular Disease". *Annals of Internal Medicine*, 137(4):273–284.

[3] Berkson, Joseph. 2014. "Limitations of the application of fourfold table analysis to hospital data". *International Journal of Epidemiology*, 43(2):511–515. Reprint.

[4] Moneta-Koehler, Liane, Brown, Abigail M., Petrie, Kimberly A., Evans, Brent J., and Chalkley, Roger. 2017. "The limitations of the gre in predicting success in biomedical graduate school". *PLOS ONE*, 12(1):1–17.

[5] Hall, Joshua D., O'Connell, Anna B., and Cook, Jeanette G.. 2017. "Predictors of student productivity in biomedical graduate school applications". *PLOS ONE*, 12(1):1–14.

[6] Pearl, Judea. 2009. *Causality*. Cambridge University Press.

[7] Deaton, Angus and Cartwright, Nancy. 2018. "Understanding and misunderstanding randomized controlled trials". *Social Science & Medicine*, 210:2–21.

[8] Pearl, Judea and Mackenzie, Dana. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

[9] Glymour, M Maria. 2006. "Using causal diagrams to understand common problems in social epidemiology". *Methods in social epidemiology*, pages 393–428.

[10] Krieger, Nancy. 2011. "Epidemiology and the people's health: Theory and context".

[11] Peters, Jonas, Janzing, Dominik, and Schölkopf, Bernhard. 2017. *Elements of Causal Inference*. MIT Press.

[12] Baron, Reuben M and Kenny, David A. 1986. "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations." *Journal of Personality and Social Psychology*, 51(6):1173.

[13] Kusner, Matt J., Loftus, Joshua R., Russell, Chris, and Silva, Ricardo. 2017. "Counterfactual fairness". In *Advances in Neural Information Processing Systems*, pages 4069–4079.

[14] Pearl, Judea, Glymour, Madelyn, and Jewell, Nicholas P.. 2016. *Causal Inference in Statistics: A Primer*. Wiley.

[15] Zhang, Junzhe and Bareinboim, Elias. 2018. "Fairness in decision-making — the causal explanation formula". In *Proc. 32nd AAAI*.

[16] Egan, Patrick J. 2020. "Identity as dependent variable: How americans shift their identities to align with their politics". *American Journal of Political Science*, 64(3):699–716.

[17] Glasgow, Joshua, Haslanger, Sally, Jeffers, Chike, and Spencer, Quayshawn. 2019. "What is race?: Four philosophical views".

[18] Bowker, Geoffrey C and Star, Susan Leigh. 2000. *Sorting things out: Classification and its consequences*. MIT Press.

[19] Fields, Karen E. and Fields, Barbara J.. 2014. *Racecraft: The Soul of Inequality in American Life*. Verso.

[20] Benjamin, Ruha. 2019. *Race after Technology*. Polity.

[21] Hacking, Ian. 2000. *The Social Construction of What?* Harvard University Press.

[22] Haslanger, Sally. 2012. *Resisting Reality: Social Construction and Social Critique*. Oxford University Press.

[23] Mallon, Ron. 2018. *The Construction of Human Kinds*. Oxford University Press.

[24] Cartwright, Nancy. 2006. *Hunting Causes and Using Them, Too*. Cambridge University Press.

[25] Hacking, Ian. 2006. "Making up people". *London Review of Books*, 28(16).

[26] Holland, Paul W.. 1986. "Statistics and causal inference". *Journal of the American Statistical Association (JASA)*, 81:945–970.

[27] Jackson, John W. and VanderWeele, Tyler J.. 2018. "Decomposition analysis to identify intervention targets for reducing disparities". *Epidemiology*, pages 825–835.

[28] VanderWeele, Tyler J. and Robinson, Whitney R.. 2014. "On causal interpretation of race in regressions adjusting for confounding and mediating variables". *Epidemiology*.

[29] Greiner, D. James and Rubin, Donald B.. 2011. "Causal effects of perceived immutable characteristics". *The Review of Economics and Statistics*, 93(3):775–785.

[30] Kohler-Hausmann, Issa. 2019. "Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination". *SSRN*.

[31] Spirtes, Peter, Glymour, Clark N, Scheines, Richard, Heckerman, David, Meek, Christopher, Cooper, Gregory, and Richardson, Thomas. 2000. *Causation, prediction, and search*. MIT Press.

[32] Morgan, Stephen L. and Winship, Christopher. 2014. *Counterfactuals and Causal Inference*. Cambridge University Press.

[33] Imbens, Guido W. and Rubin, Donald B.. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

[34] Angrist, Joshua D. and Jörn-Steffen, Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

[35] Hernán, Miguel and Robins, James. 2019. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.

[36] Simpson, Edward H. 1951. "The interpretation of interaction in contingency tables". *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241.

[37] Hernán, Miguel A, Clayton, David, and Keiding, Niels. 2011. "The Simpson's paradox unraveled". *International Journal of Epidemiology*, 40(3):780–785.

[38] Zhang, Lu, Wu, Yongkai, and Wu, Xintao. 2017. "A causal framework for discovering and removing direct and indirect discrimination". In *Proc.* 26*th IJCAI*, pages 3929–3935.

[39] Russell, Chris, Kusner, Matt J., Loftus, Joshua R., and Silva, Ricardo. 2017. "When worlds collide: Integrating different counterfactual assumptions in fairness". In *Advances in Neural Information Processing Systems*, pages 6417–6426.

[40] Chiappa, Silvia. 2019. "Path-specific counterfactual fairness". In *Proc.* 33*rd AAAI*, volume 33, pages 7801–7808.

[41] Kilbertus, Niki, Rojas-Carulla, Mateo, Parascandolo, Giambattista, Hardt, Moritz, Janzing, Dominik, and Schölkopf, Bernhard. 2017. "Avoiding discrimination through causal reasoning". In *Advances in Neural Information Processing Systems*, pages 656–666.

[42] Nabi, Razieh and Shpitser, Ilya. 2018. "Fair inference on outcomes". In *Proc.* 32*nd AAAI*, pages 1931–1940.

[43] Chiappa, Silvia and Isaac, William S.. 2019. "A causal bayesian networks viewpoint on fairness". *arxiv.org*, arXiv:1907.06430.

[44] Kasirzadeh, Atoosa and Smart, Andrew. 2021. "The use and misuse of counterfactuals in ethical machine learning". In *Conference on Fairness, Accountability, and Transparency*, pages 228–236.

[45] Krieger, Nancy. 2014. "On the causal interpretation of race". *Epidemiology*, 25(6):937.

[46] —. 2014. "Discrimination and health inequities". *International Journal of Health Services*, 44(4):643–710.