Path-Specific Discrimination

Naftali Weinberger

November 1, 2022

Abstract

The distinction between direct and indirect effects plays an increasingly prominent role in studies of discrimination, algorithmic fairness, and measuring disparities. This paper clarifies the interpretation of such effects within the context of non-parametric mediation methods. These effects should not be understood as decomposing the total effect into the independent contributions of direct and indirect paths, but instead concern the result of allowing a change in the causal variable to be transmitted via some causal paths and not others. Given this, distinct changes in the sensitive attribute (e.g. from black to white vs. from white to black) correspond to distinct effects, and talk of a path's "contribution" is therefore ambiguous. I consider some implications of this change-relativity for discussions of discrimination and fairness and further highlight a neglected substantive question regarding whether discrimination is linked to paths or to path-specific effects.

1 Introduction

Among social scientists studying racial discrimination by police, there has recently been a heated debate surrounding a study that found only limited discrimination (Fryer Jr, 2019). This debate has led to increased attention the methodology of testing discrimination claims using both experimental and non-experimental data. One crucial issue is that in contexts of discrimination, which cases end up in the data are typically not random with respect to race. Knox et al. (2020) emphasize that in studies on police use of force, the records contain only individuals who are stopped by the police in the first place. But plausibly, whether one is stopped is not random with respect to race, since one's race may influence whether one is stopped. We thus have a situation in which race potentially influences the use of force via two distinct pathways: it can influence both whether one is stopped (and thus potentially subject to force) as well as one's chances of being subject to force if stopped (fig. 1). One must account for both of these influences in order to draw reliable conclusions from police records.

The causal model within which the influence of race is mediated by whether an individual is stopped is called a *mediation model* and facts about how race influences being stopped via particular causal avenues are known as *path-specific effects*. An example of a path-specific effect would be the direct effect of race on force that is *not* mediated via being stopped. I'll make this more precise, but as a first approximation, this corresponds to the influence of race on being subject to force among people who are in fact stopped. In the Knox et al. case, the fact that race influences use of force via two paths has implications for how to properly interpret the data. This reveals that causal reasoning is relevant for interpreting discrimination statistics even in contexts in which one is not explicitly considering a possible intervention.

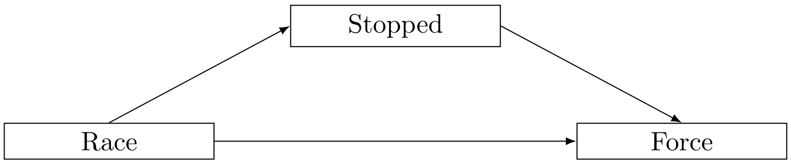

Figure 1: Knox et al. (2020) Example

The present paper explores the relevance of mediation models for studying discrimination. Cases that are appropriately modeled using mediation methods are ubiquitous in the discrimination literature. For instance, in Bickel et al.'s (1975) widely cited study of possible gender discrimination in Berkeley admissions, discrimination is properly analyzed as a direct effect of gender on admission, conditional on the department to which one applied. Moreover, it is not a coincidence that in Pearl's (2001) groundbreaking paper on mediation, a hypothetical case of employment discrimination provides a key illustration of a direct effect. The link between path-specific effects and discrimination, which may initially appear to be a mere curiosity, is now playing a more systematic role in two burgeoning research fields. First, path-specific effects have been used in formal analyses of algorithmic fairness (Zhang and Bareinboim, 2018; Chiappa, 2019; Plecko and Bareinboim, 2022). Second, they have been central to the study of racial health disparities (Graetz et al., 2022; Jackson, 2018).

This paper will present a systematic discussion of the relevance of mediation models for analyzing discrimination. Despite prior work on the topic (see e.g. VanderWeele and Robinson, 2014), there are several reasons why such a discussion is needed. The first is partly pedagogical. Scientists employing mediation methods are still very much influenced by the mediation methods from Baron and Kenny (1986) – as opposed to Pearl's (2001) non-parametric methods. I will argue that the former methods, although suitable for their intended domain, produce the wrong intuitions for understanding path-specific effects more generally. This has already been discussed by Weinberger (2019), but since mediation remains obscure to philosophers and confusing to scientists, a more explicit attempt at addressing misunderstandings is warranted. Second, and more constructively, a proper understanding of path-specific effects opens the door to novel questions that, I will argue, the existing literature has not even begun to address. Mediation methods enable one not merely to differentiate between direct and indirect effects, but further reveal that there are various versions of each. There are two versions of the direct effect – natural and controlled – and for each of the direct and indirect effect there are at least two possible effects to measure depending on the particular change in the causal variable. Although different studies employ different effects, often opting for the controlled direct effect, there has been little systematic defense of choosing one effect over another, or even discussion of whether and why path-specific effects are of interest in the first place. A precise philosophical analysis of *what* path-specific effects are will clarify the advantages of choosing particular effects over others and will reveal why this choice is sensitive to normative and legal issues surrounding discrimination.

Perhaps the most surprising conclusion of the paper is that if one treats racial discrimination as corresponding to a natural direct effect in a particular model, anti-black and pro-white discrimination correspond to logically independent counterfactuals (this point also generalizes to other races and other categories). That is, the truth of the causal counterfactual corresponding to whether (e.g.) a particular black job candidate was discriminated against has no implications regarding the truth of the corresponding counterfactual about a white candidate was privileged based on her race. I provide evidence that disentangling these two counterfactuals matters even for relatively informal debates over discrimination and has a further analogue to a distinction between two types of errors in the algorithmic fairness literature. But this is just one insight that arises from careful attention to mediation techniques. The paper also reveals that there is a difference between claiming that discrimination corresponds to the existence of a direct path and claiming that it corresponds to a direct effect, and that the choice of the latter comes with a substantive normative commitment regarding the relationship between discrimination and harm. Finally, the paper emphasizes that mediation techniques should not be understood as providing an additive decomposition of the total effect into the independent contributions of the direct and indirect path. While this last point ought to be clear to those who are well-versed in non-parametric mediation techniques, it is easy to overlook and should be kept in mind by those using mediation for the decomposition of disparities.

2 Some Motivating Cases

In this section I present some preliminary scenarios that can be represented with mediation models. Doing so will help the reader develop some intuitions regarding the types of cases to which mediation techniques may be fruitfully applied and will furnish examples that will guide the subsequent discussion.

Let's begin with a hypothetical example. Imagine that a job applicant is denied a job and the question arises as to whether she was racially discriminated against. One might think that, at least from a causal perspective, the relevant question is simple: did the person's race cause her not to

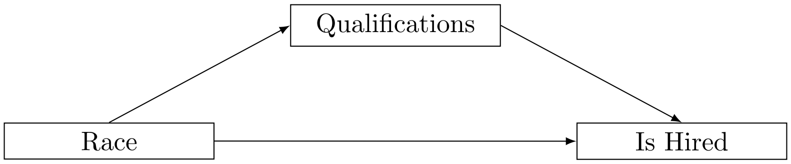

Figure 2: Employment Discrimination DAG

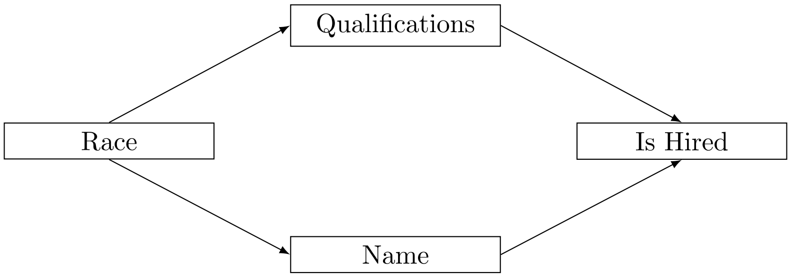

Figure 3: Audit Study

be hired? But it turns out not to be so straightforward. It could be that, as a result of prior racial discrimination, the person was denied opportunities to develop job-relevant qualifications and that this accounted for her not getting the job. If this were the whole causal story, then we presumably would not say that this particular employer discriminated against the applicant, even though race influenced her not getting the job. Consequently, what we want to know is whether race influenced hiring, even when holding the applicant's qualifications fixed.

We may model this using three variables: *race*, *job-relevant qualifications*, and whether the applicant is *hired* by the employer. For simplicity, we will imagine that the only two possible values of the race variable are black and white, and that the candidate's job-relevant qualification might be high, medium, or low. We will presume that the position in question is one in which race would not count as a job-relevant qualification (e.g. it is not an audition for a role in *Hamilton*). Pearl (2001), among others (Zhang and Bareinboim, 2018), has used this type of example as a basis for illustrating the relevance of path-specific effects. The idea is that what matters for discrimination here is not whether there is *any* effect of race in being hired, but rather whether race *directly* influences being hired via an avenue not going through qualifications. Using mediation terminology, this amounts to saying that what matters is not whether there is a *total effect* but rather a *direct effect* (see section 3).

This hypothetical example provides a basis for understanding how racial discrimination is tested using *audit studies*, in which discrimination is tested by varying a racially-informative cue. For example, suppose that the application process only involves submitting a resume and that the experimenter varies the racially relevant information by manipulating the name on the resume (Bertrand and Mullainathan, 2004). The causal model for this is given in figure 3. Here manipulating the name from a white-sounding name to a black sounding name while holding qualifications fixed corresponds to an *indirect effect* of race on being hired via name. The indirect effect in this model corresponds to the direct effect in the previous one. Both direct and indirect effects are instances of *path-specific effects* and the possibility of representing the same causal quantity across different models is an underappreciated feature of mediation models (see appendix).

This preliminary example will allow me to flag some questions that I will be putting to the side in this paper. There is an ongoing discussion about how to causally analyze demographic variables such as race, given that one arguably cannot experimentally manipulate an individual's race. A popular proposal is to argue that what is manipulated is not race, but rather the discriminator's perception of race (Greiner and Rubin, 2011). Kohler-Hausmann (2018) views this response as inadequate and, more generally, sees the use of causal models for discrimination as misguided. In the given example, the employer's perception of race may not be fully distinct from their perception of qualifications, as they may interpret purportedly similar qualifications differently based on race. To the extent that

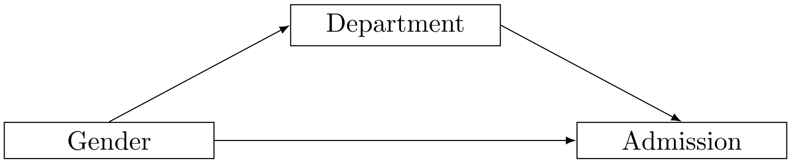

Figure 4: Berkeley Example (Bickel et al., 1975)

doing so is illegitimate, Kohler-Hausmann would argue, this is to be established using normative arguments regarding what counts discrimination, rather than via a counterfactual test regarding what happens when one varies race and qualifications independently. Weinberger (2022) argues that without a counterfactual test it will be impossible to empirically test claims about discrimination, though his argument presupposes rather than establishes that there is a coherent way to distinguish race from qualifications. The issue thus calls for further discussion, though here I will focus differentiating the various effects defined by mediation models without addressing the thorny questions of whether and how such effects are sensitive to the way one defines the race variable.

A second scenario I will consider is the one from Knox et al. (2020) that was mentioned in the introduction. There race influences being stopped in two ways: by influencing whether one is likely to be stopped and by further influencing whether someone who is stopped is likely to be subject to force (fig. 1). Note that here the mediator (i.e. the mediating variable corresponding to whether one is stopped) *interacts* with race, in the sense that the effect of one variable depends on the value of the other. In the case where one is *not* stopped, force will not be employed, regardless of race. We see that the representation of two distinct causal arrows does not imply that the two causes make additive contributions to the effect, but only that manipulating each cause has some influence on the effect for some fixed value of the other.

In this model, one is likely to be interested either in the total effect or the direct effect. Whereas the total effect captures the effect of race along all paths, the direct effect is the effect of race on being subject to force given that one is stopped (or not). Here the indirect effect corresponds to the influence that race would have on force via being stopped, were there to be no influence via the direct path. Different effects will be of interest in different contexts. A city choosing whether to mandate de-escalation training for police will be primarily interested in the direct effect. In contrast, a sociologist studying how racial disparities are influenced by disparate police behavior may care more about the total effect.

As a final illustrative example, consider the widely-cited study of graduate admissions at Berkeley (Bickel et al., 1975). The curious feature of that case was that while in the university as a whole female applicants were less likely than males to be accepted to a graduate program, *within every single department* female students were at least as likely to be admitted. This has become a paradigm example of Simpson's paradox (Sprenger and Weinberger, 2021) and the standard explanation of the overall negative correlation between gender and admission is that women were applying to departments with lower acceptance rates. Accordingly, even though women were not disadvantaged within any department, learning that an applicant was a woman provided evidence that they were applying to a more competitive department and thus had a lower chance of admission (compared to applicants to other departments) regardless of their gender.

A simplified model for this case is given in figure 4. There gender influences admission indirectly via department choice, and to the extent that gender has further influence within departments, this corresponds to various direct effects (corresponding to different departments). Although mediation is rarely emphasized in the context of this example,[1] it is essential for understanding its relevance for Simpson's paradox. The standard causal analysis of the paradox is that it results from conflating causal and evidential reasoning (i.e. whether gender causes or is evidence for admissions) and that once one eliminates confounding, the paradox should disappear (see Pearl, 2014). The issue is that in the present case the *department* variable is a mediator, and one need not (and should not) condition on mediators to de-confound the total effect of gender on admissions. To properly understand the case,

[1] Zhang and Bareinboim (2018); Barocas et al. (2019) are noteworthy exceptions

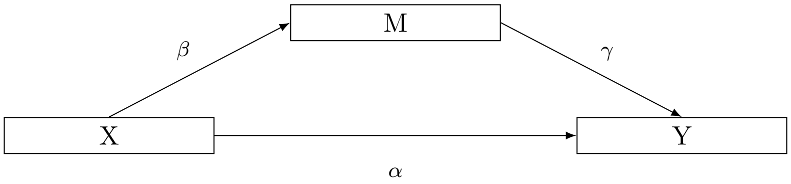

Figure 5: Linear Additive Model

it is crucial that one treats the discrimination hypothesis as one concerning whether there is a *direct* effect of gender on admissions, since identifying such an effect *does* require conditioning or intervening upon the mediator (in a particular way).

3 Defining Path-Specific Effects

For ease of illustration, it helps to focus on cases with three variables and two causal paths, such as those in figures 1, 2, and 4. In such cases, we may generically refer to the causal variable (e.g. race), X, as the *treatment*, the effect variable, Y, as the *outcome* and the intermediate variable, M, as the *mediator*. A *causal path* is a set of connected causal arrows all going in the same direction (from cause to effect). In the cases considered, there are two causal paths: the indirect path via the mediator, and the direct path not via the mediator. Talk of direct/indirect *paths* is not equivalent to talk of direct and indirect *effects*. While paths are qualitative features of causal graphs, effects are quantities (with magnitudes). The existence of a direct path entails that there is some way to change the value (or probability) of the outcome by intervening to vary the value of the treatment while holding the mediator fixed at a particular value. As long as there is *some* value of the mediator for which the treatment makes a difference for the outcome, there is a direct path. Direct *effects*, in contrast, concern the amount by which the the outcome (or its probability) will change given specific interventions on the mediator (where different values of the mediator correspond to distinct direct effects).

The most influential approach to mediation is that of Baron and Kenny (1986), which applies to linear models with no interaction. The path-specific effects in such models are relatively straightforward to define, but this simplicity comes at the cost of producing the wrong intuitions about what such effects are. Consider a linear additive model with treatment X, mediator M and outcome Y, in which the linear parameter for $X \to Y$ is α the parameter for $X \to M$ is β and the parameter for $M \to Y$ is γ (see fig. 5). Here the contribution of the direct path is given by α, that of the indirect path by $\beta\gamma$, where the former corresponds to the influence that X would have on Y were there to be no indirect influence. Accordingly, Baron and Kenny (1986) define the direct effect as α and the indirect effect as $\beta\gamma$. The resulting picture is one in which the total effect is the sum of the direct and indirect effects, which correspond to the independent contributions of the two paths. This picture does not generalize, since once one allows interaction between the mediator and treatment in their effect on the outcome, one can no longer talk about *the* effect along the direct path without specifying a particular value (or distribution) of the mediator, and thus what happens along one path depends on (and cannot be isolated from) what happens along the other one. Concretely, if one takes the given linear model and adds a single interaction term $\delta\, XM$ in the structural equation determining the value of Y, then when $M = 0$ the effect of a unit change of X on Y is α and when $M = 1$ it is $\alpha+\delta$. If one thinks of mediation analysis as a matter of decomposing the total effect into two independent contributions corresponding to the direct and indirect paths, no such decomposition will be available in cases of interaction.

A preliminary response to interaction is to acknowledge that the degree of influence of the treatment on the outcome along the direct path always depends on the value of the mediator, and thus that there will be as many direct effects as there are values of the mediator. For each value of the mediator, one can define a *controlled direct effect* (CDE), where $CDE_{x,x'}(M = m)$ corresponds to the change in the outcome resulting from changing the treatment from x to x' while intervening to set the mediator to $M = m$. Using potential outcomes notation, in which $Y_{X=x,M=m}$ denotes the value of Y resulting from intervening to set X to x and M to m, the CDE of changing X from $X = 0$ to $X = 1$ while setting M to m is as follows:

$$(1) CDE_{0,1}(M = m) = Y_{X=1,M=m} - Y_{X=0,M=m}$$

The question then arises as to which CDE (i.e. which setting of M) is relevant in a context. Consider again the scenario from Knox et al. (2020) (fig. 1). Suppose that a particular non-suspicious individual is such that she never would be stopped in this scenario, regardless of her race. Here the controlled direct effect in which $M =$ *stopped* corresponds to whether this individual's race would have led to the use of force under an alternate scenario in which police were *forced* to stop this individual (not taking into account race or other factors they would ordinarily use to make the decision). But why should *this* alternate scenario be relevant to the never-stop individual? Given that, under the existing policies, this individual would never be stopped, information about this alternate scenario does not necessarily say anything about what happens along the direct path in the scenario that actually obtains.

We should give up on trying to understand path-specific effects as additive decompositions of the total effect.[2] A more fruitful way understand such effects is as indicating the way that particular *changes* in the treatment would influence the outcome, were that change allowed to be transmitted only via certain paths (Weinberger, 2019). Let's break this down slowly. First, note that path-specific effects are always relative to a *change* in the value of the treatment. For treatment with two values, $X = 0$ and $X = 1$, the path-specific effects of changing the treatment from 0 to 1 are different from those of changing it from 1 to 0, for reasons that will become clear. The notion of "transmission" builds on the fact that within a causal model, the value or distribution of the variables in a model is determined by the values of their causes (and an error term). In potential outcomes notation, $M_{X=0}$ indicates the value that the mediator would take on were X set to 0 via intervention, and $Y_{X=0,M_{X=0}}$ indicates the value that the outcome would take on if X were set to 0 and the mediator has the value it does when $X = 0$. For a change to be transmitted is simply for its effect variables to vary as they would in response to the change. So if X is changed from 0 to 1, the mediator will change from $M_{X=0}$ to $M_{X=1}$. For the change in X from 0 to 1 *not* to be transmitted to a particular mediator is for that mediator to maintain the value $M_{X=0}$.

With this background, we now turn to the *natural direct effect.* The natural direct effect of changing X from 0 to 1 is the change in the (probability of the) outcome that would result from changing X from 0 to 1, while intervening to hold the mediator at $M_{X=0}$. Formally,

$$(2) NDE_{0,1} = Y_{X=1,M_{X=0}} - Y_{X=0,M_{X=0}}$$

In the Knox example, if $X = 0$ is black and $X = 1$ is white, then (1) gives the counterfactual: "How would being white (as opposed to black) change one's chances of being subject to force, where one to be stopped as one would were one to be black". In other words, one considers the effects of changing race on use of force, without allowing that change to influence whether one is stopped. To be clear, evaluating this effect is not a matter of intervening on the mediator so that it has *no* relationship to the treatment – to set M to $M_{X=0}$, one must know how the mediator depends on the treatment. Rather, one makes the mediator behave as if it were not responding to the *change* in the treatment. The fact that the mediator is set as a function of the treatment is crucial for understanding why it is relevant to the individual (or individuals) being considered. It ensures that one is considering whether *this* individual would have been stopped had they been of a different race. This is an important difference between the natural and controlled direct effects.

While the natural and controlled direct effects are conceptually different quantities, in some cases they may coincide. Suppose that when $X = 0$, $M_{X=0} =$ *stopped.* Then the natural direct effect given by (1) will correspond to the controlled direct effect for $M =$ *stopped.* Accordingly, a randomly chosen CDE might happen to correspond to a desired natural direct effect, but to determine whether this is so, one must consult the formula for the NDE to see whether the mediator is set to the appropriate value. The difference between natural and controlled direct effects becomes even sharper when one considers heterogeneous populations instead of individuals (or homogeneous populations). For populations, identifying $NDE_{0,1}$ is a matter of setting the mediator to $M_{X=0}$ for each individual in the population – where this value may differ across individuals. In contrast, the CDE sets the mediator to the same value for *every* member of the population.

[2] Readers wondering how this claim is compatible with existing mathematical results decomposing total effects into terms containing direct and indirect effects are directed to section 6 below.

The reason that $NDE_{0,1}$ and $NDE_{1,0}$ are two different quantities is that while the former requires setting the mediator to $M_{X=0}$ the latter sets it to $M_{X=1}$, and as a result of interaction, this can change the magnitude of the effect of X on Y. Recall that the direct effect does not indicate what would happen if the causal effect along the direct path were wholly independent of what happens on the other path, but rather what would happen were a particular change not to be transmitted along the indirect path. But depending on which change is being considered – whether from 0 to 1 or the reverse – the value of the mediator in the no-change scenario will be different.

In talking about the counterfactuals related to path-specific effects, it is helpful to employ the notion of a "default" value of the treatment. For instance, we might stipulate that the default value of the treatment is $X = 0$ and, correspondingly, the default values of the mediator and outcome are $M_{X=0}$ and $Y_{X=0,M_{X=0}}$, respectively. When considering changes from $X = 0$ to $X = 1$, we can refer to $X = 1$ as the non-default value of the treatment, and define the non-default mediator and outcome accordingly. This way of talking allows us to more generally define the natural direct effect as the change in the outcome that would result from changing the treatment from its default to its non-default value, while holding the mediator fixed at its default value. To generate the equation for the natural direct effect for the reverse change (from 1 to 0), one simply needs to treat X=1 as the default value instead, yielding the following:

$$(3) NDE_{1,0} = Y_{X=0,M_{X=1}} - Y_{X=1,M_{X=1}}$$

The specification of the default is entirely arbitrary and extrinsic to the model. Nevertheless, given that path-specific effects are change relative, talk of the "non-change" value of the treatment at a default value can be useful when describing the complex counterfactuals defining these effects.

The *indirect effect* corresponds to the result of changing the mediator as it would change, were the treatment to be varied from its default to its non-default value, while maintaining the treatment at its default value. For instance, given a change from $X = 0$ to $X = 1$, the indirect effect would be:

$$(4) IE_{0,1} = Y_{X=0,M_{X=1}} - Y_{X=0,M_{X=0}}$$

Because one varies the mediator from its default to its non-default value, the indirect path behaves as it would were it responding to a hypothetical change in the treatment. But because the treatment is held fixed at its default value, no change is transmitted along the direct path. There is no "controlled" version of the indirect effect. That it is possible to identify the indirect effect without intervening on mediators along all the other paths is one of the major advances of Pearl's (2001) non-parametric causal mediation techniques.[3] Given a proper understanding of the natural direct effect – and path-specific effects more generally – the concept behind the indirect effect follows, well, naturally.

4 Discrimination and Privilege

In the previous section I emphasized that for a treatment variable with the values 0 and 1 there are two distinct natural direct effects depending on whether the change is from 0 to 1 or vice versa. This might initially seem like a technical oddity with no further significance. In this section I will explain why this distinction matters in contexts of discrimination. Returning to the employment discrimination example depicted in figure 2, I will explain why the natural direct effect of being black (as opposed to white) is logically distinct from the natural direct effect of being white (as opposed to black) and provide evidence that the difference between these counterfactuals plays a role even in non-technical discussions of discrimination.

When the applicant is black, it will be useful to treat *Race=black* as the default value of the treatment. The NDE is the difference in the chance of the applicant's being hired that would result from the applicant being white (instead of black) were they to have the qualifications they do have as a result of being black. In other words, one holds their job-relevant qualifications fixed at their actual value. Why is it important here to consider the *natural* direct effect? Suppose that the candidate does not get hired due to lacking the relevant qualifications. Even if one were then to learn that black

[3] Though see Robins and Greenland (1992) for some earlier results in a similar direction. The ability to define the direct effect "intrinsically" (i.e. without intervening on mediators along other paths) bears an intriguing resemblance to David Lewis' (1974) definition of causation as the ancestral of dependence – meaning that a chain of events $X_1 \to X_2 \to X_n$ such that each link in the chain counterfactually depends on its predecessor suffices for causation.

candidates with high qualifications are less likely to get the position than similar white candidates – and thus that there is a controlled direct effect for *Qualifications=high* – this would not show that *this* individual was discriminated against. More generally, since the individual would still not have gotten the position had they been of a different race, they do not have a claim to having been discriminated against by this employer.[4]

Now let's consider the natural direct effect of being white (as opposed to black). To make this effect vivid, it helps to imagine a white candidate who is highly qualified and who gets the position based on those qualifications. Nevertheless, had the same candidate been black with the same qualifications, they would have been less likely to get the position. This corresponds to there being a positive NDE for the default of *Race=white*. The existence of such an NDE does not entail that any individual was discriminated against. In other words, that this individual's race made some difference in their getting the job does not mean that there was some other individual who would have gotten the position except for their race. Nevertheless, the presence of a natural direct effect does indicate an illegitimate role of race in the decision making process. Even if no one was necessarily discriminated against, we can say that the individual who got the position was *privileged* based on their race.

To distinguish between $NDE_{black,white}$ versus $NDE_{white,black}$, we will refer to the first as the *discrimination* counterfactual, and the second as the *privilege* counterfactual. The possibility of distinguishing between these two independent quantities constitutes a further advantage of considering natural, as opposed to controlled, direct effects.

The distinction between discrimination and privilege matters not only in quantitative contexts, but also in relatively informal discussions of discrimination. When the author Ta-Nehisi Coates was interviewed on *The Daily Show with Trevor Noah* after the US presidential election in 2016, he was asked about the assertion that Donald Trump's election could not be attributed to racism, since some Trump voters had previously voted for Obama. His response:

> If I have to jump six feet to get the same thing that you have to jump two feet for – that's how racism works. To be president, Obama had to be scholarly, intelligent, president of the Harvard Law Review, the product of some of our greatest educational institutions, capable of talking to two different worlds. Donald Trump had to be rich and white. That's the difference.

In other words, those denying the role of race were pointing to the following counterfactual:

Discrimination: Given Obama's actual qualifications, his race did not make a difference in whether certain individuals voted for him.

In contrast, Coates is pointing to the distinct counterfactual:

Privilege: Given Trump's actual qualifications, had he been black these individuals would not have voted for him.

The distinction between these two counterfactuals can be spelled out in an identical manner as those in the job hiring case. Simply replace the variable *is hired* in the model with *is voted for* in figure 2 and the two counterfactuals correspond to the two natural direct effects from the hiring example. Of course, mediation models do not by themselves specify which path-specific effect is relevant in a particular context. Nevertheless, we see that the framework provides tools for distinguishing among counterfactuals that would be difficult to disentangle without a formal framework. Failure to do so can result in the participants in a debate not even realizing that they are appealing to different counterfactuals, and that evidence for one is not evidence against the other.

4.1 Application to Algorithmic Fairness

The distinction between discrimination and privilege further ties in to discussions of algorithmic fairness. Barocas' et al.'s (2019) introduction to statistical fairness criteria contains a distinction that

[4] To say that the employer's action here is legitimate is *not* to claim that it is unproblematic that the candidate was denied the position as a result of lacking qualifications they would have had had they been of a different race. This does point to the existence of unjust inequalities in the broader society. The claim here is simply that this particular employer did not illegitimately discriminate.

relates to the causal counterfactual one just discussed. One of the three fairness criteria they consider is called *separation*, which requires that the sensitive attribute (race) be probabilistically independent of the decision (is hired) conditional on what one is seeking to measure (qualifications). With our variables, this corresponds to X being independent of Y conditional on M. They note that the separation criterion entails that both the *true positive rate* (the probability of being hired conditional on qualified) and the *false negative rate* (the probability that one is hired conditional on *not* being qualified) is the same across both groups, and mention the possibility of requiring only one of these rates to be equal (Barocas et al., 2019, p. 48). I will now explain how the criteria of equality of true and of false positives can be aligned with discrimination criteria corresponding to discrimination and privilege, respectively.

A small caveat: The variables in the employment example do not exactly correspond to those of interest in Barocas et al.'s discussion, since denying a qualified candidate a job is not necessarily a matter of false classification. To make the cases analogous, our causal model would have to consider the effect of qualifications on the employer's judgment of whether the person is qualified. Putting this issue to the side, let's consider the relationship between Barocas et al.'s probabilistic criterion and causal counterfactuals. For the sake of conceptual clarity, let's assume that there is no confounding (e.g. omitted common causes) and thus that one can treat the conditional probabilities as indicating how the probabilistic consequents would respond to hypothetical interventions on the antecedents. Moreover, let's assume that the data-generating model is the mediation model we've been considering. Equality of true positives then corresponds to $Y_{X=black,M=qualified} = Y_{X=white,M=qualified}$, ensuring that no one who was qualified was discriminated against based on their race. Equality of false positives corresponds to $Y_{X=black,M=unqualified} = Y_{X=white,M=unqualified}$, ensuring that no one who was unqualified was privileged based on their race.

The cases in which privilege corresponds to false negatives are a subset of the full set of cases I'm inclined to consider. Above I claimed that even a qualified white candidate might be privileged if they would have been less likely to get it had they been black. Unsurprisingly, pure probabilistic statements will be unable to express the full range of counterfactual possibilities. Nevertheless, the links drawn here illustrate the relevance of the discrimination/privilege distinction for discussions of algorithmic fairness. Note that equality of true positives is equivalent to equality of false negatives. Our discussion thus reveals that the counterfactuals corresponding to false negatives and false positives are independent. In statistics, it is elementary that the aims of avoiding false negatives and avoiding false positives are distinct. It is therefore unsurprising that a similar distinction arises in the causal counterfactual context.

5 Why go Natural?

The advantage of using natural, as opposed to controlled, direct effects is that doing so may be relevant to identifying the counterfactual of interest. In the Knox et al. (2020) example, the natural direct effect indicates how an individual's chance of being subject to force would change if she and the police were behaving as they were in the original scenario, but she was stopped (or not) as she would have been had she been of a different race. If, under this alternate scenario, she would *not* have been stopped, then the controlled direct effect of her being stopped is of limited relevance for understanding the influence of this person's race in the original scenario. At best it tells us about her chance of being subject to force under conditions in which police changed their policy about who to stop such that they would have stopped her.

The hiring scenario is in fact much more nuanced than I have thus far suggested. I've claimed that for an individual who was not hired to plausibly claim that they personally were denied a job due to racial discrimination, they need to be able to argue that had they been of a different race, but still had their qualifications, they would have been less likely to get the job. This corresponds to the natural direct effect, as opposed to some controlled direct effect for some arbitrary qualifications level. Nevertheless, there is a substantive normative and legal question regarding whether and when establishing discrimination requires establishing that particular individuals or groups were discriminated against. One might argue that in the employment example, showing that the employer took race into account would suffice for demonstrating that she acted in a discriminatory manner, whether or not anyone was harmed by the discrimination. If so, what would matter is only whether there is a direct path – and thus *some* controlled direct effect – not whether there is a natural direct effect.

Here I do not aim to weigh in on what the correct way to think about discrimination is here. I simply highlight that the choice between identifying discrimination with a direct effect and identifying it with a direct path corresponds to a normative question regarding whether (and in what contexts) establishing discrimination requires establishing that there were individuals who were harmed by the discriminatory action. To the extent that the answer is yes, one needs to appeal to a natural direct effect (or an indirect effect).

The difference between paths and path-specific effects is relevant to further clarifying the pioneering account offered by Zhang and Bareinboim (2018) and significantly extended in Plecko and Bareinboim (2022). Zhang and Bareinboim show how the variation in an outcome can be decomposed into a direct, indirect, and a spurious effect. This last effect corresponds to the variation due to a common cause and the others are defined along the lines of the definitions given here. Despite their use of path-specific effects, their explicit claims linking the models to claims about discrimination focus on the existence of *paths* in the models. Notably, their properties 1 and 2 (p. 5) link claims about whether there is direct or indirect discrimination to the presence (or absence) of direct or indirect paths in a model ("indirect discrimination" refers to discrimination via a proxy, such as neighborhood in the case of redlining). They note that the presence of such paths can be detected by establishing the presence of a non-zero direct or indirect effect. But if it were merely the existence of the causal path that were of interest, this would be a roundabout way to establish this. And once one grants that is not the existence of paths, but also the magnitude of the effects that matters, this warrants further discussion of which effects matter in which context. I present these considerations not as a criticism of Zhang and Bareinboim's general approach, but rather as an indication of some important questions that it has yet to address systematically.

6 Decomposition Analysis

In my review of mediation methods, I argued that outside of the linear additive case, one should not think about mediation as decomposing the total effect into the distinct contributions of the paths. Some readers may have been inclined to respond that Pearl (2001) *does* give a decomposition of the total effect into direct and indirect effects (and we've seen that Zhang and Bareinboim (2018) offer a similar decomposition). Specifically:

$$(5) TE_{0,1} = NDE_{0,1} - IE_{1,0}$$
$$(6) TE_{0,1} = NDE_{1,0} - IE_{0,1}$$

Let's focus on (5). This mathematical decomposition should *not* be understood as a decomposition into two independent contributions corresponding to each path. Note that the total effect of changing the treatment from 0 to 1 is not the sum of the direct and indirect effects of going from 0 to 1. Rather it involves subtracting the indirect effect of going from 1 to 0.

The decomposition could be validated simply by unpacking the terms and showing that it is true, but here I will attempt to give the reader and intuitive sense of what is going on. For ease of exposition, lets assume that changing the treatment from 0 to 1 raises the probability of the outcome, and that $Y = 1$ corresponds to the outcome's "occurring". The first term, $NDE_{0,1}$ concern the probability that the cause would *still* bring about the outcome if the change were only transmitted via the direct path. The second term $-IE_{1,0}$ can be rendered intuitive as follows. Imagine that the outcome *did* occur. What is the chance that it would not have occurred if it had not been transmitted via the indirect path. In other words, the first term captures whether the change is sufficient for bringing about the effect and the second captures whether the indirect path is necessary. (5) then says that in the cases where the direct path was *not* sufficient, it is because the indirect path was necessary.

The decompositions in (5) and (6) are significant insofar as they reveal that although there are four path-specific effects ($NDE_{0,1}, NDE_{1,0}, IE_{0,1}, IE_{1,0}$), knowing both direct effects plus the total effect is enough to cover all of them. There are thus fewer independent quantities than might appear at first. But they nevertheless underscore that one should not simply talk about the "contribution" of a path, but rather must relativize path-specific effects to particular changes in the treatment. Neither $NDE_{0,1}$ nor $NDE_{1,0}$, for example count as *the* contribution of the direct path.

In Zhang and Bareinboim (2018), the decomposition is not simply that of the total effect into direct and indirect effects, but rather the total variation into direct, indirect, and spurious effects

(SEs)[5]. *Total variation* is the statistical difference between the demographic groups with respect to the outcome. The decompositions are as follows:

$$(7) TV_{0,1}(Y) = P(Y|X=1) - P(Y|X=0) = SE_{0,1}(Y) + IE_{0,1}(Y|X=1) - NDE_{1,0}(Y|X=1)$$

$$(8) TV_{0,1}(Y) = P(Y|X=1) - P(Y|X=0) = NDE_{0,1}(Y|X=0) - SE_{1,0}(Y) - IE_{1,0}(Y|X=0)$$

These mathematical decompositions provide an important tool for evaluating policies that aim to reduce the total variation by targeting one of the effects, and Zhang and Bareinboim provide illustrative examples of how this works. While I take no issue with any of their illustrations or formal proofs, I do worry that talk of decomposition invites a conflation between the purely mathematical decompositions provided and the notion that there is a unique composition in terms of the activities of the paths. Such a conflation is encouraged by their claim that "[t]he counterfactual measures can explain how much of the observed disparity is due to their corresponding (unobserved) causal mechanism" (p. 7). Note that in (7) and (8) the "corresponding" path-specific and spurious effects (e.g. $NDE_{0,1}$ and $NDE_{1,0}$) reflect independent quantities. One cannot talk unambiguously about the amount of disparity due to a particular causal path, since the effects will be different depending on whether one is going from $X = 0$ to $X = 1$ or vice versa.[6] Furthermore, given some arbitrary intervention on a path (e.g. a CDE) there may be no decomposition available.

7 Indirect Effects and Partially-Backtracking Counterfactuals

In the cases considered so far, the natural direct effect was of primary interest. In this section I will suggest a novel use for indirect effects. Instead of interpreting direct effects in terms of their relationship to the total effect, one can instead focus on the particular way that one intervenes on the mediator and to understand this intervention as corresponding to what I will call a "partially-backtracking" counterfactual. Standard causal counterfactuals do not "backtrack" (Lewis, 1974), in the sense that when one reasons about how variation in a cause leads to variation in its effect, one does not suppose that the variation in the cause must reflect variation in *its* causes. This forward-looking feature of counterfactuals is reflected within causal models by the use of *ideal interventions* that render the intervened-upon variable independent of its causes. But the interventions used to evaluate path-specific effects are not like this. In manipulating a mediator to identify an indirect effect, one needs to explicitly consider its relationship to the treatment in order to know how to vary it. But the mediators do not fully backtrack, in the sense that one does not vary the treatment itself. I will now suggest that this type of counterfactual may be of use in contexts where one must account for discrimination occurring prior to the action being studied. I will illustrate this using Hausman's (2012) discussion of affirmative action in college admissions.

Hausman does not aim to settle the question of whether affirmative action is justified, but rather to differentiate between better and worse arguments both for and against it. Crucially, he rejects the position that it is always wrong to take race into account in hiring or university admissions and thus that such policies are "racism in reverse". He denies that there is a general moral principle that such decisions must take into account facts other than qualifications, and claims that the harm of discrimination must be understood in the context of the broader historical patterns of social stigmatization. That said, he does not defend such policies as a form of reparations. Rather he suggests they can be motivated on egalitarian grounds of striving to equalize opportunities.

But doesn't preferential admission in favor of minorities require treating white and black college applicants *un*equally? How, then, can it be defended on egalitarian grounds? Hausman's key point is that egalitarianism does not require equal opportunities to college admissions, but rather equal opportunities over a lifetime. Hausman imagines a white applicant who is denied admissions to a college, and complains that they were treated unfairly because they would have been admitted had the school not had their policy. He suggests that the question the student should ask instead is:

[5] The spurious effect $SE_{0,1}(Y)$ is defined as $P(Y_{X=0}|X=1) - P(Y|X=0)$.

[6] Zhang and Bareinboim evaluate most path-specific effects relative to the baseline of $X = 0$, which they stipulate corresponds to the the hypothesized disadvantaged category in a study. This is a reasonable choice for the purposes of studying discrimination, as opposed to privilege, but it potentially obscures the change-relativity of path-specific effects.

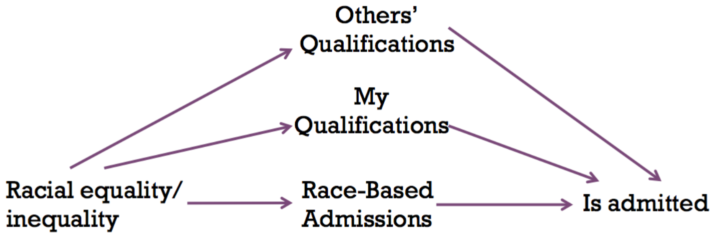

Figure 6: Mediation Model for Hausman Counterfactual

> How do my odds of being admitted to the university with its preferential admissions policy compare to what my odds would have been if the university had no such policy and I and all the other applicants had the qualifications we would have had if opportunities before college had been equal?

The underlying idea is that to the extent that the student's qualifications above other students are a result of prior unequal opportunities, the aim of promoting lifetime equal opportunity could justify admitting students who have fewer qualifications as a result of this disadvantage.

We can represent the counterfactual that Hausman presents using a mediation model (Figure 6). In it, the treatment is whether a society is racially equal or unequal, and the mediators are the applicant's qualifications, the qualifications of the other candidates, and whether the university adopts a preferential admissions policy. The counterfactual Hausman specifies corresponds to the indirect effect via preferential admissions for the change of the treatment from an equal to an unequal society. That is, for the mediators other than *preferential admissions*, one sets them to the value they would take on under a fair distribution of opportunities, while changing *preferential admissions* from the value it would have under a fair distribution (the policy would not be adopted) to the value it has under the existing unfair distribution.

Why does this indirect effect identify the counterfactual that Hausman treats as relevant to evaluating the policy's fairness? Given Hausman's understanding of fairness as being equal opportunities across a lifetime, a set of policies that privilege one group at one stage and another group at a later stage can still be fair despite not giving all groups equal opportunities at every stage. Accordingly, in evaluating the fairness of a particular action, one cannot just look at whether it gives all individuals equal opportunities moving forward, but must also take into account past inequalities of opportunity and the effects of these inequalities on the present. The mediation model presented does so by taking as the "treatment" a variable for whether initial opportunities were equal or unequal. This serves as a "normative baseline" for evaluating the fairness of actions. In the example of preferential admissions, if the aggrieved student would not have had a worse chance of admission under a scenario in which initial opportunities were equal, then she cannot claim that the policy violates her claim to equal lifetime opportunities. This is *not* to claim that there might not be other reasons for rejecting the policy, or to weigh in on whether it is compatible with existing equal protection laws. Instead, Hausman is claiming that such policies cannot be faulted on egalitarian grounds, as they do not reduce but rather promote equality of opportunity.

One might wonder whether Hausman's position specifically supports *race*-based preferential admissions, rather than other policies for mitigating the effects of past inequality. Additionally, there are subtle questions about how an egalitarian should specify a normative baseline. For instance, is the aim to reduce only inequality resulting from social institutions, or also "natural" inequalities? And is there a non-arbitrary way to draw this distinction? Causal models are unable to address these normative questions, and one should not expect them to. The present modeling exercise reveals that mediation models can serve as a flexible tool for contexts in which one wants to evaluate the fairness of an action in a way that takes into account historical factors that led to it. In cases where one understands fairness not merely as forward-looking procedural fairness, but rather as fairness over a time-frame beginning prior to the action or policy in question, one can build a mediation model in which the treatment variable defines the starting point of the time-frame.

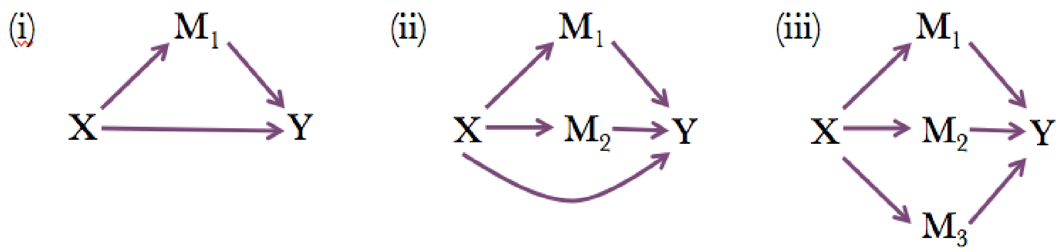

Figure 7: Caption

A Path-Specific Effects Across Models

The above characterization of path-specific effects in terms of the transmission of a change along a set of paths is helpful for keeping track of when a direct effect in one model corresponds to the same quantity as an indirect effect in another (or vice versa). Consider the DAGs in figure 7. Whereas DAG (i) contains a single mediator M_1, DAG (ii) includes an additional mediator M_2. Note that the direct effects in (i) and (ii) do not refer to the same quantity. The direct effect in (i) corresponds to the effect transmitted via M_2 as well as the direct path in (ii). How do we know that these are the same? The answer is that even when M_2 is not included in DAG (i), the direct effect corresponds to the effect transmitted along that path, and this includes mediators along that path not included in the model. Now suppose we include a third mediator M_3 and that there is no direct effect not via these three mediators (DAG (iii)). Now the the direct effect in (i) corresponds to the indirect effect via M_2 and M_3 in DAG (iii), and could be identified by simultaneously varying M_2 and M_3 to transmit the relevant change. Additionally, if one were remove M_1 to draw a graph with just M_2 and M_3, the direct effect in the resulting graph would correspond to the indirect effect via M_1 in (i). More generally, path-specific effects are equivalent across graphs if either: (A) They are indirect effects for a set of paths characterized by the same mediators, or (B) there exists a set of paths as characterized by (A) and the effect in each graph corresponds to the transmission of a change *not* along these paths.[7]

References

Barocas, S., M. Hardt, and A. Narayanan (2019). *Fairness and Machine Learning*. fairmlbook.org. url: http://www.fairmlbook.org.

Baron, R. M. and D. A. Kenny (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology 51*(6), 1173.

Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review 94*(4), 991–1013.

Bickel, P. J., E. A. Hammel, and J. W. O'Connell (1975). Sex bias in graduate admissions: Data from berkeley. *Science 187*(4175), 398–404.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, pp. 7801–7808.

Fryer Jr, R. G. (2019). An empirical analysis of racial differences in police use of force. *Journal of Political Economy 127*(3), 1210–1261.

[7] In principle there can be path-specific effects that are neither direct nor indirect, such as the effect transmitted via M_2 and the direct path in (ii). In such cases, however, there does exist some DAG (i.e. (i)) in which the same effect is a direct effect. Note that the DAGs presented here do not involve direct causal relationships between the mediators. This considerably simplifies the discussion, though is adequate for our aims in this paper.

Graetz, N., C. E. Boen, and M. H. Esposito (2022). Structural racism and quantitative causal inference: a life course mediation framework for decomposing racial health disparities. *Journal of Health and Social Behavior*, 00221465211066108.

Greiner, D. J. and D. B. Rubin (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics 93*(3), 775–785.

Hausman, D. M. (2012). Affirmative Action: Bad Arguments and Some Good Ones. In R. Shafer-Landau (Ed.), *The ethical life: Fundamental readings in ethics and moral problems*, pp. 432–445.

Jackson, J. W. (2018). On the interpretation of path-specific effects in health disparities research. *Epidemiology 29*(4), 517–520.

Knox, D., W. Lowe, and J. Mummolo (2020). Administrative records mask racially biased policing. *American Political Science Review 114*(3), 619–637.

Kohler-Hausmann, I. (2018). Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev. 113*, 1163.

Lewis, D. (1974). Causation. *The journal of philosophy 70*(17), 556–567.

Pearl, J. (2001). Direct and Indirect Effects. In J. Breese and D. Koller (Eds.), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420.

Pearl, J. (2014). Comment: Understanding Simpson's Paradox. *68*(1), 8–13.

Plecko, D. and E. Bareinboim (2022). Causal fairness analysis. *arXiv preprint arXiv:2207.11385*.

Robins, J. M. and S. Greenland (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 143–155.

Sprenger, J. and N. Weinberger (2021). Simpson's Paradox. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University.

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

Weinberger, N. (2019). Path-specific effects. *The British Journal for the Philosophy of Science 70*(1), 53–76.

Weinberger, N. (2022, February). Signal manipulation and the causal analysis of racial discrimination.

Zhang, J. and E. Bareinboim (2018). Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.