CAUSAL INFERENCE ACROSS POPULATIONS

By

Naftali Weinberger

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy
(Philosophy)

at the
UNIVERSITY OF WISCONSIN-MADISON
2015

Date of final oral examination: 5/12/15

The dissertation is approved by the following members of the Final Oral Committee:

Daniel M. Hausman, Professor, Philosophy (Chair)
Elliott Sober, Professor, Philosophy
Malcolm Forster, Professor, Philosophy
Felix Elwert, Associate Professor, Sociology
Peter Steiner, Assistant Professor, Educational Psychology

Acknowledgements

The philosophy department at the University of Wisconsin has provided me with an amazingly supportive environment within which to pursue the research for this dissertation. I am indebted to the professors in the department for devoting so much time to advising me and introducing me to the world of professional philosophy, and to the graduate students for creating a collegial environment in which the success of individuals was always perceived to be the success of all. Here I would like to thank the particular individuals without whom this project could not have been such a success.

First and foremost, I would like to thank Dan Hausman, my thesis advisor. He has been unbelievably helpful at every stage of this process, reading multiple drafts of each chapter and providing thorough comments. Although I have never taken a seminar with Dan, through engaging with him I was able to get a foothold into the dense literature on causation. I feel very fortunate to have had him as an advisor. Next, I owe a great deal to Elliott Sober, who was invaluable for helping me make the transition to graduate school and was always willing to talk with me about my work. Malcolm Forster challenged me to think about the fundamental assumptions of causal modeling, and this manuscript is much richer as a result of our many long conversations over beer. Finally, I'd like to thank Peter Steiner and Felix Elwert for allowing me to sit in on their causation course and for meeting with me often to discuss my work. Their input helped me see how I could pitch the dissertation to a much broader audience.

I owe a special debt to Reuben Stern, who has collaborated with me on several projects and helped me work through some thorny issues at the end of the dissertation. Shannon Nolen provided constant feedback and invaluable support and deserves much of the credit for the lucidity of chapter 5. At every stage of graduate school, I benefitted from discussing my work

with Hayley Clatterbuck, and I would like to especially thank her for her support as we navigated the job market this year.

The following people were also very helpful at providing feedback on parts of the dissertation: Elias Bareinboim, Judea Pearl, Nancy Cartwright, Fabienne Peters, Kosuke Imai, and Olav Vassend. Thanks to Adam Gamoran and the Interdisciplinary Training Program at the Wisconsin Center for Educational Research for allowing me to participate in the Fall 2012 seminar on causal mediation techniques in education. I would also like to thank the staff at Barriques Coffee Trader and the owners and staff of Johnson Public House (especially Lenin) for keeping me caffeinated at key points in the project. Finally, thanks to my parents and to my brother Michael for their constant support and encouragement.

Abstract

This dissertation addresses the problem of when one can infer that a causal factor that has an effect in one population would have a similar effect in others. For example, if reducing class sizes increases educational outcomes in one neighborhood, would reducing them in another raise outcomes by a similar amount? This is known as the problem of extrapolation. This dissertation reviews the prior literature on extrapolation, explains how extrapolation relates to other forms of causal inference, and presents techniques for extrapolating more reliably.

To model extrapolation, I rely on recently developed causal modeling techniques, which use graphs to represent causal relations among variables. I build on the work of Judea Pearl and Elias Bareinboim, who provide graphical methods for determining when it is possible to extrapolate causal quantities across populations given assumptions about how those populations differ. I argue that their account explicates one type of extrapolation, but that there are extrapolative inferences that go beyond their account.

The central positive contribution of the dissertation is that it makes precise the sense in which knowing how a cause brings about its effect facilitates extrapolation. In cases where a cause influences its effect via multiple paths, newly developed "causal mediation techniques" enable one to precisely quantify the way that the cause influences its effect via each of the paths. These techniques aid extrapolation, since the causal quantities identified by these techniques are invariant across a range of ways that two populations may differ. Moreover, the conditions across which these quantities are invariant cannot be represented within Pearl and Bareinboim's framework.

In discussing the problem of extrapolation, I touch on several central philosophical issues. First, characterizing the problem requires one to elucidate the relationship between causal

and statistical inference. Second, the causal mediation techniques I advocate shed light on recent debates about mechanistic explanation. Finally, the study of the conditions under which causal relationships generalize is essential for understanding the nature of causal relationships and their role in scientific theories.

Table of Contents

Chapter 1: Introduction

A woman walks into a bar and orders a gin and tonic. She insists that the gin be poured first. The bartender looks perplexed. She explains that she has the unusual talent of discerning which ingredient was poured first – even after the drink has been mixed – and she wouldn't even think of consuming a drink with the tonic water poured first. How can we test whether she in fact has this talent? R.A. Fisher (1935) provides a method for doing so. Being British, his own example involves tea and milk rather than gin and tonic, but the idea is the same. Make several drinks and randomly assign some of the drinks to be mixed one way and the others to be mixed in the opposite manner. If the woman is right more often than would be expected based on guessing, this provides reason to think that she can tell the difference.

Suppose, surprisingly, that one gives the woman 100 drinks – on different days, one would hope – and she is right about 95 of them. This would provide extremely good evidence that the order in which one pours the drinks causally influences the woman's verdicts. While this experiment no doubt succeeds beyond the bartender-researcher's wildest dreams, it by no means guarantees that the woman would have similar success in slightly modified contexts. The test results are consistent, for example, with the possibility that had the room been a few degrees warmer (on average), the woman would entirely lose her predictive powers.

More generally, establishing that a cause obtains in one context, person or population does not entail that there will be a similar causal effect in situations that differ from the one in which the causal claim was established. In the frivolous case of the tea lady, we might be more than satisfied living with this uncertainty, since any false predictions about the woman's performance in another context will have minor consequences. In many real scenarios, the consequences of a false prediction are both tragic and expensive. If a cancer drug has benefits in

a human study population, will it have similar effects in the broader population of cancer patients? Alternatively, will a drug that works in rats also work in humans? To give a non-medical example, will a policy that decreases poverty in one city also do so in other cities? All of these questions concern whether a cause found in one context will *extrapolate* to another context. This dissertation explores the degree to which one can address questions involving causal extrapolation using the currently (and increasingly) popular causal modeling frameworks developed by Pearl (2009), and Spirtes, Glymour and Scheines (2000). These frameworks use Directed Acyclic Graphs (DAGs) to represent the causal relations between random variables.

In approaching the problem of extrapolation, I will generally assume that there is a population regarding which we have knowledge of a causal relation or set of causal relations that are of interest. This is called the *study population*. To extrapolate is to make an inference about the nature of the corresponding causal relationships in a *target population* (or a set of target populations) regarding which one does not have the same degree of causal knowledge. A qualitative extrapolation is an inference regarding whether the causal relationship is present in the target population and also regarding its direction of influence – does the cause promote or inhibit its effect? A quantitative extrapolation is an inference regarding the magnitude of the effect in the target population. For example, the magnitude of the effect of having the flu on body temperature might be to raise it by 3°F. In contexts where the causal relationships among dichotomous variables are probabilistic, the magnitude of the effect is the degree to which the cause raises the probability of its effect.[1]

Causes do not typically act in isolation, but depend on the presence of various background factors for their activity. Striking a match causes it to light only if there is oxygen in

[1] I will remain neutral regarding whether the probabilities in causal models are metaphysical or epistemic. That is,

the room and the tip of the match is not wet. In more complicated scenarios, there can be an indefinite number of unknown background factors. Will implementing the Common Core curriculum lead to higher standardized test scores? Even if the curriculum is effective in some places, its success in others will depend on the presence of adequate educators, on students having adequate time to do homework, and on facts about the prior education of the students, and a variety of other factors that policymakers might not have the resources to measure, or might not even be aware of. What makes extrapolation difficult is that the background factors in the target population can differ from those in the study population and one has no way of knowing that one has measured all such factors. Worse, one has no way of knowing how the unmeasured factors influence the causal relationship of interest. Let's refer to this as *the problem of unknown unknowns* (with apologies to Mr. Rumsfeld).

I can think of four general responses to this problem. The first is the *minimal sufficient condition* approach. According to this approach, one should try and find a set of background factors that are jointly sufficient for bringing about the effect and then evaluate whether these factors obtain in the target population. The second is the *natural kinds* approach. The idea behind this approach is that if we find that the effect of *C* on *E* is very sensitive to an indefinite number of background conditions, we should try and re-characterize the relationship with different variables *C'* and *E'* that are not similarly unreliable. The third is the *inductive* approach. On this approach, in cases where we cannot find a set of minimal sufficient conditions for a cause bringing about an effect, we should seek evidence that inductively supports the belief that the causal relationship will obtain in the target population. The fourth is the *mediation* approach. According to this approach, in order to determine whether the effect of *C* on *E* will generalize, we need to determine *how* the cause brings about its effect. This is called the mediation

approach, since a *mediator* is a variable that is causally intermediate between a cause and its effect (i.e *C*→*Mediator*→*E*). This is the approach that I primarily pursue in the dissertation. I further discuss the other three approaches and my reasons for focusing on mediation below.

The mediation approach could also be referred to as the *mechanisms* approach, which is the way that Daniel Steel characterizes his account of extrapolation. My reason for calling it the mediation approach is that there is currently a large philosophical literature on mechanisms and I do not want to be construed as adopting the assumptions that are common in this literature (see chapter 5). This literature explores the way that the behavior of a physical mechanism is explained by the contributions of its components and it is generally assumed that this explanation must appeal to something other than the causal relationships between the components of the mechanism. In contrast, the relationship between a cause, its effect, and a mediator is characterized entirely by the causal relationships among these variables. If there is more to being a mechanism component than having a property that is causally in-between the input and output of the mechanism, this additional element plays no role in my account.

It is not difficult to think of cases in which learning how a cause brings about its effect enables one to extrapolate across scenarios. Suppose that among ex-convicts, being employed promotes a lower rate of recidivism. Here are two plausible stories that could help explain this causal relationship. One, employed individuals have more funds and therefore are less tempted to engage in criminal activities that help them procure goods that they could not otherwise afford. Two, employed individuals have less free time, and therefore less time to spend on illegal activities. Both of these stories could be true, and there could be other explanations for the effect of employment on recidivism. If, implausibly, we were to learn that the effect of employment on recidivism were entirely mediated through increasing ex-convicts' cash-in-pocket – that is, if

there are no variables on paths other than that of *employment→cash-in-pocket→recidivism* – this would have important policy implications for the conditions under which this effect generalizes. For example, the magnitude of the effect would presumably depend on how much newly employed ex-convicts are paid in a particular area. In areas where they only receive a meager salary, employing ex-convicts would be a less effective means of reducing recidivism.

While it is clear that learning how a cause brings about its effect facilitates extrapolation, it is not clear how the mediation approach could resolve the problem of unknown unknowns. In the recidivism example, the problem is that the effect of employment on recidivism might depend on an indefinite number of factors of unknown influence. Here it is extremely plausible that we would never know the myriad of background factors that one would need to know to determine whether a particular individual will end up back in prison. Even once we are told that the effect of employment on recidivism in an individual is mediated by the amount of cash he has, these background factors can still make a difference for the effect of employment on cash-in-pocket for that individual and in the effect of cash-in-pocket on whether he returns to prison. More generally, if we are uncertain how much the effect of X on Y varies across populations, it is unclear how measuring a mediator between X and Y reduces this uncertainty.

Part of the answer to this question is that in some cases, the relationship between X and M in the target population is more easily ascertainable than the effect of M on Y. In the present example, one could get a reasonably good estimate of the effect of employment on cash-in-pocket by looking at salaries. Prior to measuring the mediator, one had no way to distinguish cross-population variation in the effect that results from salary differences and cross-population variation that results from differences in other background factors. Once one measures the mediator, one can isolate the cross-population variation that is due to variation in the effect of

employment on cash-in-pocket. In the simplest case, learning that the magnitude of the effect of the treatment on the mediator does *not* differ between the study and the target population will increase one's confidence that the total effect of employment on recidivism will be similar across the populations. Of course, saying that discovering that the $X \rightarrow M$ relationship is invariant across populations should make you *more* confident that the $X \rightarrow M \rightarrow Y$ relationship will be similar across the populations does not tell you *how* confident you should be. But this still counts as progress.

The primary limitation of the approach just sketched is that it relies on the assumption that the effect of employment on recidivism is entirely mediated by the amount of funds that the ex-convicts have. As I noted at the outset, however, there is another mediator that could also make a difference in this effect. Namely, being employed could reduce recidivism in part by reducing one's free time. Moreover, there could be many other ways that employment affects recidivism. Many of them we might not know about or, even if we do, we won't know how to measure them. Hopefully, being employed increases one's self-esteem, which in some contexts would reduce one's chance of committing crimes. Yet, it is difficult to measure latent psychological states and to do so in such a way that the variable one measures is the relevant one. There are many ways of measuring self-esteem, and one would have to measure the one that is influenced by employment and which influences recidivism.

It gets worse. Even if one *could* measure enough mediators such that there is one corresponding to every way that employment affects recidivism, the mediators can interact in producing their effects. For example, the effect of an increase in cash-in-pocket on recidivism might depend on the amount of free time one has. It could be that an increase in cash will reduce recidivism among employed people with little free time, but increase recidivism among

unemployed people who have a lot of free time. This is plausible if the crime in question is drug use. Increasing free time incentivizes drug use, while increasing cash-in-pocket makes it possible to obtain drugs. So in order to know about the effect of employment on recidivism going through a particular mediator, it appears that one needs to know about the activity of all of the other mediators. Yet, one might not know what these even are. In trying to make headway on the problem of unknown unknowns by measuring mediators, we appear to have run into a problem that is just as intractable.

Fortunately, this problem can be solved. Or at least that is what I intend to show in this dissertation. In order to solve it, we need to get a clearer picture of the relationship between the effect of *C* on *E* going through all mediators and path-specific effects going through particular mediators. Philosophers have long been aware that the total effect of *C* on *E* can differ from the path-specific effects going through particular mediators. Perhaps the most famous example of this is Hesslow's (1976) thrombosis case. Birth control pills raise one's risk of thrombosis by producing a certain chemical in the blood, but they lower one's risk of thrombosis by decreasing one's chance of getting pregnant, since pregnancy itself is a risk factor for thrombosis. Accordingly, birth control pills exert both positive and negative component effects on thrombosis. Whether the total effect is positive or negative depends on the relative strengths of the component effects.

Despite the great deal of attention that has been devoted to Hesslow's example, to my knowledge no philosopher has produced a correct general account of what it means for a path-specific effect to be positive or negative, much less explained how one could measure its magnitude. Yet, this question was answered in the causal modeling literature in 2001 in Pearl's

article “Direct and Indirect Effects”. Consider the causal model for the recidivism example in figure 1.

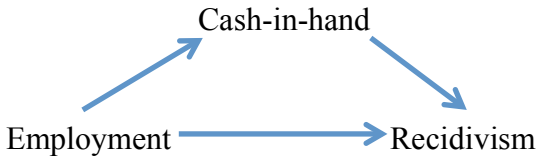

Figure 1

In the graph in figure 1, the path going from employment to cash-in-hand to recidivism is the *indirect path*; the other path is the *direct path.* The *direct effect* is the effect that employment would have on recidivism if there were no indirect path, and the *indirect effect* is the effect that employment would have on recidivism if there were no direct path. The effect going through all paths is the *total effect.* The definitions of direct and indirect effect are model relative, since the direct effect is the effect that is not due to any mediators that are included in the model. It is possible to measure the direct and indirect effects and to specify their contributions to the total effect using *causal mediation techniques.*

Here in the introduction I will say very little about how causal mediation techniques work. The crucial feature I would like to highlight here is that in order to identify the direct and indirect effects it is not necessary to measure a mediator along every path between employment and recidivism. The direct path corresponds to the effect of all mediators that are not on the indirect path. What this means is that in identifying the indirect effect, one is able to determine how employment would influence recidivism via cash-in-hand even if all of the unknown paths corresponding to the direct one were disabled. This is the reason that measuring mediators enables one to make headway on the problem of unknown unknowns. Not only is it possible to

evaluate the counterfactual contribution of particular paths in the absence of the others, but one can do so without specifying anything about the other paths.

The way that identifying the indirect effect in figure 1 facilitates extrapolation is analogous to the way that measuring the mediator *cash-in-hand* would facilitate extrapolation were there to be no direct path at all. By comparing the effect of employment on cash-in-hand in two populations, one isolates a potential source of cross-population variation. As noted above, even if the two populations do not differ at all in the effect of employment on the mediator, there is still no guarantee that the indirect effect will be similar in the target population, or even that it will probably be similar. What one *can* say is that any variation in the indirect effect across populations is due to variation in the effect of the mediator on recidivism. At first glance, this claim may seem trivial. Since the indirect path is the concatenation of the causal arrow from employment to cash-in-hand and the causal arrow from cash-in-hand to recidivism, and we are assuming that we know about the relationship in the target population corresponding to the first arrow, isn't it obvious that all remaining variation corresponds to the second arrow? This only seems obvious, however, if one forgets that as in fact there *is* a direct path and that *employment* and *cash-in-hand* can interact in their effect on *recidivism*. To say that all cross-population variation in the indirect effect is due to the effect of *cash-in-hand* on recidivism is to exclude the possibility that it is due to the direct influence of *employment* on *recidivism.*

I have now briefly sketched the way that I will ultimately use causal mediation techniques as a means to facilitate cross-population inferences. It will take the rest of the dissertation to fill in the details. Before getting to causal mediation techniques, I will first review the prior literature on extrapolation. I will then introduce causal mediation techniques, argue that

the role they play in extrapolation is not just an application of previously existing methods for extrapolating using causal graphs, and then, finally, explain how they facilitate extrapolation.

In the remainder of this chapter, I discuss some philosophical issues that will arise repeatedly throughout the dissertation. I start by briefly considering the relationship between extrapolation, causal inference and statistical inference and present a difficulty with representing extrapolations using DAGs. I then say a bit more about my use of the term "populations" and about the relationship between populations and subpopulations. Next I explain why my presupposition that the study and target populations can be represented within a single DAG is a substantive assumption. I then discuss the approaches to extrapolation that I do not pursue in the dissertation and provide a sense of the challenges related to those approaches. I conclude with a brief synopsis of the following chapters.

Extrapolation, Causal Inference and Statistical Inference

While it is clear that there are situations in which researchers face the challenge of knowing about the magnitude of a causal relationship among one group of individuals and not knowing whether that relationship is similarly strong in another group, it turns out to be surprisingly difficult to characterize this problem within the DAG framework. In this section, I discuss the project of making causal inferences from probability distributions and explain why it is difficult to characterize extrapolation within the DAG framework. I provide a resolution to these difficulties in chapter 4. While I briefly introduce a few elements of the DAG framework in this section, I will provide a more thorough introduction to this framework later in the dissertation.

DAGs enable one to represent one's knowledge of the causal relationships among a set of variables. A DAG is a graph in which the nodes are variables and the edges are arrows representing causal relationships among the variables. It is directed, since all of the edges are

represent asymmetric causal relationships, and acyclic: one cannot get from a variable back to itself via a set of directed edges. A DAG is associated with a joint probability distribution over the variables in the DAG.

Using DAGs, it is possible to determine when a causal relationship between variables can be *identified* from the probability distribution. For a causal quantity to be identifiable is for its value to be uniquely determined from the probability distribution given the assumptions embedded in the DAG. For example, if *C* causes *E*, and there are no common causes of *C* and *E* in the DAG, then the effect of *C* on *E* is identifiable. It corresponds to the correlation coefficient for *C* and *E.* In contrast, if there is an unmeasured common cause of *C* and *E*, then the effect of *C* on *E* is not identifiable. The degree of correlation between these variables is not be a reliable guide to the magnitude of the effect, since it is biased by the presence of the common cause.

Correlations – and other features of the probability distribution – are not observed, but rather inferred from finite samples. Using statistics, one can make inferences regarding when one's sample is large enough that the relative frequencies of traits in the sample are a good guide to their relative frequencies in a hypothetical population from which one is sampling using an unbiased sampling process. In causal inference, it is common to assume that one knows the joint probability distribution for a population. In assuming that one has knowledge of *the probability distribution* for a population – rather than just knowledge of the relative frequencies of traits in the population – one bypasses all questions about how one infers the probability distribution from finite data.

Causal inference from probability distributions is often contrasted with causal inference from experiments. In the present discussion, I do not intend to be making this contrast. Even in a well-executed experiment with perfect randomization, the experiment is only able to measure the

magnitude of the causal effect if one has a large enough sample to limit sampling variation by the desired amount. The purpose of randomly assigning some test subjects to take a drug and some to take a placebo is to render whether one receives the treatment or the placebo random with respect to an individual's causally relevant properties. With a larger sample, the treatment and control groups will with increasing probability be 'balanced' with respect to all properties that might make a difference in the magnitude of the effect. With a smaller sample, there is a significant chance that even though the assignment is random, the two groups might differ in causally relevant properties. Whether the sample is large enough to effectively eliminate these differences is a statistical question. Causal inference from experiment resembles causal inference from probability distributions in that it abstracts away from statistical questions of how to determine whether one has an appropriately large sample.

The causal effects that are identified using DAGs and probability distributions are typically *average* effects. For the effect of C on E to be identifiable, it is not necessary to measure *every* cause of E, but only every common cause of C and E. This is the case even though there can be causes of E that make a difference in the magnitude of the effect of C on E. These causes are what I referred to earlier as *background factors*. The reason that one does not need to measure these background factors is that they are allowed to vary randomly in the population. Even though individuals in the population with different combinations of background factors will have different effect magnitudes, the average effect will be the effect across the different individuals.

In the dissertation, I typically assume that one knows the DAGs for both the study and target populations and that one is able to establish causal directionality (e.g. that if there is a direct causal relationship between variables X and Y, X causes Y and Y does not cause X). In fact,

I will generally assume that the two populations are representable by the *same* DAG. The question of extrapolation is that of inferring the magnitude of the causal effect in the target population based on one's knowledge of its magnitude in the study population. In order for two populations with the same DAG to differ in the magnitude of an effect, they must have different distributions of background factors. If the two populations had the *same* distributions of background factors, they would have the same causal effects with the same magnitudes, and the problem of extrapolation could not arise.

To understand extrapolation within the DAG framework, one must clarify what it is for two populations to have different distributions of background factors. Of course, if one knew what the background factors were, it would be trivial to specify how they differ across the populations. But one does not know what they are. In identifying an average effect from the probability distribution, one identifies the average effect across these unknown factors. Yet, it is non-trivial to specify *which* distribution of background factors one should average over in estimating the probability distribution for a particular sample.

Consider, for example, a drug trial performed in Madison, Wisconsin to determine the effect of a drug on cholesterol. One could represent this effect in a DAG with the variables *takes drug* and *cholesterol* level. One would also presumably include variables (covariates) that one believes to make a difference in the effect of the drug on cholesterol, but here we will consider the simple two-variable model to keep things simple. Even if one includes some variables that make a difference for the effect, one will almost never be able to include all of them – there will still be some unmeasured background factors. In estimating the probability distribution for *takes drug* and *cholesterol level* in the sample, which distribution of background factors is one trying to average over? The distribution of factors in Madison? In Wisconsin? In the Midwest? Nothing

in the DAG itself differentiates among these options. Nevertheless, the claim that one knows the probability distribution for a population presupposes some answer to this question.

By way of illustration, consider the proposal that one should average over the widest possible distribution of background factors. If we were to adopt such an idealized notion of a probability distribution, this would define the problem of extrapolation out of existence. Any DAG with the same set of variables would be defined relative to the *same* set of background factors, and would therefore have the same causal effects. For example, suppose that Wisconsin and Michigan differ in their consumption of aged cheddar and that the drug is more effective in reducing cholesterol in people who eat more aged cheddar. It intuitively seems like effect of the drug on cholesterol would differ between the two states. If, however, we represented this effect using a DAG containing just the variables for the drug and cholesterol, the average effect in this DAG would not correspond to the effect in Wisconsin or Michigan, but rather the effect given the average consumption of aged cheddar across the states. In order to make sense of the problem of extrapolation, the distribution for the variables in a DAG cannot average over the widest possible distribution of background factors, but rather, it must average over a more narrowly defined set.

A more realistic proposal is that when one seeks to estimate the distribution for the variables in one's sample, one has some rough idea of the broader population whose distribution one is trying to estimate. For example, one might take the sample in the study to be representative of people in Wisconsin, and remain agnostic regarding whether it is representative of a broader population including Michiganites. This proposal seems correct, though note that in specifying that one's sample enables one to estimate the average effect across background factors in Wisconsin, but not across factors in the broader population, one is assuming that it is

possible to extrapolate from the members of one's study to the Wisconsin population, but not to the broader population. If we assume that one's probability distribution averages over the background factors for the Wisconsin population, it is clear why the distribution may fail to apply to Michigan – the distribution of background factors might be different in Michigan. Yet, if two populations are distinct only if they differ in their distributions of background factors, why doesn't this rule out the possibility of extrapolating?

Here my aim is simply to flag some of the difficulties of understanding the problem of extrapolation within the DAG framework. In Chapter 4, I will explain how one can use DAGs to distinguish between populations in such as way that the magnitudes of the effects in the populations may differ, without ruling out the possibility of extrapolation.

Individuals and Inferences Between Populations and Subpopulations

In the dissertation, I use the term "populations" in a broad sense such that anything that can be represented by a probability distribution counts as a population. For instance, one might talk of populations of *events*. If striking a match causes a match to light in a certain percentage of cases, we can talk about the population of match strikings. Once one characterizes a set of events in terms of variables with a joint probability distribution, it is straightforward to think of that set of events in terms of a representative population of instances of those event types.

In the broad sense that I use the term "population", it also makes sense to discuss the population corresponding to an individual of a particular type. If one characterizes an individual by a set of properties, the probability distribution for those properties provides information about the correlations between the traits of that individual. In the same way that one cannot determine the probability of a coin's landing heads based on a single flip, one cannot determine the probability distribution over the properties of an individual of a certain type based on the history

of a token individual of that type. For example, an individual characterized by a particular set of variable values might have a 30% chance of developing heart disease if she eats a lot of red meat. In an infinite population of individuals who share the same values for those variables, 30% of them who eat a lot of red meat will develop heart disease. The sense in which it makes sense to think of an individual as corresponding to a population is that the probability distributions for that individual can be spelled out in terms of the population that would result from randomly sampling from individuals of that type.

Since DAGs represent the causal relationships among variables, and variables represent properties, the causal relationships I discuss are relationships between properties. These are sometimes referred to as "type-level" causal relationships. In saying that I will be considering type-level causal claims, I do not mean to commit myself to the view that type-level causal claims are fundamentally different from token causal claims. I intend for everything I say in what follows to be compatible with (but not to presuppose) the view that type-level causal claims are generalizations over so-called token causal claims (Hausman, 2005). I will not commit to a position in the debate regarding whether there is one or many concepts of cause. If there is a type of causal claim that is fundamentally distinct from and unrelated to the type-level causal relations represented by DAGs, the following discussion will not apply to such claims.

One can divide up a population corresponding to the probability distribution for variables $V_1,...,V_n$ into *subpopulations* by stratifying it based on the value of a variable V_{n+1}. For example, if one has identified the effect of smoking on cancer in a population of individuals, one can stratify that population based on age, to yield groups such as 'smokers between the ages of 20 and 30'. There is an important relationship between the causal relationships among the individuals in the population and individuals in its subpopulations. The magnitude of the effect

of *C* on *E* in a population is a weighted average of the effects in all subpopulations (Weinberger, 2015). To illustrate, the effect of smoking across all age groups will be the average of the effect in each age group, weighted by the size of each group.

The fact that an effect in a population is an average across subpopulations has an implication for extrapolation. *C* cannot be a cause of *E* in a population, but not in any of its subpopulations. If there were no effect of *C* on *E* in any subpopulation, there would be no effect in the population as a whole. Of course, there could be great variance in the effect of *C* on *E* across populations. Just because *C* influences *E* in the population, it does not follow that *C* will similarly influence *E* – or influence *E* at all – in a particular subpopulation. Yet, the fact that *C* cannot cause *E* in a population without also causing *E* in at least some populations reveals that knowledge of the effect in a population is *evidence* that there will be a similar effect in a subpopulation. If one suspects that there is a large amount of variance in the effects among the subpopulations, one will count it as very *weak* evidence. Nevertheless, the mathematical relationship between effects in populations and in subpopulations ensures that learning about the effect in a population will provide some information about the effect in subpopulations.[2]

Presuppositions of DAGs

Throughout the dissertation, I will assume that both the study and target populations can be represented using acyclic graphs. This is a substantive assumption. Hausman, Stern, and Weinberger (2013) show that not every system of variables can be given a graphical causal representation see also Druzdzel and Dash, 2001, for a similar analysis). Several philosophers, including Hausman (1998), have claimed that the causal relationships between variables depend

[2] Note that the proposition that *C* causes *E* in some of a population's subpopulations does not entail that *C* causes *E* in the population. It could be the case that *C* has a positive effect on *E* in some populations and a negative effect in others, and that there is no average effect.

on those variables being instantiated within a particular type of system. For example, the ideal gas law (PV=kT) does not by itself determine whether, e.g., temperature increases cause volume increases. However, relative to a system in which the gas is in a sealed container, temperature does cause volume. Hausman et al. present a mechanical device that allows one to switch from a system with one causal structure to a system with another. They argue that there is no DAG that accurately represents the whole system.

What Hausman et al.'s example reveals is that it is possible to have two systems with the same set of variables such that although each system can be represented with a DAG, there is no way to represent interventions that change one system into the other. If the study population has causal relations corresponding to one of the systems and the target population has a set of relations corresponding to the other, DAGs will not be useful for extrapolating across populations in such a case. Further investigation is needed to determine the conditions under which two populations cannot be represented in a single DAG.

An even more basic point is that in order to represent two populations with *N* variables in a DAG with the same *N* variables, the variables must be the same in both populations. This requirement becomes problematic when one extrapolates across systems that are extremely different. One of the areas where extrapolation has been most discussed is with respect to animal models. When are rats a good model for the way that humans will respond to a certain type of treatment? Suppose one discovers that exercise is more effective in non-obese rats than in obese rats. Will this result apply to humans? The problem with this extrapolation is not merely that there might be background factors that vary between rats and humans. A more immediate problem is that of whether *obesity* counts as the same variable when measured in rats and in

humans. Unless it does, it is unclear how one can make a meaningful comparison between the populations.

Even before one considers extrapolations across populations, the limitations on what can count as a causal variable *within* a population are more stringent than one might at first suspect. For each variable within a causal model, it cannot be the case that there are different surgical interventions – that is interventions that influence other variables in the model only *via* influencing that variable – such that different ways of changing the values of that variable lead to different downstream effects. This requirement places constraints on what can count as a causal variable in a model. If a drug is more effective when administered intravenously than when it is administered orally, one could not have a variable corresponding to whether one receives the drug, since different ways of receiving the drug lead to different downstream effects. The correlate of this for extrapolation is that when considering the DAGs for two populations, one must ensure that the variables are defined relative to the same types of interventions. If both populations include a variable for receiving the drug, but only one receives it intravenously, the effects will differ across the populations. The reason for this is not that the populations differ in background factors, but that the effects being considered in the populations are different effects.

In this dissertation, I consider cases in which the causal relationships in populations differ as the result of differences in background factors. While it may seem trivial to say that an effect that differs across two populations must differ as a result of differences in background factors, many assumptions must be made before one can declare that an effect is in fact the same effect across both populations. Sadly, I have little to say here about when one is justified in adopting the presuppositions discussed in this section. The proposals in this dissertation will fail to apply

to cases where two populations correspond to systems that are not representable using a single DAG or to systems with different variables.

Other Approaches

In the overview, I enumerated four approaches to extrapolation: the minimal sufficient conditions approach, the natural kinds approach, the inductive approach and the mediation approach. I have already sketched the mediation approach, which is the one that I will pursue in the dissertation. I did not begin this project with a commitment to the mediation approach. Here I will present some of the obstacles I ran into in trying to develop the other three. Perhaps someone else will be able to develop them more successfully.

The minimal sufficient condition approach seeks to find the factors necessary for a cause to bring about its effect. The primary limitation with this approach is that it is often not feasible to provide a complete set of factors that are sufficient for bringing about the effect. If this approach is to be at all useful, one must clarify how finding background factors facilitates extrapolation even when one does not have the complete set of factors. There are, in fact, some cases where it is useful to know about the contribution of a background factor even when there are other unknown factors. For example, if one knows that a particular factor is *necessary* for a cause bringing about its effect, then it follows that if that factor is absent, the cause will fail to bring about its effect no matter what other factors are present.

In chapter 2, I consider Cartwright and Hardie's account of extrapolation. This account combines the minimal sufficient conditions approach and the natural kinds approach. While Cartwright and Hardie are able to show how finding a necessary cause enables one to avoid making a bad extrapolation, I argue that their approach is much less useful for determining when one *can* extrapolate. I do not see a way to develop the minimal sufficient condition approach so

that it does not have this limitation. The fundamental problem is the problem of the unknown unknowns. There are many possible background factors and outside of the special cases where a factor is known to be either necessary or sufficient for the cause to bring about its effect, one has no way of knowing what difference would result from cross-population variation in these factors.

The natural kinds and inductive approaches strike me as being more promising than the minimal sufficient condition approach, but developing either would itself require a book-length treatment. The natural kinds approach is motivated by the observation that the development of new scientific theories has often involved revisions in the set and number of properties that are believed to generate a phenomenon. Newton's theory replaces celestial and terrestrial forces with a gravitational force that explains both motions of objects towards the Earth and of the planets around the Sun. Lavoisier explains oxidation as involving the addition of oxygen, rather than the removal of phlogiston. In the same way as characterizing a pill as "acetaminophen" rather than "Tylenol" allows one to get certain extrapolations for free – since the inference from the claim that Tylenol works to the claim that some other drug with acetaminophen works no longer counts as extrapolation – major revisions of scientific concepts lead to similar expansions of one's ability to extrapolate. Physicists do not need to measure the gravitational constant every time they calculate a new trajectory. According to Newtonian mechanics, the force exerted by its object is always proportional to its mass.

I will have little to say about how one determines which variables to use in one's causal model. The types of causal models I will be using are clearly sensitive to the way that one specifies the variables in one's model. For example, the identification of causal models with probability distributions depends on there being a correspondence between whether variables are causally related and whether they are correlated. Yet, whether two events are correlated is not

invariant across all ways of describing them. Flipping a coin is uncorrelated with its landing heads, unless one provides an extremely fine-grained description of how one flipped the coin. The question of how one's causal model is sensitive to one's choice of variables is conceptually prior to that of how one can determine whether the causal relations among the variables in one population also obtain among the variables in another. While there has been some preliminary work on the question of when one can aggregate variables in a causal model (Iwasaki and Simon, 1994), most contemporary discussions of DAGs take the variable set as given. Since extrapolation is hard enough even given a variable set, I will take the variables in a model as given as well. Questions related to variable selection are difficult and largely unexplored. A better understanding of how models are sensitive to variable specification would greatly contribute to our understanding of extrapolation.

There is one part of the dissertation that does make a contribution to the study of variable selection in causal models. Mediation models enable one to evaluate the relationship between the total effect going through all causal paths and component effects going through particular paths. In the recidivism example, it is possible to evaluate the effect of employment on recidivism that is mediated by cash-in-hand and compare it to the total effect of employment on recidivism going through all paths. The total effect can be given in a model with just two variables (*employment* and *recidivism*). The indirect effect through cash-in-hand is given in a model containing at least three variables. Thus, mediation techniques allow one to draw a connection between models with distinct numbers of variables.

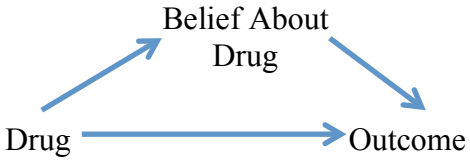

Figure 2

Mediation techniques also enable one to distinguish between causally relevant and irrelevant properties of a treatment. Consider the question of whether a drug works through a chemical pathway or only as a placebo. Knowing that a drug is effective in bringing about an outcome is not enough to establish that it brought about the outcome as a result of its chemical composition. It could be that the drug worked primarily as a result of making the patient believe that they have been treated. Through mediation techniques, one can hold one's belief about the drug fixed in order to determine whether the drug is effective through non-psychological pathways. Suppose that the effect is entirely due to the indirect path (figure 2). This would reveal that the drug works in virtue of its psychological properties, rather than in virtue of its chemical properties. This is a further way in which mediation techniques contribute to variable selection and the question of how to distinguish between different candidate properties.

The inductive approach is closely related to the natural kinds approach. At least since Goodman, it has been widely accepted that whether one can infer that future *F*'s will be *G*'s based on the prior observation of *F*'s that are *G*'s depends on the nature of the predicates *F* and *G.* To use Goodman's terminology, certain predicates are "projectable", which means that observations of *F* that are *G's* confirms the hypothesis that unobserved *F'*s are also *G*'s. So headway on either the natural kinds or the inductive approach would contribute to the other.

The central challenge to developing an inductive approach to extrapolation is that it is unclear how causal inference relates to inductive/statistical inference. As already emphasized, it is standard in causal inference to assume that one knows the probability distribution for the variables in a population, and thus to bypass statistical questions regarding how one knows that two variables are correlated in general based on one's finite sample. As I noted above, in cases of

extrapolation, it is conceptually difficult to disambiguate between the problem of not knowing the probability distribution for the study population and the problem of knowing the probability distribution of the study population, but not knowing if the probability distribution for the target population is different. We will return to this question at the end of chapter 4.

Chapters Summary

Now that I have introduced the topic, I will briefly present the organizational structure of the dissertation.

Chapters 2-4 review the prior literature on extrapolation. Chapter 2 examines Nancy Cartwright and Jeremy Hardie's account of extrapolation in *Evidence-Based Policy: A Practical Guide to Doing it Better.* Their account is a combination of (what I have called) the minimal sufficient conditions and the natural kinds approaches. Chapter 3 evaluates Daniel Steel's account in *Across the Boundaries: Extrapolation in Biology and the Social Sciences*. Steel refers to his approach as the "mechanisms approach", which I refer to as the "mediation approach". Chapter 4 considers Judea Pearl and Elias Bareinboim's *transportability* approach to extrapolation, which is the most sophisticated approach to date. The transportability approach does not fall neatly into any of the four categories I have presented. It does enable one to make headway on the mediation approach and it presents an opportunity to revisit some important questions about the relationship between extrapolation, causal inference and probabilities. After chapter 4 there is a short section that takes stock of which questions have been answered in the prior literature and which remain open.

Chapters 5-7 develop the mediation/mechanisms approach to extrapolation. Chapter 5 discusses the philosophical literature on mechanistic explanation. I dispute the claim that

mechanistic phenomena call for a non-causal form of explanation by showing that the features of mechanisms that allegedly elude causal explanation can be modeled using mediation techniques. Chapter 6 provides a more rigorous introduction to causal mediation techniques and considers one way in which they are useful for extrapolation. Namely, knowing the relative contributions of distinct causal paths enables one to measure the maximum effectiveness of a policy that seeks to disrupt one of the paths. Chapter 7 discusses the relationship between causal mediation techniques and transportability and argues that mediation techniques track a form of cross-population invariance that cannot be represented within the transportability framework. It then advances further proposals for how mediation techniques facilitate extrapolation.

Chapter 2: Cartwright and Hardie on Extrapolation

The first account of extrapolation I will consider is that of Nancy Cartwright and Jeremy Hardie in *Evidence-Based Policy: A Practical Guide to Doing it Better*. This book is primarily addressed to policymakers. Policymakers are increasingly privileging randomized control trials (RCTs) as the best evidence for causal claims. In an RCT a researcher randomly assigns subjects into treatment and control groups. If the randomization is successful, then the difference in outcomes between the two groups provides an unbiased estimate of the average causal effect of the treatment on the outcome for the population in the study. Yet, positive RCTs only show that a causal relation obtains in the study's population. Cartwright and Hardie present assumptions that enable one to extrapolate to target populations. Their aim is to encourage policymakers to think about whether these assumptions are met in particular cases and to avoid extrapolating in cases where the assumptions fail.

In the previous chapter, I identified four general approaches to extrapolation: minimal sufficient condition approaches, natural kinds approaches, inductive approaches and mediation approaches. We can see elements of all four approaches in Cartwright and Hardie's account. Their primary account is a combination of a minimal sufficient conditions approach and a natural kinds approach. Additionally, by arguing that knowing how a cause brings about its effect facilitates extrapolation, they point to the need for a mediation approach. Finally, they provide an account of what counts as evidence for an extrapolation. I argue that this account fails, but it nevertheless constitutes an attempt to develop an inductive approach.

The approach that Cartwright and Hardie develop most extensively is the minimal sufficient conditions approach. They instruct policymakers to consider whether the background factors that are necessary for a cause to bring about its effect are present in the target population.

While this approach is helpful for avoiding bad extrapolations in cases where a necessary condition for the effect is absent, the authors have less to say about when one *can* extrapolate. Of course, if one knew that all of the factors that constitute a minimal sufficient condition were present in the target population, one could extrapolate. But one does not generally have this knowledge, and the authors are silent regarding what one should do if one knows only *some* of the relevant factors.

This chapter introduces one element of the DAG framework for causal modeling. Cartwright and Hardie represent causal relationships between a variable and its causes using what they call "causal principles". These causal principles are identical to what I (and others) elsewhere refer to as *structural equations.* Within the DAG framework, the value of each variable is determined by structural equation representing that variable as a function of its direct causes in the graph (more on this in chapter 3). Cartwright and Hardie do not use DAGs, but my discussion of causal principles will elucidate some important features of structural equations. Notably, whether the causal relationship represented by a structural equation generalizes is sensitive to how one specifies the variables in the equation.

This chapter is organized as follows. Section 1 presents Cartwright and Hardie's effectiveness argument, which is a deductive argument for the conclusion that a policy will work in the target population. I show that establishing the premises of this argument would require both minimal sufficient condition and natural kinds approaches. Section 2 argues that the authors develop only the first of these approaches and that their account only licenses extrapolations in a limited set of cases. Section 3 presents one way to develop their idea that knowing how a cause brings about its effect aids extrapolation. Section 4 criticizes the authors' account of evidence. Section 5 concludes.

1. The Effectiveness Argument

Cartwright and Hardie model extrapolative inferences as having the form of a deductive argument, which they call *the effectiveness argument* (45). The conclusion of the argument is that a particular factor that had a positive causal effect in the study population will have a positive causal effect in at least some members of the target population. This is a weak conclusion that is compatible with the policy having a net negative effect. Although it is not sufficient for establishing that a policy is effective in a population, it is necessary.

The effectiveness argument contains three premises. Premise 1 is that a factor, *X*, has a positive effect on an outcome in one population. This is what an ideal RCT establishes.[3] It is a mistake to infer from the first premise that *X* will have a similar effect in other populations; two additional premises are required. Premise 2, which requires further elaboration, is that *X can* play a similar causal role in the intended population. Premise 3 states that the *support factors* necessary for *X* playing this role are present in the target population. Support factors for *X* are other factors required for *X* to have its effect.

Premises 2 and 3 block two ways that a causal claim can fail to generalize from one population to another. To illustrate, a study in Tamil Nadu established that educating mothers promoted healthier infants. Unfortunately, a similar intervention in Bangladesh failed to improve infant health. Why? The authors suggest that what explains the difference is that in Bangladesh mother-in-laws (rather than mothers) are in charge of distributing the food in the family. Premise

[3] Some have criticized RCTs on the grounds that we have no assurance that the populations will be even approximately balanced in studies with small samples (Worrall (2007); See Reiss (2013), chapter 11 for discussion). Cartwright and Hardie purposely put this issue to the side. They assume that RCT are valid for the test population and ask whether their results can be generalized to other populations.

2 does not obtain, since educating mothers plays a different causal role in Bangladesh than in Tamil Nadu. Educating mother-in-laws, in contrast, would play a similar causal role.

Even if educating mothers did play a similar causal role in Bangladesh, the intervention could fail if certain support factors were absent. Educating the mother might have no impact on infant health if the family lacks an adequate food supply. Causes do not typically work in a vacuum, but rather require other factors to bring about an effect. Borrowing J.L Mackie's terminology, causes are *INUS conditions.* An INUS condition is an *I*nsufficient but *N*ecessary part of an *U*nnecessary but *S*ufficient condition for an effect. In other words, when *X* is an INUS condition for *Y*, *Y* obtains if and only if *BX* v *Z* is true, where *BX* is a minimal sufficient condition for *Y*, and *Z* is a disjunction of other minimal sufficient conditions for *Y*. Within this framework, one can easily see that *B* is a support factor for *X*, since only in conjunction with *B* does *X* bring about *Y*. The authors, like those concerned to identify causes, pick out one factor (*X*) as *the* cause, but there is no non-pragmatic distinction between causes and support factors. When *X*'s support factors are not present, premise 3 does not obtain and the policy may not have its intended effect.

Although premises 2 and 3 in the effective argument are intuitively distinct, one must refer to what the authors call *causal principles* to make this distinction precise. Here is the causal principle for Tamil Nadu[4]:

$$\text{(TN) } I = a_1 + a_2 I_0 + a_3 B_m E_m + a_4 Z$$

The lowercase '*a*'s are coefficients and the uppercase letters are random variables - *I* refers to infant health, I_0 is infant health at an earlier time, E_m is education of the mother, B_m are the support factors for E_m, and *Z* represents all other causes of *I* that do not interact with $B_m E_m$. The

[4] I have altered the notation of the causal principles in several ways to improve clarity. All of the coefficients are adjustable parameters, so a_1 in one principle need not have the same value as a_1 in another.

equation represents how the infant health would change if one were to intervene on one of the right-hand-side variables while holding the others constant.[5] The difference between a failure of premise 3 and a failure of premise 2 is as follows. Premise 3 is false if the value of B_m differs in the two populations. Premise 2 is false if the variable E_m does not appear in the causal principle for one of the populations. According to the authors, the educational intervention failed in Bangladesh because premise 2 was false. Bangladesh has the following causal principle:

$$\text{(BD) } I = a_1 + a_2I_0 + a_3B_{ml}E_{ml} + a_4Z$$

Where E_{ml} refers to the education of the mother-in-law. Since (BD) does not contain a variable for E_m, premise 2 does not obtain. But what determines whether E_m appears in Bangladesh's causal principle?

Another way to ask this question is to ask why there need to be *two* causal principles (one for each population). Consider the following combined causal principle, which applies to both populations:

$$\text{(C) } I= a_1 + a_2I_0 + a_3B_mE_m + a_4B_{ml}E_{ml} + a_5Z$$

(C) contains both E_m and E_{ml}, so premise 2 is satisfied. Since the values of the support factors can differ between the populations, the effects of E_m and E_{ml} can differ as well (as, in fact, they do). If one represents Bangladesh using (BD), premise 2 does not obtain, but if one represents it as (C), it does. Absent some reason for choosing (BD) over (C), the truth of premise 2 will be objectionably language dependent.

One reason to prefer (BD) to (C) is that if one models the difference between the populations with (C), one misses the fact that the policy's success depends not on which particular member of a family one educates, but rather on whether one educates the person with

[5] Chapter 3 further explains what it means to 'intervene' on a variable.

power over the family's food distribution. At one point the authors suggest that for each population, the relevant causal principle should look as follows:

$$(P)\ I = a_1 + a_2I_0 + a_3B_{pw}E_{pw} + a_4Z$$

The subscript *pw* means "person with the power". I'd like to suggest that instead of considering (P) as an alternative to the distinct principles for each population ((TN) and (BP)), we should rather think of it as an alternative to (C). Like (C), (P) applies to both populations, but only (P) captures the common causal role played by the variables E_m and E_{ml} in (C).

Cartwright and Hardie talk as if one can determine whether premise 2 obtains by considering whether a factor appears in a population's causal principle, but populations do not wear causal principles on their sleeves. A population can have one causal principle relative to one set of measured variables, and a different principle relative to another set. The insight behind premise 2 is that choosing one variable set over another can aid extrapolation. This insight has been neglected in the literature on causation. In order to make this point, however, one needs to separate the cases in which one compares two populations using a single model from those in which one compares two ways of modeling the same population. Premise 3 concerns the way that two populations could differ relative to a single way of specifying the variables. Premise 2 concerns the question of whether the factor under consideration would be a variable in the optimal model.[6]

The insight that whether a causal relation generalizes depends on how one specifies the variables corresponds to what I referred to in chapter 1 as the *natural kinds* approach to extrapolation. In Tamil Nadu, the variables for the education of the person in power and the

[6] More must be said about how to choose among competing causal models. The question raised here regarding whether one should use two population-specific causal variables or a single causal variable for both is related to the question of why models with fewer adjustable parameters are preferable to those with more (Forster (2007), Forster and Sober (1994), Whewell (1840)).

variable for the mother's education both refer to the education of the same individual. Yet, which variable one chooses makes a difference for whether the causal relationship generalizes to Bangladesh. To establish premise 2, one would need to provide an account of how to choose between alternate specifications of a variable.

While establishing premise 2 involves providing a natural kind approach, establishing premise 3 requires a minimal sufficient conditions approach. Premise 3 says that all of the support factors necessary for *X* bringing about its effect are present. The support factors for *X* are the factors that combine with *X* to form a minimal sufficient condition.

3. How Useful is the Account?

Cartwright and Hardie suggest that a policymaker should perform two searches – a *horizontal search* and a *vertical search* – prior to implementing a policy. These searches correspond to premises 3 and 2, respectively.[7] In a horizontal search, one considers whether the support factors in the study population obtain in the target population as well. In a vertical search, one thinks about whether one has described the cause at the right level of description.

How useful are these searches for determining whether a policy will succeed? Cartwright and Hardie describe an intervention to improve reading scores by means of reducing class size that was successful in Tennessee, but failed in California. A horizontal search would have revealed that California was missing support factors that were present in Tennessee. Specifically, unlike Tennessee, California had a shortage of both teachers and classroom space. In cases like this, where one knows some of the necessary conditions for a policy to work, horizontal searches are clearly useful. In situations where both populations appear to have the

[7] The authors do not explicitly note the correspondence between premise 3 and a horizontal search and between premise 2 and a vertical search.

conditions strictly necessary for bringing about the effect, horizontal searches are less useful. Would it have been worth performing the intervention had California had enough teachers to implement it, but fewer teachers-per-student than in Tennessee? All else being equal, one would guess that this would reduce the efficacy of the intervention, but all else is never equal. Perhaps the teachers in California are better on average and this compensates for the negative effects of the higher student-to-teacher ratio. Alternatively, maybe good teachers can only do so much if the classes are too big. One rarely knows what all of the support factors are, and even if one did, this knowledge would be insufficient for determining how varying these factors changes the effect. For this reason, horizontal searches are better suited for ruling out policies in which support factors are absent than for justifying policies when they are present.

The limitation just described regarding horizontal searches corresponds to a more general limitation of minimal sufficient condition approaches. Such approaches are useful when one is either able to find a full set of factors constituting a minimal sufficient condition or one is able to find particular factors that are necessary for *X* to bring about its effect. They do not appear to be useful in other cases. If one only knows some of the support factors for *X*, and these factors are not sufficient, the minimal sufficient conditions approach provides no guidance.

We've already seen an example of a vertical search in the Tamil Nadu case. The principle "educate the person in power" extrapolates to Bangladesh; "educate the mother" does not. The level of abstraction at which we describe a causal factor is important. How can we translate this insight into practical advice? By abstracting away from the properties of a population we end up with claims that apply to a wider range of populations, but not all ways of abstracting work equally well. In the Tamil Nadu case, switching from "educate the mother" to the more general "educate the person in power" worked, but why should we abstract to *this* general principle. Why

not "educate the person who supervises the child" (supposing that mothers play this role in Tamil Nadu)? This principle is as abstract as the one they suggest and it yields different advice for applying the lessons from Tamil Nadu to Bangladesh. How can one know which principle to adopt just by looking at Tamil Nadu? Without some guidance regarding which ways of abstracting are preferable, vertical searches do not yield a verdict on whether a causal relation extrapolates to the target population. Cartwright and Hardie identify this need, but they do not provide guidance concerning how to satisfy it.

4. Mechanisms and Mediation

Horizontal and vertical searches enable a policymaker to use her background knowledge in considering whether a policy will work. The authors say little about how to determine if one has reliable background knowledge in the first place. Consider the case they discuss of a nurse who is able to quickly detect whether an infant has a certain disease (131-2). Since this disease is treatable only if it is detected early, the hospital would like to teach the nurse's skill to other nurses. Through careful deliberation, the nurse discovered that she detects the disease through monitoring whether the infant changes color, shows heightened activity, and has reduced appetite. Assuming that the nurse is correct about how she makes her diagnoses, it will be possible to teach the other nurses how to make similar diagnoses by looking for these changes. In this case, the nurse was in fact correct, and the hospital was able to teach other nurses to make better predictions. Yet, even though the nurse's judgment was reliable, there is little reason to think that people's causal judgments are generally reliable, especially when one is implementing a complicated policy. This is why we need RCTs in the first place. It would therefore be unsatisfactory if extrapolation relied entirely on causal intuitions.

Fortunately, some of the assumptions that license an extrapolation are testable. Consider the following model for the hospital case (figure 1):

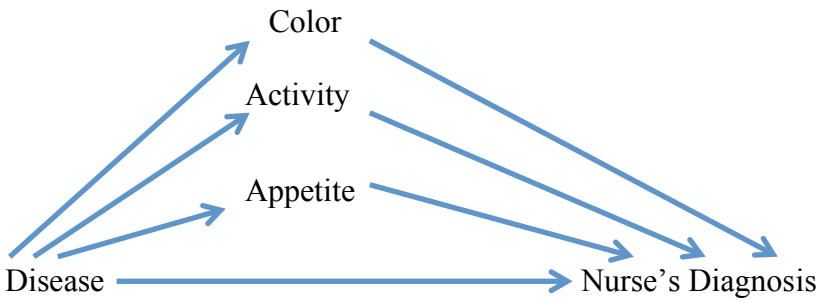

Figure 1

This model represents the possible causal relations between the variables. It includes three measured variables on the path from the disease to the diagnosis. These measured variables are called *mediators*. The arrow going directly from the disease to the diagnosis represents all the causal paths between the treatment and the outcome that do not go through the measured mediators. Using *causal mediation* techniques, one can determine how much each path contributes to the total effect. Doing so requires more complicated experimental designs than standard RCTs (Imai et al. 2011). Initially, one might think that one could measure the influence on a path going through a mediator by randomizing the mediator. The reason this does not work is that when one randomizes the mediator, one severs the causal connection from the treatment to the mediator. Randomizing the mediator enables one to estimate the effect of the mediator on the outcome, but this is not the quantity one wants to estimate in causal mediation. The desired quantity is the causal contribution of the path going from the treatment to the mediator to the outcome, but randomization disrupts this path. Despite this complication that arises in measuring the relative contributions of the different paths, they are in principle measureable (Pearl 2012) and social scientists have developed preliminary experimental designs for measuring them (Imai

et al., 2013). The nurse's hypothesis about how she makes her predictions can be verified by measuring the contributions of the paths going through the mediators.

Causal mediation techniques aid in extrapolation, since testing a hypothesis about the way a cause operates in the study population often enables one to predict whether it will work in other populations. If the nurse's predictions were largely based on infant color, then other people capable of detecting these color changes would probably make similarly good predictions.

A central thesis of *Evidence-Based Policy* is that knowing *how* a cause works (which requires more than knowing the support factors and the causally relevant description) is essential to knowing whether it will generalize. This is the central idea behind what I called the mediation approach. The authors say little about how we can learn what we need to know. Causal mediation techniques help answer this question. In the later chapters of this dissertation, I will provide a more precise account of how mediation techniques facilitate extrapolation.

5. Extrapolation and Induction

The effectiveness argument contains a set of assumptions that, if true, would justify an extrapolation. In addition to presenting these assumptions, the authors also give an account of evidence for when an extrapolation is justified. According to this account (19), any evidence *e* for a premise in the effectiveness argument is also evidence for the conclusion of the argument. This account is untenable, since evidential relevance is not, in general, transitive; just because *e* is evidence for a premise that is evidence (relative to an argument) for a conclusion does not entail that *e* is evidence for that conclusion (Hesse, 1970). That a card is red is evidence that it is the queen of hearts, which entails that it is a queen. But that a card is red is not evidence that it is

a queen. Fortunately, none of their claims about extrapolation depends on this theory of evidence.

While the authors' theory of evidence does not make a difference for their subsequent claims about extrapolation, their claim to be providing such a theory matters rhetorically. The authors want policymakers to abandon the view that RCTs count as the only evidence for causal claims. Their rhetorical strategy is to claim that they are not advocating for a weakening of the standards for establishing a causal claim, but rather seeking evidence that establishes a different type of claim (i.e. a claim about the effect in other populations). It is therefore significant that their account for what counts as evidence for an extrapolation fails.

The intransitivity of evidential relevance is one reason why it is difficult to develop an inductive approach to extrapolation. All of the accounts I evaluate in the dissertation are deductive; they provide conditions under which one's assumptions entail some fact about the magnitude of the effect in the target population. If evidential relevance were transitive, then one could easily transform these accounts into inductive accounts, since any evidence for the extrapolation-licensing assumptions discussed by these accounts would be evidence for extrapolation. The failure of the transitivity of evidential support blocks any simple way to use these accounts to make claims about evidence, confirmation, or induction.

6. Conclusion

Cartwright and Hardie provide a helpful starting point for the subsequent discussion. Through evaluating their account, I have motivated the four approaches to extrapolation I discussed in chapter 1 and noted some of the difficulties that arise in pursuing them. I will not further pursue the minimal sufficient conditions account, since I cannot see a way to avoid the limitations I

present here. In the rest of the dissertation, I develop a mediation approach. Cartwright and Hardie suggest that knowing how a cause brings about its effect facilitates extrapolation, but there remain many open questions regarding both how we learn that a cause brings about its effect and how such knowledge enables us to predict the magnitude of the effect in the target population. In the following chapter, I turn to Daniel Steel's account, which provides the most extended philosophical discussion of these questions to date.

Chapter 3: Steel's Mechanism's Account

In *Across the Boundaries: Extrapolation in Biology and Social Science*, Daniel Steel provides an account of how extrapolations can be justified in particular sciences. He provides a condition that licenses some extrapolations in biology, though he is less sanguine about the prospect of developing a similarly successful account for certain social sciences (his most developed examples come from anthropology). There are several features of his account that are similar to the one I will present. He relies on the causal modeling techniques developed by Pearl (2009), Spirtes, Glymour and Scheines (2000) (henceforth SGS) and Woodward (2003). He also emphasizes the importance of considering variables on a causal path from a treatment to an outcome, as I will when I consider causal mediation techniques. Given the similarity both in Steel's project and his approach to dealing with extrapolation, an analysis of his account will reveal both which problems have been solved and which ones require further inquiry.

In several respects, Steel's positive account of extrapolation anticipates the methods of Pearl and Bareinboim, which are the subject of the next chapter. Pearl and Bareinboim's account is both more precise and more general than Steel's. Nevertheless, Steel is unique in presenting not just an account of when one can extrapolate, but an extended discussion of the inferential challenges that are particular to extrapolation and of the types of approaches that are capable of addressing these challenges. He argues that any account of extrapolation must address a problem he calls 'the extrapolator's circle'. Addressing this problem involves showing how it is possible to know about a causal relationship in a target population without learning so much about that population as to render extrapolation from the study population unnecessary. He rejects solutions to this problem that rely on 'simple induction' – assuming that the effect in the target will be the same as that in the study population. Instead, he pursues a *mechanisms-based approach* to

extrapolation, on which learning *how* a cause brings about its effect helps one determine whether the causal relationship between these variables generalizes across populations. This idea is also at the center of my own approach, although I refer to it as a mediation approach.

In this chapter, I present three criticisms of Steel's account. First, I argue that Steel's approach to extrapolation is not substantially different from simple induction. Instead of providing methods for determining when one can assume that two populations are similar, his account only shows how to use presumed causal similarities between populations in order to extrapolate. Second, Steel does not explain how it is possible to extrapolate in cases where two populations differ causally, despite his claims to the contrary. He claims that one can extrapolate across populations provided that one extrapolates qualitative – as opposed to quantitative – causal claims, but the quantitative/qualitative distinction is irrelevant to the epistemic question of how one can extrapolate an effect across populations that differ in their background factors. Third, Steel's main theorem for extrapolating positive causal relevance claims does not actually license extrapolations, and it is unrelated to the other elements of his account.

This chapter is organized as follows. Section 1 introduces the causal modeling techniques that Steel utilizes. Section 2 evaluates Steel's solution to the "extrapolator's circle". Section 3 criticizes Steel's attempt to show how to extrapolate across causally heterogeneous populations by limiting extrapolation to qualitative causal claims. Section 4 critically evaluates Steel's "mechanisms-based approach" to extrapolation. Section 5 concludes.

1. Structural Causal Models and Interventionist Accounts of Causation

Steel dubs his account the 'mechanisms-based' approach to extrapolation. Mechanisms are defined as "entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (Machamer, Darden, Craver, 2000, 3).

The mechanisms-based approach to extrapolation weds this notion of a mechanism to Woodward's interventionist account of causation. This account – which builds on the work of Pearl (2009) and Spirtes, Glymour and Scheines (2000) – models the causal relations between variables using directed acyclic graphs (DAGs). Graphs consist of nodes and edges. In a DAG, the nodes are variables and the edges are arrows. The graph is directed, since the arrows represent asymmetric causal relationships. It is also acyclic: one cannot get from a variable back to itself via a connected set of arrows.

In a DAG, there is an arrow (i.e a directed edge) between two variables iff it is possible to change the value of variable to which the arrow points through an *ideal intervention* on the variable before it (see below). The variable at the tail of an arrow is referred to as the *parent* of the variable coming after it. DAGs correspond to sets of *structural equations* in which each variable is a function of its parents (and, typically, an error term representing omitted causes of that variable). A DAG combined with a corresponding set of equations and the distributions of the variables to which no arrows are pointed determines the probability distribution for the variables in the graph. One can think of the DAG as representing the physical process that generates this probability distribution. Steel refers to such processes as *causal structures.* A central thesis in the book is that within the domain of biology, the causal structures that generate probability distributions just *are* biological mechanisms. I will say more about this thesis shortly.

A virtue of Woodward's interventionist account of causation is how easily it distinguishes between correlations that reflect direct causal influences of one variable on another and those that are merely the result of a common cause of the two variables. Consider the familiar example of a barometer. Here's the DAG for the variables of atmospheric pressure (*A*), the barometer reading (*B*) and whether there is a storm (*S*):

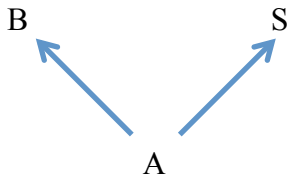

Figure 1

The variables *B* and *S* are correlated due to having *A* as a common cause. Yet *B* is not a cause of *S* (or vice versa). For the interventionist, the fact that *B* does not cause *S* is reflected in the fact that one cannot influence the probability of there being a storm by intervening on the barometer reading. To intervene on *B* is change its value in such a way that it no longer depends on its prior causes. An intervention on *B* in this system can be represented as follows:

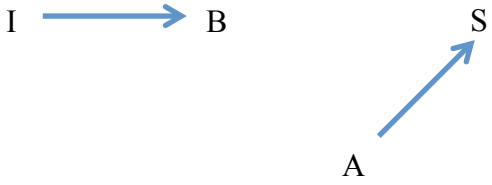

Figure 2

In figure 2, *I* represents a mechanism through which one intervenes on the barometric reading. If one sets the barometer reading based on the outcome of a coin toss, the reading will now depend on a variable that is both probabilistically and causally independent of the atmospheric pressure and any other causes. Note that when you intervene on *B*, this "breaks" any arrows going into *B*, since the value of *B* is entirely determined by the intervention, leaving no room for other influences. *B* is not a direct cause of *S* since there is no possible intervention on *B* through which one can change the value of *S*. More generally, *C* is a *direct cause* of *E* in variable set ***V*** iff there

is at least one intervention on *C* that changes the value of *E*, while intervening to hold all other variables in ***V*** fixed.

We need to make the notion of an intervention more precise. Colloquially, to intervene on a variable is merely to change it, but in this loose sense it is not the case that there is no way to change *S* by changing *B*. If the way that one changes *B* is by changing *A*, doing so could also change *S*. The interventions that count for distinguishing causation from mere correlation are *ideal interventions,* which Steel, following Woodward, defines as follows:

> [*Ideal Intervention*]: *I* is an ideal intervention on X ∈ **V** if and only if it is a direct cause of X that satisfies these three conditions:
> (a) *I* eliminates other influences upon X but otherwise does not alter the causal relations among **V**.
> (b) *I* is a direct cause of no variable in **V** other than X.
> (c) *I* is exogenous. (Steel, 2008, 13-14)

I is exogenous iff it is not an effect of any variable in **V** and does not share a common cause (either latent or measured) with any variable in **V**. In the barometer example, *I* fulfills (a) and (b) because it eliminates all of the arrows going into *B* without influencing any other variables and it is exogenous, since it has no (included)[8] causes. It should be clear that this definition of an ideal intervention could not be used to provide a non-reductive explication of "direct cause" that does not rely on causal concepts. Evaluating conditions (a) – (c) requires one to already have causal concepts. Nevertheless, interpreting the arrows in DAGs as direct causes in Woodward's sense turns DAGs into powerful tools for modeling how a probability distribution will change as a result of interventions.

---

[8] DAGs need not include all causes of a variable – unmeasured causes are typically omitted from the graph – but in order for it to provide a reliable guide to the probabilistic independencies that obtain in a population one must include any variable that is a common cause of two variables in the DAG. The requirement that one include all common causes is known as *causal sufficiency.* Steel claims that his account does not require assuming causal sufficiency (13 fn. 1), but since he takes the DAGs he presents to have implications for which variables will be uncorrelated it appears that it does. Moreover, he assumes the causal Markov condition, which is known to fail for variable sets with omitted common causes. Assuming causal sufficiency, any node in a DAG that has no explicitly represented causes is exogenous.

One can relate DAGs to probability distributions using the *Causal Markov Condition* (CMC). Let's call any variable *Y* that is a downstream effect of *X* – that is, *X* is either a parent of *Y*, or a parent of a parent of *Y* etc. – a *descendant* of *X*. The CMC states that every variable in a DAG is uncorrelated with all of its non-descendants conditional on its parents. For example, consider two variables *A* and *B* that share a complete common cause *C* – that is, *C* is the only parent shared by *A* and *B*. If there is no direct causal relationship between *A* and *B*, then conditional on *C*, *A* and *B* will be uncorrelated. A useful consequence of the CMC is that if one has not conditioned on any variables, then any two variables that are correlated must be causally related either directly (one causes the other), indirectly along a path, or via a common cause. This is known as the *Principle of the Common Cause,* which was originally suggested by Reichenbach (1956). Part of what the CMC captures is that causal chains are "memoryless". To know the probability of a variable taking on a certain value, it is sufficient to know the values of its parents and learning about the values its "grandparents" (parents of parents) provides no further information.

The CMC specifies for a given DAG which variables must be *uncorrelated*, but does not entail anything about which variables will be *correlated.* It is common scientific practice, however, to infer from the fact that two variables are uncorrelated that they are not causally related. This inference presupposes that if two variables *are* causally related, then they *will* be correlated. In the causal modeling literature, this assumption is made precise in the form of the *Causal Faithfulness Condition* (CFC). CFC states that the *only* variables in a DAG that will be uncorrelated are those that the CMC entails will be uncorrelated. All other variables will be correlated. While the CMC entails that all correlated variables will be causally connected, the CFC entails that all causally connected variables will be correlated. The CFC is violated

whenever two causal paths between a cause and effect cancel out exactly, so it clearly is not universally true. It may nevertheless be justified as a defeasible rule of scientific inference.

CMC and CFC are common assumptions in the causal modeling literature. Together, they entail that two variables are probabilistically related iff they are causally related.[9] Steel provides interesting defenses of these conditions, though I will not go into the details here. Note that CMC and CFC do not have anything specifically to do with extrapolation. They are standard assumptions made whenever one relates a probability distribution to a DAG.

Steel takes his account to rely on both the interventionist approach of causation and the recent literature on mechanistic explanation. He argues that in biology the causal structures that generate probability distributions are biological mechanisms. His argument does not involve showing that any particular biological mechanism can be causally modeled, but rather works by claiming that biological mechanisms have the types of properties that we expect to give rise to probability distributions that can be causally modeled. Here I will not examine the details of this argument, since as far as I can tell nothing in his account depends on the identification of causal structures with mechanisms. When Steel talks of particular mechanisms he is referring to particular *causal paths* between variables, where a causal path between two variables is a set of connected edges all pointing in the same direction. There can be multiple causal paths between two variables.

Of course, by stipulating that mechanisms are causal paths we lose the ability to relate Steel's account to the broader literature on mechanistic explanation. Later on in the dissertation

---

[9] Here I've avoided providing a more technical definition of what it is for two variables to be causally related in a graph. To do so requires further terminology. Two variables *X* and *Y* are *adjacent* in a graph iff *X* is a direct cause of *Y* or *Y* is a direct cause of *X* in the graph. A *path* from *X* to *Y* is a series of adjacent variable connected by edges going in any direction (when I want to indicate that all of the edges in a path go in the *same* direction, I will refer to it as a *causal path*). *Y* is a collider on the path $X - Y - Z$ iff $X \rightarrow Y \leftarrow Z$. Roughly, two variables *C* and *E* are causally related iff they are connected by at least one path on which there is no unconditioned collider (i.e. no collider upon which one has not probabilistically conditioned).

(Chapter 5), I will come back to the question of whether this literature has anything to contribute to the problem of extrapolation. We don't need to resolve this question here to evaluate Steel's account.

2. The Extrapolator's Circle

Steel argues that any account of extrapolation must be able to address what he refers to as *the extrapolator's circle*: extrapolation requires the assumption that the study and target populations are similar, but to justify this similarity one must have knowledge of the target population. The concern is that this circle is vicious, since the knowledge of the target required for justifying the extrapolation is precisely what one wants to learn from the extrapolation. Steel rejects theories of extrapolation that are unable to break the circle as inadequate. For example, some scholars have suggested that the key to extrapolation is discovering the mechanism underlying a causal relation in the study population (e.g. Wimsatt, 1998). While Steel is obviously sympathetic to this suggestion, he is notes that simply appealing to mechanisms does nothing to break the extrapolator's circle. To break the circle, one would need to explain how one could learn that the mechanism in the target population resembles that in the study population without learning so much about the target as to render extrapolation unnecessary.

One way to break the circle is *simple induction*: assume that the magnitude of the causal relationship in the target population will be the same as that in the study population. Steel acknowledges that simple induction is sometimes useful, but argues that it is too limited to provide a general strategy for extrapolating. The problem with simple induction is that we typically extrapolate in cases where we expect there to be some differences between the model and the target, so simple induction will lead one to make a fallacious inference. For example, suppose that 15% of Americans cannot stand cilantro – if they eat any amount of it, they will feel

nauseous. In contrast, everyone born in Japan can eat cilantro without negative consequences. The argument against simple induction looks as follows. If we measure the effect of cilantro consumption on nausea among Americans, we will find that there is a causal effect. To apply simple induction would be to assume that there will be a similar effect among Japanese people, but this would lead to a false conclusion. We therefore should not rely upon simple induction.

Steel's dismissal of simple induction is overly hasty. Whether simple induction works depends on which effect one measures. Recent research has identified the gene that causes cilantro aversion. Suppose that there is only a single such gene and all and only people who have the gene hate cilantro. Instead of simply measuring the effect of cilantro on nausea in Americans, one might instead measure the effect in two groups of Americans: those with the gene and those without it. One might discover that in the former group consuming cilantro always causes nausea and in the latter group it never does. To apply simple induction to this case would be to assume that cilantro will cause nausea in all and only Japanese people who have the gene. Given the assumptions stated here this would be the correct result. Cilantro is not a cause of nausea in Japan, since people in Japan do not have the gene.

Of course, none of this saves simple induction in its unrestricted form. It is a bad policy to always assume that some arbitrarily specified causal effect generalizes to a target population. My point here is that even at this stage in the discussion, we can identify two distinct ways one might model extrapolation. For Steel, we have a fixed causal relationship C that by hypothesis may differ between two populations and the challenge is that of determining when it does. An alternative is to assume that there are some causal relationships that are the *same* across the populations. Accordingly, the problem is that of finding out which relationships these are.

Steel's own response to the extrapolator's circle relies on what he calls "comparative process tracing". In comparative process tracing, one determines whether a causal relationship between *C* and *E* in the study population will obtain in the target population by comparing the mechanism in the target population to that in the study population. According to Steel, this method counts as a solution to the extrapolator's circle, since one does not need to compare the mechanisms in their entirety, but rather only the parts of the mechanism that are likely to differ across the populations.

Steel explains comparative process tracing by reference to a hypothetical mechanism with the following structure:

Figure 3

Figure 3 represents a causal path from *C* to *E*. Comparative process tracing is applicable when one has fully examined the entire mechanism from C to E in one population (usually a laboratory population, such as mice, which is referred to as the *model*) and one wants to make inferences about the corresponding mechanism in a target population where one has less information about that mechanism. There are two ways that comparative process tracing aids one in this inference. First, if one believes that there are parts of the mechanisms at which the model and the target are unlikely to differ causally, one does not need to examine those mechanisms in the target population. Second, sometimes it will be possible to make inferences about whether earlier stages of the mechanism differ based on whether later stages of the mechanism do. For example, suppose that one suspects that two populations differ in the causal relationship between *Y* and *A*, but is unable to investigate this relationship in the target population. Since a change in the effect

of *Y* on *A* would be reflected in a change in the value of *Z* in the populations, one can verify whether the effect of *Y* on *A* varies between the populations by looking at *Z.*

Let's start with the second proposal first. As Steel notes, in order for this to work, there cannot be any causal path from *Y* to *Z* not going through *A*. What Steel does not note is that this inference requires that the two populations not differ in background factors that make a difference in the effect of *A* on *Z.* If there were, then it could be the case that the populations differ in *Z* even though they do not differ in the effect of *Y* on *A.* The second proposal resembles the first in requiring knowledge regarding where the populations are likely to differ. More importantly, it requires assumptions that they do *not* differ.

At first glance, the first proposal that in order to extrapolate using comparative process tracing one must have a theory about the probability that two mechanisms will differ at particular points is not much of a step forward. Without a story regarding how we could learn about the likely differences between mechanisms, it merely pushes the relevant question back a step. This concern is somewhat allayed once one notices that in the biological sciences under consideration, scientists do typically have a lot of data concerning phenotypic differences among organisms. It is a virtue of Steel's account that he provides real life biological examples in which comparative process tracing enabled extrapolation.[10]

Even granting that there is some empirical basis for making claims about where the relevant differences are between model and target mechanisms, Steel still owes us an account of *how* our empirical evidence supports such claims. The evidence obviously cannot be sufficient for establishing the relevant causal claims in the target population, since otherwise there would

[10] Julian Reiss (2010) raises some doubts regarding whether Steel's examples of extrapolations in biology are historically accurate.

be no extrapolation problem. So why think that the evidence we have gathered from, say, toxicology, enables us to draw reliable conclusions about the target mechanism? Steel writes:

> Of course, it might be questioned whether the data presently available to toxicologists constitute a representative sample. However, that is a standard problem of statistical sampling rather than a difficulty specifically raised by extrapolation, such as the extrapolator's circle. (90)

As I argued in chapter 1, it is in general correct to separate causal inference problems from problems about statistical sampling. In using DAGs, one generally assumes that one knows which variables are correlated. To say that two variables are correlated is not just to say that they co-occur in one's sample, but rather that they co-occur in the population of which one's sample is representative. Thus, causal inference generally abstracts away from statistical questions of how one learns about correlations. Yet, in the context of extrapolation, the assumption that one's data from the model organisms is sampled from the same distribution as the data one from which would collect from the target population is contentious. The worry that a causal relationship might differ between the populations presupposes that the model might *not* have the same probability distribution as the target.

The question of whether one can justify certain extrapolations using the types of sampling assumptions that statisticians invoke is a difficult one that I will return to in chapter 4. We are certainly not entitled to assume at the outset of the inquiry that one can do so. One might be inclined to conclude from this that Steel has not in fact broken the extrapolator's circle.[11] Steel could reasonably respond that comparative process tracing *does* allow one to break the circle, since it allows one to carry over some of the information learned from the model

[11] See Howick et al. (2013, 285-6) for an argument that Steel does not break the circle. They argue that because one needs to examine data from both the model and target population in order to establish the points in the mechanism at which the populations might diverge, Steel requires one to have causal knowledge of the target even at points where the populations are assumed to be similar. While I am sympathetic to their argument, here I grant that Steel does break the circle in order to argue that if his solution succeeds, the problem of the extrapolator's circle can be trivially solved. One could combine my argument here with that of Howick et al. to yield a disjunctive conclusion: either Steel does not break the circle, or the problem of the extrapolator's circle can be trivially solved.

mechanism in investigating the target. The fact that we can only do so by assuming that the mechanisms are similar in certain respects does not undermine Steel's claim to have broken the circle.

The easier it is to solve the problem posed by the extrapolator's circle, the less clear it is that the problem is an interesting one. Once one abstracts away from the biological details in Steel's examples, his solution to the circle is that one can combine observations of a target population with assumptions about how it differs from the model to get a set of statements that deductively entail that a causal relation extrapolates. If, however, one can solve the circle by adopting premises that entail that the causal relation extrapolates, it becomes mysterious why it is useful to think about the circle as a central challenge for extrapolation. It is trivial that one does not need to look at the whole mechanism in the target population if one is allowed to *assume* that the unobserved parts of the mechanism do not differ from the corresponding parts in the model. To the degree that the extrapolator's circle seems like a genuine epistemic problem, it is because one's background information falls short of guaranteeing that a causal relevance relation obtains in the target population and it is unclear what would justify the needed ampliative inference. In fact Steel's use of the phrase 'simple induction' suggests that extrapolations are ampliative inferences. Yet Steel's approach appears to be deductive.

As we will see in the next chapter, Steel's discussion of comparative process tracing contains an important insight. One significant way that causal models facilitate extrapolation is by encoding information about which causal quantities are invariant to changes in specific parts of the model. More specifically, cross-population variation in one variable in a model will correspond to changes in only some of the causal relationships in the population represented by that model. When a causal relationship is invariant to all suspected sources of cross-population

variation, that relationship can be extrapolated. Pearl and Bareinboim (2012) provide an account of how one can use the invariance properties of models in order to extrapolate. Their account is much more general than Steel's, but it confirms Steel's insight that we can extrapolate by incorporating background information about how populations differ.

Given Steel's solution to the problem of extrapolation, his framing of the problem in terms of the extrapolator's circle and simple induction is misleading. Steel rejects simple induction on the grounds that the causal relationship of interest may differ across populations. Yet his solution to the extrapolator's circle essentially relies on using simple induction for the parts of a mechanism that one believes to be invariant between the model and the target. At the points where the mechanisms do potentially diverge, one does need to look at the target mechanism. Steel's approach does not allow one to bypass the assumption that many causal relationships are identical between the model and the target. What it does do is show one how to combine the assumption that *some* of the relationships are invariant with partial information about the target mechanism in order to extrapolate. As we will see in our discussion of Pearl and Bareinboim, this approach is extremely fecund.

3. Extrapolating Positive Causal Relevance

In the introduction, Steel presents two challenges that must be addressed by any account of extrapolation. The first is the extrapolator's circle. The second is that of "how it can be possible to extrapolate from model to target even when some causally relevant differences are present" (4). I will refer to this as the *heterogeneity problem.* Steel is interested in this problem in part because of its relevance to animal models. Scientists interested in knowing whether a drug has a certain negative effect in humans will first test it out on organisms (often rats or mice) that are

assumed to be a good model for humans. No one would doubt that there are causally relevant differences between humans and rats that will influence how they respond to a particular drug. Yet, if scientists thought that the effect in rats did not provide evidence about the effect in humans, it would be pointless to experiment on rats. The challenge is to explain how it is possible for the results of experiments on rats to provide evidence for the effect in human populations despite the presence of causally relevant differences between rats and humans.

Comparative process tracing does not solve the heterogeneity problem, since in any case where one suspects that the model and the target differ one cannot extrapolate, but must inspect the target mechanism. Steel's solution to the heterogeneity problem is that if one only extrapolates *qualitative* causal claims, one can successfully extrapolate even if the causal relation is not exactly the same in the model and the target. He concisely articulates this solution in the introduction:

> The central point is that the closeness of match required between model and target depends upon the specificity of the causal claim that one wishes to extrapolate. In particular, a total absence of causally relevant disanalogies is not required for extrapolating claims about positive and negative causal relevance. (8)

In other words, while the problem of heterogeneity *would* undermine an attempt to extrapolate maximally specific quantitative claims about the magnitude of an effect, they do not similarly undermine less specific qualitative claims about the direction of the effect. In this section, I argue that this is not an adequate response to the problem.

Outside of a short appendix, Steel only discusses the extrapolation of qualitative causal claims. Unlike quantitative causal claims like "Smoking increases one's risk of cancer by 30%", qualitative causal claims such as "smoking causes cancer" assert that a treatment is positively relevant, negatively relevant or neutral to an outcome. For a dichotomous variable such as "smokes/does not smoke" these three types of relevance are easily definable and they form

jointly exhaustive categories. For variables with more than two values, these categories are not well defined without making further stipulations; Smoking two packs a day is positively relevant to cancer relative to smoking one pack a day, but negatively relevant to cancer relative to smoking three packs a day (Hitchcock, 1993). Steel provides stipulations that allow one to extend these categories to non-dichotomous variables. The precise stipulations will not matter in what follows, but it is worth noting that the extended categories of relevance are not jointly exhaustive. To give just one example, a treatment that alters the distribution of an outcome without changing its expected value will be neither positively nor negatively relevant to the outcome, but it is not causally irrelevant.

Steel's choice to focus on the extrapolation of qualitative claims is essential for his solution to the heterogeneity problem. The reason that it is possible to extrapolate in cases where two populations are causally different is that *C* can be positively relevant to *E* in several populations even if the magnitude of the effect differs across the populations. For example, one can extrapolate the claim that "pesticides cause cancer" from mice to humans even if the magnitude of the effect of pesticides on cancer differs between the two populations. Were Steel considering the question of when one can infer that the magnitude of a causal relationship is the same across populations, this solution to the heterogeneity problem would obviously not be available.

As I've already noted, one of Steel's primary reason for considering the heterogeneity problem is that it arises in the context of animal models. In particular, he is concerned with responding to LaFollette and Shanks' (1995) claim that if there is *any* causally relevant difference between the model and the target, one cannot extrapolate. His response is that although causal differences can lead to differences in the exact magnitude of an effect, the effects

in the populations could still be qualitatively the same, in that both could be positively or negatively relevant. While this response does show that LaFollette and Shanks' position is too strong, it does little to address the problem of heterogeneity. The problem, as I understand it, is that of how one can learn anything about the target population in cases where there are causally relevant differences between them. Of course, if one had some knowledge about the factors that make a difference and about how much of a difference they make, one could easily use this knowledge to extrapolate. But the problem of extrapolation arises precisely because one does not have such knowledge.

In the next section, we will evaluate Steel's proposal for how one can extrapolate claims of positive causal relevance across populations. What I want to emphasize here is that the fact that the claims he is extrapolating are qualitative as opposed to quantitative can play no epistemic role in the account. The fundamental problem is that the effects in the model and the target – at whatever grains they are defined – are potentially different and we need to give a reason to think that they are not *so* different as to render our knowledge of the model useless for thinking about the target.

To illustrate this point, suppose that a cause raises the expected value of its effect by .3 in the study population and one characterizes it as being positively causally relevant to its effect. To know that the cause is also positively causally relevant in the target population, one needs some way of knowing that the magnitude of the effect is not more than .3 less than the effect in the study population. If one did not know at least this, one could not extrapolate positive causal relevance across the populations. What this shows is that even if one decides to only make qualitative causal claims, one can only extrapolate such claims if one has some knowledge of the magnitudes of the causal effects. One's knowledge might be imprecise. But unless one is

tracking the magnitudes of the effects, one will not be able to extrapolate even qualitative claims. More generally, to solve the problem of heterogeneity, one needs to specify how a causal effect can be similar in two causally different populations. The fundamental question is not whether one characterizes them as being quantitatively or qualitatively similar, but how one can know that they are similar at all.

4. The Mechanisms-Based Approach to Extrapolation

Comparative process tracing involves comparing a mechanism found in one population to that mechanism in other populations. To extrapolate the claim that a certain substance causes cancer, it is not enough to show that it does or does not cause cancer via a particular mechanism, since there could be other mechanisms involved. To bridge this gap, Steel utilizes the formal apparatus underlying Woodward's account of causation to develop a precise sufficient condition under which a qualitative causal claim can be extrapolated. The condition is the antecedent of *the extrapolation theorem,* which states that if a treatment is positively causally relevant to an outcome via all of the mechanisms and combinations of mechanisms from the treatment to the outcome that actually occur in a population, then the treatment will be positively relevant to the outcome just in case there are some members of that population for whom some of those mechanisms are not *disrupted*. For a mechanism to be disrupted is for there to be an intervention that destroys the causal connection between the starting point and the termination point of the mechanism. If there is no mechanism between two variables, then it is trivial that all mechanisms between them are disrupted.

The condition in the antecedent of the extrapolation theorem is stringent. It does not apply if there are any mechanisms that exert a negative influence unless those mechanisms

always co-occur with another mechanism such that the joint effect of the two mechanisms is positive. Moreover, the extrapolation theorem does not appear to have anything to do with extrapolation. It is a condition stating that if the distinct mechanisms within a population have a certain property, then the total effect will be positive *within that same population.* The only method given for comparing mechanisms across populations is comparative process tracing.

The sense in which the extrapolation theorem concerns extrapolation is that when one considers a population in which every subpopulation has some subset of undisrupted mechanisms between *X* and *Y* and the antecedent of the theorem obtains, one can extrapolate from the subpopulations to the population as a whole. Specifically, one can infer that as long as there is *some* subpopulation in which not all of the mechanisms are disrupted, *X* will be positively relevant to *Y* in the population as a whole. Even though this involves an inference from a claim about subpopulations to one about the whole population, to call this as an extrapolation is misleading. Since the effect of *X* on *Y just is* the effect going through all mechanisms, the assumption that all combinations of mechanisms lead to a positive effect guarantees that there is a positive effect in the population as a whole. The claim that the mechanisms in conjunction produce a positive effect is *assumed* rather than inferred from independent claims about subpopulations.

Steel devotes two sentences to explaining how the extrapolation theorem relates to the other parts of his account:

> Comparative process tracing…would be the basis for the claim that there is a mechanism from *X* to *Y* in P [the population considered by the extrapolation theorem]…Thus, the extrapolation theorem illustrates how the step from extrapolating a mechanism to extrapolating positive causal relevance can be made. (113-4)

The idea here seems to be that one uses comparative process tracing to establish the existence of particular mechanisms in the population and then appeals to the extrapolation theorem to make

an inference about the combined behavior of the mechanisms. The problem with this proposal is that one cannot use DAGs to represent particular mechanisms in isolation, as I will explain.

Suppose that exercise reduces the risk of heart disease both by reducing body fat and by causing one to sleep better (figure 4):

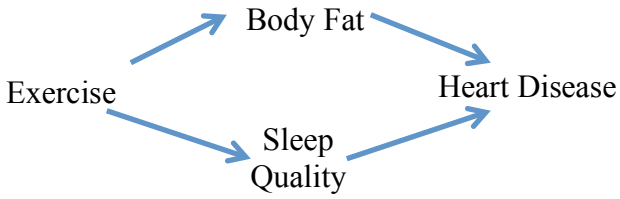

Figure 4

There are two distinct mechanisms here. If this were the correct DAG, one could not use the following DAG to represent the effect of exercise on heart disease via reducing body fat:

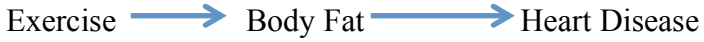

Figure 5

According to this DAG, exercise is not a direct cause of heart disease. Figure 5 entails (by the Causal Markov Condition) the false conclusion that conditional on body fat, exercise and heart disease will be uncorrelated. To correct for this, one needs to substitute figure 4 or the following (figure 6):

Figure 6

The DAG in figure 6 would be an appropriate one to use if one were primarily concerned with the effect of exercising on heart disease through its reducing body fat. The effect of exercise on heart disease through any other mechanisms (in this case sleep quality) would be accounted for by the direct path from exercise to heart disease.

In his discussion of comparative process tracing, Steel presents a DAG with a single causal path. This DAG entails that there is no path from the first node to the last one other than the one represented, so there is no need for a principle saying how to combine this path with others. If this DAG were false and there *were* other paths, then one would need to consider those paths in evaluating the contribution of the path he describes, since the activity of each path alone is not necessarily a reliable guide to the behavior of the combined paths.

The feature of DAGs just discussed rules out the following way of relating comparative process tracing to the extrapolation theorem. Suppose that one evaluates the extrapolation theorem with respect to a DAG in which there are three causal paths between variables *X* and *Y*. One cannot break this DAG into three DAGs – one for each path – and then use comparative process tracing to evaluate the contributions of the paths individually. If one represents one of the causal paths, one needs to indicate that there are others – even if one does not include any variables along the other paths, but simply has an arrow from *X* directly to *Y*.

It remains unclear to me how comparative process tracing is supposed to relate to the extrapolation theorem in Steel's account. Here I will put this question to the side, since the extrapolation theorem does not appear to be useful for extrapolation, however they are related. An important point that emerges from the present discussion is that even if Steel can justify using comparative process tracing in the single-path case that he presents, he has not provided a way of relating the single-path case to more complicated cases in which there are multiple paths. The

problem is that variables along different causal paths interact, so there is no simple way to evaluate the contributions of individual paths in isolation. Later on in the dissertation (Chapters 5 and 6), I explain how one can use causal mediation techniques to evaluate the contributions of individual paths in cases where there is interaction. In Chapter 7, section 5, I explain how one can use these techniques to generalize Steel's treatment of the single-path case.

5. Conclusion

Steel's account relates to the rest of the dissertation in two ways. First, as I have noted, Steel's account anticipates Pearl and Bareinboim's account in several respects (I will elaborate upon this point after presenting their account). Second, in the later chapters of the dissertation, I further develop the mechanistic approach with the help of causal mediation techniques.

The discussion of Steel thus far has yielded few positive results. His account of comparative process tracing sheds some light on when one can extrapolate mechanisms across populations, but there remain important unanswered questions regarding how one learns about the points at which the mechanisms are likely to differ. His extrapolation theorem does not appear to license any genuine extrapolations. Nevertheless, the results of the dissertation will validate his general approach. In particular, in chapter 7 I defend the thesis that that in cases where all of the mediators between a cause and effect are on a single path, measuring the effects between some of the variables along that path facilitates extrapolation. This idea is very similar to the one that Steel defends in discussing comparative process tracing. Where my analysis goes beyond his is in providing a more rigorous account of the relationship between single-path and multi-path cases.

Chapter 4: Transportability

Consider two possible causal explanations for why Canadians are more liberal than Americans. One, a higher percentage of Canadians live in cities than do Americans and city-dwellers tend to be more liberal. Two, Canadians are genetically predisposed towards liberalism, so both urban and rural Canadians are more likely to be liberal than their US counterparts. Both of these explanations are compatible with the observations that Canadians and city-dwellers are more liberal on average. If, however, the first one is correct, this has an important implication for extrapolating causal claims from Canada to America (or vice versa). The first explanation entails that if the US were to have the same distribution of people living in cities as Canada does, then it would be as liberal as Canada. Consequently, one could determine the percentage of liberals in Canada without actually measuring the effect of living in a city on Canadians' political views by taking the data regarding the propensities of American city-dwellers and non-city-dwellers to be liberal and giving a weighted average based on the ratio of city-dwellers to non-city-dwellers in Canada.

The inference in the previous paragraph might seem trivial. It was only possible to extrapolate between the countries on the unrealistic assumption that they only differed in a single respect. Yet, there are ways the countries might have differed such that one would not have been able to extrapolate. For example, if the second explanation is correct, then even were the countries to have the same urban/rural distribution, one would have to estimate the effects of living in a city on one's politics separately for each country. One could not assume that American urbanites are as likely to be liberal as Canadian urbanites, since the genetic difference produces a difference in the probability that members of these groups are liberal.

The simple example just presented reveals that whether it is possible to extrapolate among populations depends on how they differ. When *C* causes *E* in two populations, it is possible to extrapolate when the populations differ in the unmeasured causes of *C*, but not necessarily when they differ in the unmeasured causes of *E.* In the case where the populations differ in unmeasured causes of *C* it is possible to extrapolate using an averaging process such as the one mentioned in the first paragraph. Such averaging processes are referred to as *adjustment.* If the first explanation is correct, one can derive the effect of living in a city on liberalism in Canada by taking the effect in the US and adjusting for the difference in urban-to-rural ratio.

Judea Pearl and Elias Bareinboim (2012) consider a type of extrapolative inference in which one uses experimental and observational data from one population to make an inference about a causal quantity in a population for which one only has the probability distribution. When such an inference is possible, the causal quantity in the experimental population is *transportable* to the target population. By *experimental data* they mean the results of randomized control tests. When a causal quantity is transportable, it is possible to identify it in the target using an *adjustment formula* that indicates which probabilistic terms must be reweighted according to the probability distribution for the target. The authors develop a procedure for distinguishing between transportable and non-transportable quantities given one's background knowledge of the differences between the populations.

Here I provide a non-technical introduction to the transportability framework. In it, I present the authors' method of representing population differences, consider paradigm cases of transportable and non-transportable quantities, and explain why transportability succeeds or fails in those cases. I do not explain how to derive adjustment formulas for transporting causal relationship. Readers interested in how to derive such formulas can refer to the technical

appendix. Here my aim is to explain the fundamentals of the account and make explicit which problems it addresses and which are left open.

This chapter is organized as follows. Section 1 reviews directed acyclic graphs, introduces the concept of identifiability, and explains how the interpretation of a causal quantity in a DAG depends on which background factors are included in the DAG. Section 2 shows how to derive distributions for subpopulations by conditioning on particular variables in a DAG and describes how Pearl and Bareinboim distinguish among populations using selection diagrams. Section 3 explains how to use selection diagrams to distinguish between transportable and non-transportable quantities. Section 4 shows how one can achieve transportability by measuring mediators. Section 5 explores the conceptual basis for transportability by asking why DAGs alone are insufficient for representing extrapolation. Section 6 argues that transportability is a special case of a more general problem of extrapolation, which I characterize. Section 7 considers the relationship between extrapolation and induction. Section 8 revisits the problem from chapter 1 about how to characterize populations. Section 9 compares transportability to the accounts evaluated in chapters 2 and 3. Section 10 concludes.

1. DAGs and Identifiability

Correlation does not imply causation, but given a model for the causal relations among a set of variables one can measure the magnitudes of (some of) the causal relations between variables. The causal relations among a variable set can be represented using a directed acyclic graph (DAG). Consider the following graph for the relationship between one's parents' socioeconomic status (SES), education and whether one likes opera (figure 1):

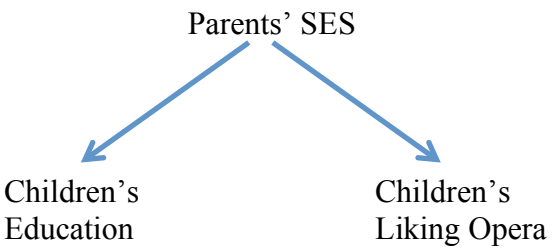

Figure 1

According to this graph, one's parents having a higher SES increases both the chance that one will be educated and that one will appreciate opera. Assuming the causal faithfulness condition, all three variables will be correlated, though there is no arrow between education and opera. This conveys that education is not a *direct cause* of liking opera (or vice versa)[12]; according to this causal model, if you take a person of a particular SES and give her more education, this will not increase the probability that she will like opera. Note that I have not provided any reasons for thinking that the graph in figure 1 is correct. Figure 1 is simply an example of how one can use DAGs to encode one's causal beliefs about a set of variables. While I introduced DAGs in the previous chapter, here I will say more about how they relate to probability distributions and structural equations.

DAGs represent the way that a set of factors generates a probability distribution. Each variable in a DAG is a random variable, which means that it has at least two possible values and each value is assigned a probability. For example, the variable "likes opera" might have two possible values – yes or no – where the probability that a randomly selected person likes opera is 20%. One might want to know not merely the probability that a randomly selected person likes opera, but whether a randomly selected person of low SES does. To answer this question, one

[12] In this graph, education is also not an indirect cause of liking opera. A necessary condition for *X* being a cause (either direct or indirect) of *Y* is that there is a series of unidirectional arrows from *X* to *Y*. *X* is a direct cause of *Y* just in case there is an intervention on *X* that changes the value of *Y* while keeping all other variables in the model fixed (Pearl, 2009; Woodward 2003). For more on interventions, see chapter 3.

must determine the probability of liking opera *conditional on* having low SES. Liking opera is correlated with SES just in case the unconditional probability of liking opera differs from its probability conditional on some value of SES. Unconditional and conditional probabilities are features of the probability distribution itself and can be measured independently of any DAG.

As already noted, DAGs make assumptions that cannot be reduced to features of the probability distribution, since they convey not merely which variables are correlated, but which variables can be used as a means to alter the values (or probabilities) of other variables. The arrow between SES and education, for example, conveys not merely that these two variables are correlated, but that one can influence an individual's level of education through an *intervention* on SES. As we saw in the previous chapter, an intervention sets a variable to a particular value in a manner that is independent of its prior causes. One can represent an intervention of this sort by deleting all arrows going into the variable upon which one intervenes. Some (Korb et al. 2004; Eberhardt and Scheines, 2008) advocate a broader notion of interventions on which some interventions do not set a variable to a particular value, but rather alter its probability distribution. Such interventions do not "break" all the arrows going into a variable, since the distribution of the variable stills depend on its non-intervention causes. The essential feature of arrow-breaking and non-arrow-breaking interventions alike is that they only influence a variable *X* through changing its causes and they only influence other variables in the model *via* changing *X*'s value or distribution.

It should be clear by now that the definition of an intervention invokes causal concepts. Accounts that use interventionist concepts to explain causal relations are not trying to reduce causes to something non-causal (such as probabilities). Nevertheless, DAGs in which the arrows between variables encode information about how the probability distribution will change in

response to interventions are extremely useful for mapping out the relationship between probability distributions and causal hypotheses.

A key notion in what follows is *identifiability.* To say that the effect of SES on education in figure 1 is identifiable is to say that one can uniquely determine the (average) causal effect of SES on education based on the probability distribution. Identifiability is always relative to a DAG. One cannot determine whether, or how strongly, SES causes education from the probability distribution alone, but given the DAG in figure 1 plus the probability distribution one can measure the magnitude of this effect. Since all of the causal relations in figure 1 are identifiable it will help to present a DAG with non-identifiable quantities:

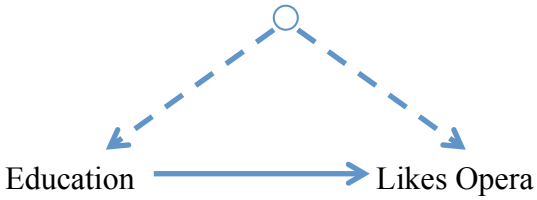

Figure 2

In figure 2, education is a cause of liking opera. The two variables are also connected by a *bidirected arc,* which represents an unmeasured common cause of education and liking opera. In figure 2 the effect of education on liking opera is non-identifiable. The reason is that even if education and opera appreciation are correlated, one cannot determine from the probability distribution how much of the correlation is due to the direct causal connection and how much is due to the unmeasured common cause.

Whether a causal quantity is identifiable is a distinct question from whether it can be *estimated* from a finite data set. Identification concerns whether a quantity in a DAG is uniquely determined by the true probability distribution; estimation concerns inferring the relevant

features of the distribution from one's data. Most theoretical discussions of causal modeling – especially Pearl's – stipulate that one already knows the probability distribution, allowing one to bypass the statistical issues related to estimation. This is a useful idealization, but it is always important to bear in mind that when we discuss probability distributions we are not talking about the relative frequencies of traits in a finite sample, but rather specifying what the frequencies would be in an infinite population. As a matter of practice, one approximates such a population by using a random sampling process and having a sample large enough to limit the amount by which the sample diverges from the true distribution by chance.

In looking at a DAG, one needs to pay attention not just to which arrows connect the variables, but also which arrows are missing. If a DAG does not contain an arrow between two variables, this means that there is no direct causal relation between them and if there is no bidirected arc between the variables, this means they do not share a common cause. It is often very difficult to rule out the possibility that two variables share a common cause. When two variables share an unmeasured common cause one can only measure the causal relation between them using a randomized control test. A randomized control test is a type of intervention in which one randomly assigns participants into a study and control groups. Since the assignment is random, whether a subject is in a control group no longer depends on the common cause and as a result one can measure the desired causal relation.[13] Were one able to assume that the two variables in question do not share a common cause, one would be able to identify the causal relationship between them without an experiment. The trend among social scientists towards

[13] Randomization is not always successful. Even if the mechanism by which the researcher assigns subjects into treatment and control groups is random with respect to the variables in the model, in finite populations the causal features of the members of each group may diverge by chance. The discussion here is limited to idealized randomized control trials, in which randomization is successful.

only accepting causal claims that have been established by a randomized control test suggests that in general they do not feel justified in ruling out the possibility of a common cause.

There is a temptation to consider the separate causes of an effect in a DAG as each contributing additively to the total effect. This temptation should be resisted. Even the simple causal relation of a fire being caused by the striking of a match depends on background conditions such as the presence of oxygen and the dryness of the match. The DAG for this case may look as follows:

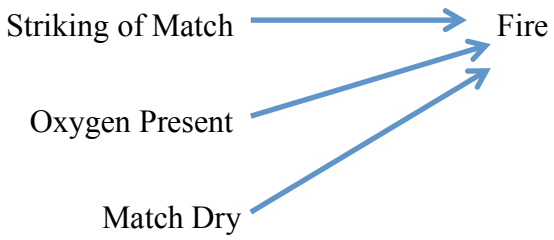

Figure 3

The DAG in figure 3 conveys only that the lighting of the fire depends on three distinct factors, but does not specify whether each one contributes a fixed amount or whether they interact. When people claim that striking a match causes a fire they are taking for granted that there is oxygen and the match is dry. The factors that are taken for granted are referred to as conditions rather than causes, but this distinction is pragmatic. A condition is just a cause that one does not represent. When one does not explicitly represent the presence of oxygen or the dryness of the match in the DAG, the variation in the effect of striking the match on whether there is a fire is captured by an *error term* that represents the unmeasured causes of fire that vary within the population.[14] Omitting these error terms will not inhibit one's ability to estimate the effect of

[14] Here I assume that all background conditions may be represented as causal factors whose influence on the model may be captured in an error term. Hausman et al. 2014 provide a case in which the relations in a causal model only obtain when certain background conditions remain fixed. In such cases, one cannot think of such background conditions as naturally varying error terms.

striking a match on the lighting of a fire, provided that none of the omitted causes is a common cause of fire or of other variables in the DAG.

The fact that DAGs allow for interactions among causes of an effect has an implication for how one must interpret the arrows in a graph. Compare the DAG in figure 3 to one in which the only variables are the striking of a match and the lighting of a fire (figure 4).

Figure 4

The effect of striking a match on lighting a fire is identifiable in both figure 3 and figure 4. Confusingly, although both DAGs contain an arrow between striking a match and lighting a fire these arrows represent different quantities. In figure 4, the arrow represents the average effect of striking the match on the lighting of a fire in a population of different lighting events. In figure 3, the separate causes interact with one another, so there is not a single quantity to estimate for the effect of striking on lighting. Instead, there is the effect of striking when there is oxygen and the match is dry, the effect of striking when there is oxygen and the match is wet and so on. In this DAG it does not make sense to consider the contribution of an arrow by itself.

The reader might be puzzled how both figures 3 and 4 could constitute adequate representations of the same process, given that figure 4 misses the variability in the effect of striking the match. Recall that the variables in DAGs implicitly have error terms that account for any unmeasured causes of that variable alone. So in figure 4, one could imagine an unmeasured variable that accounts for all other causes of fire, including oxygen and the dryness of the match. These background factors will vary among the members of the population. Although one never measures these factors, as long as one's sampling process is representative and the sample is

sufficiently large, the variation of these factors in one's sample will be the same as the variation of these factors in the population from which one is sampling. Among strikings of the match in one's sample, the effect of the striking will vary as a result of variation in these factors. But assuming that the distribution of these factors is same as the distribution in the broader population, the average effect of striking the match on the lighting of the fire will be the same as it is in the general population. In other words, one's estimate of the effect of striking the match will be unbiased, though there will a lot of variation in the magnitude of the effect that will not be accounted for in the model. In figure 3, one includes additional variables that account for more of the variation in the effect of striking a match. The effect variable in figure 3 would still have an error term accounting for any remaining variance, though this error term would be different from that in figure 4, since it would not capture the variation in the effect that results from the variation in the values of *oxygen present* and *match dry*.

It is of obvious importance for extrapolation that the magnitude of an effect can depend on unmeasured and varying background conditions. The reason why a causal effect measured in one population may not generalize to other populations is that the background conditions may vary across populations. When the relationship between *C* and *E* depends on a third variable *Z*, *Z* is known as an *effect modifier* (Vanderweele and Robins, 2007). Causes of *E* that interact with *C* are always effect modifiers.

Each variable in a DAG is related to its direct causes via a structural equation. These equations indicate that each variable is a *function* of its direct causes. Crucially, the relationship between a variable and its causes may have any parametric form. This is why one should not assume that the causes of an effect make additive contributions. Such an assumption presupposes that there is no interaction term in the function relating the effect variable to its causes and thus

places a restriction on the parametric form of the function. Causal inference that does not make any assumptions about the parametric form of structural equations is called *non-parametric* causal inference. In principle, when a causal effect is identifiable, one can determine the functional relationship between an effect and its causes from the probability distribution alone; one does not need to make any parametric assumptions.

The question of whether one should make parametric assumptions in interpreting a causal model is distinct from the question of whether one should make parametric assumptions in *estimating* a causal quantity from finite data. In practice, social scientists with limited data sets often do make parametric assumptions when it comes to estimating causal quantities. This is compatible with adopting a non-parametric approach to causal inference. Pearl's approach to causal inference is non-parametric insofar as one does not need to make any parametric assumptions to determine whether a particular quantity Q is identifiable from the probability distribution. Whether one should use parametric or non-parametric methods for estimating Q from one's data depends on the nature of one's data set.

2. Populations, Subpopulations and S-nodes

A useful way to think about the difference between causal inference using DAGs and extrapolation is to note that while a DAG represents a probability distribution for a single population, extrapolative inferences are inferences across populations. This idea is basically correct, provided one notes a few caveats regarding the claim that DAGs represent single populations. First, since a DAG represents how a probability distribution would change as a result of interventions, it is not quite right to say that it only represents a single probability

distribution; it also represents the distributions that would result from interventions.[15] Second, and more importantly, although a DAG represents a probability distribution for an entire population, one can learn information about subpopulations by conditioning on values of variables in the DAG. Let's consider an example of how this works.

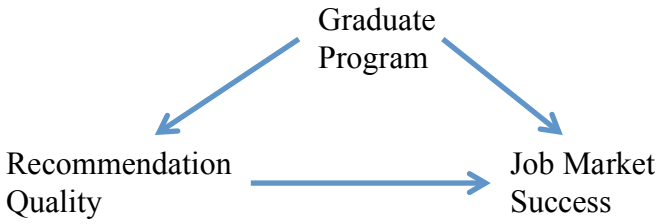

Figure 5

Suppose one wants to estimate the effect of recommendation quality on how one does on the academic job market. These variables share a common cause, since being in a particular graduate program influences both the quality of one's recommendation and one's chances of job market success. On the (unrealistic) assumption that there is only this one common cause of the two variables, one can calculate the average effect of recommendation quality on job market success by looking at the probability of success conditional on recommendation quality for each graduate program and then taking a weighted average of the conditional probabilities in different departments. For concreteness, suppose that there are only two programs, Oxford and Cambridge and that the probability of getting a job given that one has received a good Oxford recommendation is .8 and the corresponding probability at Cambridge is .6. If the probability that a student attends Cambridge is equal to the probability that she attends Oxford, then the probability that one will get a job given that they received a good letter is .7. To determine the causal effect of receiving a good recommendation one would have to compare this number to the

[15] See Sprites et al. 2000, p. 51, for a precise characterization of the relationship between the manipulated and unmanipulated graphs

probability of getting a job given that one does *not* receive a good letter.[16] This could be calculated in the same manner. When one looked at just the Oxford students or just the Cambridge students one effectively conditioned on each of the values of the variable for graduate program. This shows that even though the DAG in figure 5 refers to the entire population of Cambridge and Oxford attendees, by conditioning on the *graduate program* variable one can learn information about an effect in subpopulations.[17]

The possibility of distinguishing between subpopulations by conditioning on variables in a DAG suggests a general strategy for representing extrapolative inferences. When extrapolating from one population to another, one can represent both populations within a single DAG and derive the probability distributions for individual populations by conditioning on the values of certain variables. This is the strategy that Pearl and Bareinboim utilize and to do so they introduce what they call *selection diagrams.*

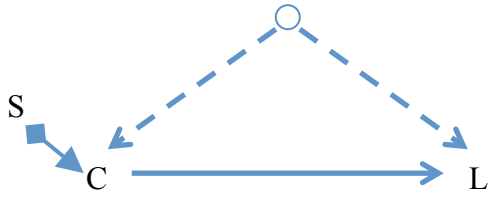

Figure 6

[16] Pearl defines the causal effect of *X* on *Y* not as the difference in *Y* given two distinct interventions on *X,* but rather as the probability distribution of *Y* given a single intervention on *X.* Accordingly, the effect of receiving good letter on getting a job would just be the probability of receiving a job given that one is assigned a good letter via an intervention. In many cases, I find it more natural to define the causal effect as the difference in *Y* given two settings of *X.* This terminological issue will not make a difference in the present chapter.

[17] As an aside, the fact that one can use DAGs such as the one in figure 5 to identify the causal effects in subpopulations stratified based on a common cause is important for understanding the debate between advocates and critics of randomized control trials (RCTs). The advantage of RCTs, clearly, is that the researcher does not need to know all of the common causes between two variables in order to measure the causal effect. The advantage of observational studies in which one does not randomize is less obvious. What this example shows is that if one measures *all* the common causes, one thereby identifies not only the average effect in the whole population, but the average effect for each subpopulation stratified on the common cause variable. In other words, *if* the assumptions in an observational study is met and one can estimate all identifiable quantities, one learns more from it than one would have learned from just an RCT.

Consider the causal model in which living in a city (*C*) causes one to be liberal (*L*) and the only difference between Canada and the US is the urban-to-rural ratio. Here I assume that there are unmeasured common causes of living in a city and being liberal. *S* is called an *S-node* (for selection) indicating all and only the mechanisms by which the two populations differ. Here I use the term "mechanisms" loosely to indicate the way variables are assigned their values. There is no variable in the DAG for whether one lives in the US or Canada. The S-node conveys that the two populations differ in the mechanisms determining the value of *C*, but it is not necessary to know what these mechanisms are. Whether Pearl and Bareinboim's procedure applies in a case depends on which variables have S-nodes pointing into them.

The selection diagram in figure 6 corresponds to the first explanation I suggested in the introduction. There I noted that if the only difference between the two countries is in the urban-to-rural ratio, then provided that one could measure the effect of living in a city on being liberal in one country, one could derive the probability distribution for the other country by simply adjusting the distribution of the *C.* Since there is a common cause of *C* and *L,* determining the effect of *C* on *L* in one of the countries would require a randomized control test. Once one did so, however, one could estimate the average effect in the other country without having to do another experiment. In Pearl and Bareinboim's terminology, the relationship between *C* and *L* is *transportable.* In fact, whenever figure 6 is the correct selection diagram for a set of populations, the relationship between the cause and effect is transportable. Of course, not all relations are transportable. If, for example, there were an S-node pointing into *L* (Figure 7)*,* there would be no way to determine the effect of *C* on *L* in any population on which one has not performed an experiment. In short, the location of the S-nodes determines whether or not a relationship is

transportable and Pearl and Bareinboim provide a general procedure for determining which causal quantities in a selection diagram are.

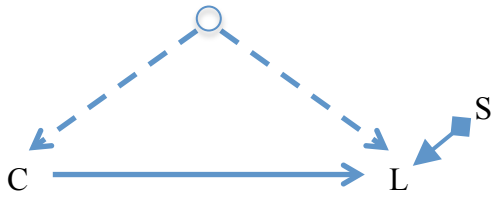

Figure 7

I introduced selection diagrams by noting that within a DAG, one can individuate populations by conditioning on different values of a variable. In selection diagrams, the "variable" one conditions upon is *S*. Given that one never measures *S* (and might not have any idea what it referent is), how can one "condition" upon it? To explain this, I need to go into a bit more detail about how Pearl and Bareinboim would deal with figure 6. We can evaluate the DAG for the population upon which one performs the experiment without considering the S-node at all, since the S-node only marks the *difference* between the populations. To evaluate the target population, we think about how the probability distribution of the study population would change were one to condition on the S-node. Here is the crucial point. The only way that variation in the S-node influences the other variables in the model is through influencing the variation in *C*. Thus, if one estimates the probability of *C* in the target population, there is no added benefit to measuring whatever variables S-represents. The reason why we can evaluate the effects on conditioning on an unknown variable is that the consequences of conditioning are entirely reflected by changes in the probabilities of known variables.

At this point it should be clear both how DAGs encode causal information about a population and how selection diagrams differ from DAGs. In the following section, I will say

more about transportability and explain which types of relationships are transportable. Before doing so, it is worth noting three features of Pearl and Bareinboim's approach. First, they represent both the population from which one extrapolates and the target population in a single diagram. This contrasts with approaches that represent different populations using separate DAGs (Steel, 2008; Cartwright and Hardie, 2013). While both approaches are legitimate, Pearl and Bareinboim's approach is better suited for explicitly representing the differences between populations. Second, the reader has probably noticed that I have not said anything about how one determines the respects in which populations differ. This is because Pearl and Bareinboim are concerned with whether extrapolation is possible given a set of assumptions, not with the question of how one justifies those assumptions. Third, although it is natural to focus on the presence of S-nodes, the strongest assumption encoded in a selection diagram is that variables *without* S-nodes share the same mechanisms across populations. Only in cases where there are at least some missing S-nodes is one able to identify causal quantities in the target that would not be identifiable without transportability methods.

3. Transportable and Non-Transportable Relations

A causal quantity is transportable from a study population to a target population just in case it is identifiable in the target population based on the probability distribution for both populations and experimental results from the study population. In figure 6, the effect of living in a city on political views was transportable from Canada to the US, since as long as one could do an experiment to determine the effect in Canada, one could then estimate the effect in the US using only facts about the probability distribution in the US – specifically, the distribution of those living in cities.

As the effect of living in a city on liberalism is transportable in figure 6, but not in figure 7, one might conclude that a causal relation is transportable when populations differ in the causal variable, but not when they differ in the effect variable. This would not be quite right, since there are cases in which it is possible to transport a relation even though populations differ in their mechanism for the effect variable. The simplest such case is presented in figure 8:

Figure 8

In figure 8, there is no common cause of *C* and *L*, so the effect of *C* on *L* is identifiable in both populations from their probability distribution alone. As the relation in the target population can be identified from that population's distribution alone, it clearly can be identified given the distributions for both populations and experimental data. In such a case, the relation is *trivially transportable.* Of course, issues of transportability only arise in practice when one cannot identify the effect in the target population from its probability distribution alone. Absent an experiment one would typically not be able to rule out a common cause of *C* and *L* and therefore not be able to identify the effect of *C* on *L* from the distribution. The relevant diagram would thus be figure 7 rather than figure 8. In general, if two variables *X* and *Y* share an unmeasured cause in two populations, then the relationship between the variables is transportable given differences in unmeasured causes of *X*, but not given differences in unmeasured causes of *Y*.

Now that we have two examples of transportable relations (figures 6 and 8), we can say something more general about why causal knowledge is ever transportable. Although there are many ways that two populations can differ, not all differences make a difference for causal

inference. In figure 6, variation in *C* does not make a difference, since the effect of *C* on *L* is invariant to variation in the causal mechanism(s) by which *C* is assigned a value. In figure 8, variation in unmeasured causes of *L* does change the effect of *C* on *L.* Fortunately it is possible to identify the effect of *C* on *L* in the new population without any knowledge of what these causes are, since the effect is identified by the probability of *L* given *C* in the new population. The fact that not every possible source of variation between populations renders one's experimental results on one of them obsolete is especially important when one considers diagrams with more complicated structures.

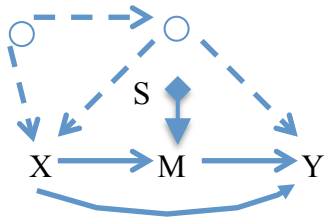

Figure 9

The diagram in figure 9 is more complicated than anything we have yet considered, but we can evaluate the transportability of the effect of *X* on *Y* using reasoning similar to that we applied to figures 6-8. First consider the effect of *X* on *M.* Since there is no common cause of *X* and *M*,[18] this relationship is identifiable in both populations and is therefore trivially transportable. Now consider the effect of *M* on *Y.* As the two populations differ only in the mechanism for *M,* this is similar to the case in figure 6 and this relationship is transportable by adjustment. In this case, identifying these two effects allows one to identify the total effect. Note that the total effect consists of the direct path from *X* to *Y* in addition to those just discussed, but

[18] Those familiar with causal modeling will be aware that there are paths that could bias the effect of *X* on *M* other than a common cause of both. A more general version of this sentence would read "since *X* and *M* are d-separated along every path other than the one going from *X* to *M*, this relationship is identifiable in both populations and is therefore trivially transportable."

this path is the same across the populations. I hasten to add that it is not always the case that if you can identify all of the direct effects in a diagram, then one can identify all effects. In Pearl's terminology, local identifiability does not entail global identifiability (2009, 94). In this case, however, knowing about the component effects does allow one to derive the total effect, for reasons I will not address here.

In figures such as 9, the formula required for transporting a causal relation across populations using adjustment will be fairly complicated (see appendix for details). Without going into the details of such formulas, it is possible to give a simple necessary condition for transportability. In cases where the causal relationship between *X* and *Y* is not identifiable in the absence of an experiment, this relationship is transportable only if there is no S-node going into *Y*.

4. Effect Modification and Causal Mediation

Pearl and Bareinboim's treatment of transportability demonstrates that given causal assumptions, it is sometimes possible to transfer the results of an experiment to a target population without having to do a new experiment. This addresses certain challenges to extrapolation. For example, the approach breaks Steel's extrapolators circle, since when one transports a causal quantity, one need not do further experiments on the target population.

Pearl and Bareinboim's approach is limited to cases where the difference between populations is not represented by an S-node into the effect variable (except in cases where the effect is trivially transportable). Intriguingly, their paper does contain tools for dealing with such cases. To utilize these tools, one must consider transportability in a slightly different manner than

the way we have been up to now. So far we have been considering the following task: Given a model for a fixed set of variables, determine which causal quantities are transportable. I now want to ask a different question: given that a quantity in a model is not transportable, is there a model with additional variables in which it *is* transportable? Let's consider a case in which there is.

A pharmaceutical company is testing the effect of a drug on heart disease. They perform a randomized control trial and discover that it is somewhat effective at lowering heart disease on average. Yet, based on this trial, they cannot determine how well the drug will work in the population as a whole or even in particular members of the trial population. They suspect that there might be widespread variation in people's responses to the drug. The correct selection diagram will therefore be the one in figure 10 and, unfortunately, the causal relationship from the drug trial is not transportable.

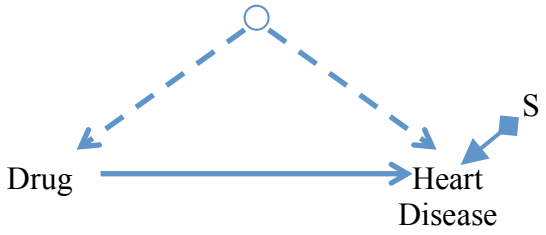

Figure 10

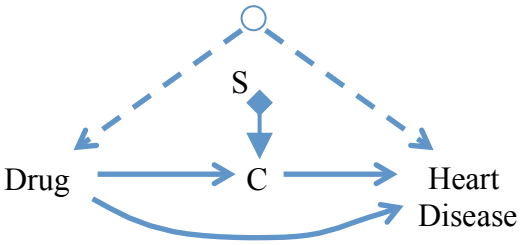

Figure 11

Both Steel (2008) and Cartwright and Hardie (2013) note that that in thinking about whether a causal relation generalizes to other populations, it helps to consider how the cause brings about its effect. Using Pearl and Bareinboim's framework, we can present one way to make this idea much more precise. Suppose that the drug reduces heart disease in part by

lowering cholesterol (*C*). Once we include *C* in the model, we might end up with figure 11 instead of figure 10. In figure 11, the relationship between the drug and heart disease *is* transportable. Note that figure 11 makes the strong assumptions that all variation in the effect of the drug is due to variation in its influence on cholesterol and that there are no common causes of drug intake and cholesterol or of cholesterol and heart disease. Nevertheless, what figure 11 reveals is that it is sometimes possible to transform a non-transportable relation into a transportable one by adding a variable.

In figure 11, *C* is a mediator. The drug example here reveals that measuring mediators can enable one to make extrapolative inferences. This might appear to support my claim that causal mediation techniques are relevant to extrapolation. However, one is able to extrapolate in this case without appealing to any of the concepts of causal mediation. It therefore remains to be seen whether mediation techniques contribute to extrapolation. However, we don't need to wait for the answer to this question to declare this example to be a victory for the mediation approach to extrapolation. Not only does this example make it clear how measuring a mediator facilitates extrapolation, but it makes precise the conditions under which it is possible to do so and it enables one to make quantitative predictions about effect in the target population.

6. Why One Cannot Extrapolate Using DAGs Alone

A recurring question in this dissertation is: what is the difference between extrapolation and ordinary causal inference? I ask this question repeatedly, since I do not presume that there is a single problem of extrapolation, so different extrapolative inferences need to be considered individually. Pearl and Bareinboim allow us to model one such inference. We can therefore pose a more specific version of the question. Why does transportability require the use of selection diagrams, as opposed to just DAGs? To ask this question is to take a step back. It is already clear

(I hope) that selection diagrams are extremely useful. Yet asking why they are will shed some light on the conceptual contribution of S-nodes. Additionally, we will uncover some extrapolative inferences that go beyond the transportability framework.

Let's consider how one would go about modeling extrapolation in a DAG without a selection diagram. Suppose that smoking (*S*) raises the probability of cancer (*C*) in all humans, but that there is a gene (*G*) that modifies the magnitude of the effect of smoking on cancer. The only way that the gene modifies this effect is by being present (*G*=g) or absent (*G*=a). Further suppose that aside from its role in modifying the effect of smoking on cancer, the gene has no other phenotypic effects. Since it is clear that the effect of smoking on cancer will differ between the population in which everyone has the gene and the one where no one does, this looks like a paradigmatic case of extrapolation. If your initial study only contains people without the gene and you predict that there will be a similar average effect of smoking on cancer in a population of people with the gene, you will be wrong. How can we model the flaw in this inference?

Let's take a step back for a moment to consider different ways you might try to model this case. Suppose you build a model in which you only observe the variables for smoking and cancer (and there are no confounders that you do not condition upon). This model will allow you to correctly identify the average effect of smoking on cancer from the probability distribution. The identified quantity is the average effect of smoking across all values of *G*. This is akin to the match-striking case in which one just has variables for the striking and the lighting and the variation in background conditions is captured by the error term. Moreover, assuming that being in the sample does not cause one to have the gene and that having the gene makes no difference for whether one is sampled, the expected proportion of people with the gene in the sample will be equivalent to the proportion of people with the gene in the general population and also

equivalent to the expected proportion of people with the gene in any sample representative of that population. Of course, in finite samples there could be different *frequencies* of people with the gene, but causal models do not consider frequencies, but rather probabilities. The *probability* of having (or not having) the gene conditional on being in the sample will be the same as the probability of having it (or not) in the general population. In short, simply measuring the values of *S* and *C* will be sufficient for getting an unbiased estimate of the average effect of smoking on cancer in the whole population and in any representative subpopulation.[19]

There is nothing new in the claim that in the case I described you can get an unbiased estimate of the effect of smoking on cancer even if there are factors that lead to variance in the magnitude of this effect. Yet, the point is worth reflecting upon, since it shows that in cases where having the gene makes no systematic difference in whether someone ends up in the sample, there is no way to use a DAG with just the variables for smoking and cancer to represent the possibility that the sample happens by chance to have an unrepresentative distribution of people with the gene. Whatever sampling variance there is will be entirely missed by causal models, since in estimating the probability distribution for a population, one makes a prediction regarding the relative frequencies of the variables in an infinite population (which has no sampling variance). Clearly, the process of inferring probabilities from frequencies is not trivial, but standard treatments of causal modeling assume that it has somehow been accomplished.

In order to consider the case in which one has measured the effect of smoking on cancer only for people without the gene, one must include a variable for the gene (or an effect of the gene) in the model. Even though in both the model with *G* and the model without it there is an

[19] The reader may have noticed that it is analytic to say for some property X that the probability of X in the population is equivalent to the probability of X in some representative sample of that population. Given that the sample population is representative, any differences between it and the whole population must be due to sampling variance. Yet, in discussing probabilities we abstract away from sampling variance.

arrow between *C* and *S*, the two arrows measure different things in each model. Once you include *G*, the arrow no longer corresponds to the average effect of smoking on cancer in the population, but rather the effect of smoking on cancer for each value of *G* (figure 12).

Figure 12 – Two possible DAGs for the effect of smoking (*S*) on cancer (C). Although there is Ian arrow from S to C in each DAG, the two arrows refer to different quantities depending on whether one includes a variable for the presence of a gene (*G*) that modifies the effect of smoking on cancer. In the DAG on the left, the arrow between *S* and *C* represents the average effect of smoking on cancer. In contrast, in order to estimate the effect of *S* on *C* in the DAG on the right, one must measure the effect for each value of *G*.

Once you include *G* in the DAG, it is theoretically possible to estimate both the effect of smoking on cancer in the population of people with the gene (the g-specific effect) and in the population of those without it (the a-specific effect). In this particular case it is hard to imagine a scenario in which one would measure the g-specific effect of smoking but not the a-specific effect. Once a scientist goes through the trouble of figuring out which people have the gene, why would she not measure the effect of smoking on cancer in both populations?

In the case considered, it does not appear likely that one would be forced to make any extrapolations. If, however, a scientist did have data on the effect of smoking on cancer only for people with the gene and she wanted to extrapolate to the population of people without it, how would we represent this scenario? Clearly, we need to include *G* in the model in order to have a way to represent the fact that the distribution of the gene in a sample of gene carriers is not representative of the general population. Yet, in including *G*, we do not want to indicate that we have information about the causal effect of *S* on *C* for people who do not have the gene. Fortunately, a case in which the scientist only has information about people with the gene can be

thought of as one in which the scientist has conditioned on the value of *G*=g. One typically conditions on a value of a variable in cases where one has information about the other values of that variable, though there is no reason not to think of cases in which one *only* has information about one value of the variable as one of conditioning as well. One could think of such cases as ones of *accidental conditioning*.

In the smoking case, I needed to contrive a scenario in which the scientist lost some data in order to get a case of accidental conditioning. Yet, such cases are more common that one might think. Imagine that a group of scientists have tested the effect of a drug among chimpanzees and want to know if it has a similar effect for humans. In this case it would be misleading to have a DAG with variables only for the drug and its effect, since it is clear that one has only measured the effect of the drug in chimpanzees. If one only included the two variables, one would get an unbiased estimate of an effect, but from the graph it would not be clear what quantity was being estimated. In order to make it clear both that we are dealing with a case in which the effect is species specific and where we have only looked at chimpanzees, we should use a DAG with a variable for species in which we have conditioned on the variable SPECIES=chimp. In the extrapolation problem in the example, the scientists are using the effect of the drug on the outcome when one conditions on SPECIES=chimp in order to predict the effect of the drug on the outcome when one conditions on SPECIES=human (Figure 13).

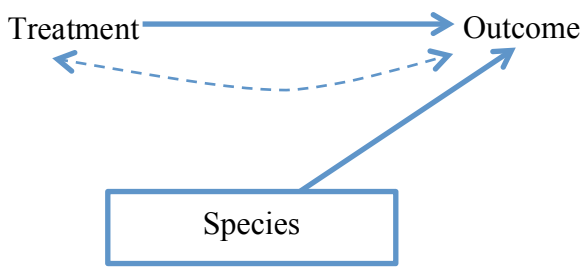

Figure 13 – the species-specific effect of the treatment on the outcome. The box around the species variable denotes that one has conditioned upon it.

In short, given the assumption that one has the probability distribution for a set of variables and the assumption that a particular causal quantity is identifiable, DAGs in principle enable one to measure the average effect across all populations characterized by the variables in the distribution. To represent the possibility that two populations might differ in the specified average causal effects, we need to explicitly represent variables whose distributions differ across the populations.

This way of characterizing extrapolation suggests a general way of representing the problem. Consider the effect of *X* on *Y* given covariate set ***C***. I include ***C*** to convey that it is possible to evaluate very fine-grained effects, but after this paragraph I will leave it implicit. Given two populations that potentially differ in the distribution of variable *Z*, such that the distribution of *Z* in the study population is $D_S(Z)$ and the distribution in the target is $D_T(Z)$, extrapolation concerns whether it is possible to infer the effect of *X* on *Y* given ***C*** and $D_T(Z)$ from the effect of *X* on *Y* given ***C*** and $D_S(Z)$.[20] In the animal experimentation example, *Z* corresponds to the species variable and the problem of extrapolation is that of inferring the effect of the treatment in humans based on the effect in chimpanzees. We can represent this case in a graph, by placing the values of *Z* on the *X*-axis, as I do in figure 14. The different dashed lines indicate possible functions relating *Z* to average treatment effect, where the point corresponding to z_s is the effect in the study population and the effect in the study population is the output of the true

[20] More precisely, extrapolation is the problem of inferring $\sum_z P^*(Y|do(X), \boldsymbol{C}, Z)P(Z)$ from $\sum_z P(Y|do(X), \boldsymbol{C}, Z)P(Z)$, where P^* denotes the distribution in the target and P in the study and the variables are discrete. If the variables are continuous, one can use integrals in the place of Riemann sum operators. Here I represent the full distribution of *Y* given its antecedents in each population, though one is often interested in the probability of *Y* given two specified values of *X* (e.g. $\sum_z[P(Y|do(X=1), \boldsymbol{C}, Z)P(Z) - P(Y|do(X=0), \boldsymbol{C}, Z)P(Z)]$ in each population.

function when $Z=z_t$. I draw dashed lines solely as a means of visually representing possible inferences. I do not yet want to say anything about the conditions under which such inferences are possible.

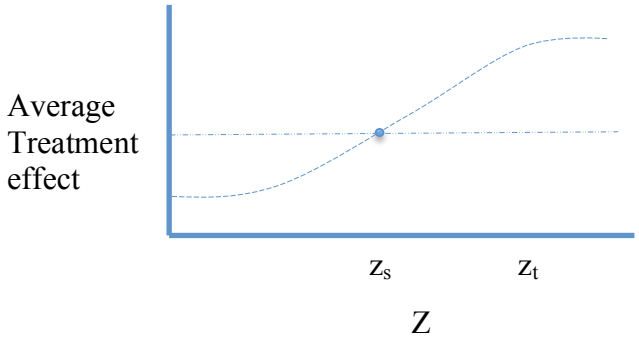

Figure 14

Here's how the characterization of extrapolation just presented relates to transportability. The basis for the transportability framework is that one does not always need to know which variable *Z* is or how it differs across the populations in order to be able to extrapolate the effect of *X* on *Y* given ***C*** across populations. This might be because *Z* does not make a difference in the effect if *X* on *Y*, or it could be because any effect *Z* has is accounted for by the variation it leads to in some other variable that one does measure. Whether it is necessary to know what *Z* is depends on which variables in a DAG are most proximate to *Z* (or any variables that correspond to cross-population differences). Accordingly, one way to think about S-nodes is as denoting the set of variables upon which one has conditioned on in order to get the probability distribution for each of the populations, when one does not know which variables one has conditioned on. All one knows is that the variation in the S-node corresponds to variation in the function determining the value of a particular variable in the DAG.

Formally, transportability is a special case of the problem of extrapolation as I have characterized it. Extrapolation is the ability to infer the effect of *X* on *Y* given variation in the value of *Z.* Transportable causal relationships are those that can be extrapolated given any variable *V* that plays the same role as *Z* in terms of how it corresponds to variation between the populations. As a result, when a causal quantity is transportable, one does not need to know what *Z* is or how it varies between the populations.

I have now argued that transportability is a special case of a more generally characterized extrapolation problem. In the following section, I evaluate two types of extrapolations of non-transportable causal quantities.

6. Extrapolation Beyond Transportability

Consider again the DAG in figure 13, in which the difference in the effect of some treatment on some outcome potentially differs across species. In this case (and any case with the same structure) the effect of the treatment on the outcome is not transportable, since one cannot know the effect in the human population without doing an additional experiment to break the bi-directed arc. An alternative way to present this inference is using a representation similar to that given in figure 14. Given a hypothesis about how the average treatment effect varies as a function of species, we would be able to infer the effect in humans from the effect in chimpanzees. Given the variables we have chosen, representing the inference in this way (figure 15) is not at all helpful. The problem is that the position of each species on the *x*-axis is entirely arbitrary, so we do not have a basis for choosing between different functions.

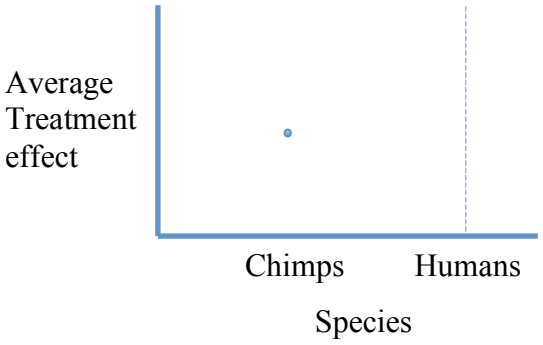

Figure 15

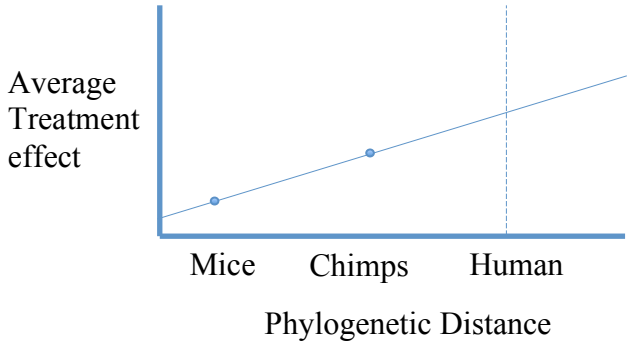

Figure 16

The problem becomes less hopeless if the variable in virtue of the populations differ is in some intuitive sense "well-behaved". In this example, one might try to make the species variable well behaved by arranging species in terms of phylogenetic distance from humans. If species were ranked in terms of phylogenetic distance from humans we might have more confidence that a curve capturing the effect of treatment on outcome in a few species will extrapolate to humans (figure 16).[21] Phylogenetic distance is not an infallible guide, but the hypothesis that more closely related organisms are more likely to respond to a drug similarly will be plausible for some effects. Here I will not attempt to give an account of what it means for a variable to be well behaved". Addressing this question is a desideratum for any "natural kinds" approach to extrapolation".

Even given well-behaved variables, there is a serious question regarding whether extrapolations such as that in figure 16 are ever justified. Figures 14 and 16 are very misleading, since they encourage the viewer to treat causal inference like a type of statistical inference even though the assumptions that ground statistical inference are not justified. The treatment effects in

[21] Steel (2008, 81) considers using phylogenetic distance as a basis for extrapolation in his discussion of simple induction. As far as I can tell, he believes that such inferences might be justified in cases where they shed light on which parts of a mechanism are likely to be similar between related organisms, but that inferences that infer causal similarity based on phylogenetic closeness alone rely upon simple induction and are therefore problematic.

different species are not like balls in urns that one can randomly sample. Moreover, species are distinguished by many factors other than phylogenetic distance and there is no a priori reason to think that these factors do not make all the difference. So we should be wary of using these figures to make any inferences. I use these figures here not to solve any questions, but rather to suggest one way that one might try to go beyond the scope of the transportability framework.

Approaches to extrapolation that seek to extrapolate based on hypotheses about the way that the effect of X on Y given $\boldsymbol{C}$ varies as a function of Z are by their nature *parametric* approaches to extrapolation. As we have seen, the transportability framework (and Pearl's framework more generally) does not make any assumptions about the parametric form of the functional relationship between a variable and its causes. In order to make inferences about the effect of X on Y given Z, one *would* have to make parametric assumptions about the relationship between X and Y and how it varies with Z. Here I have suggested that determining whether such parametric assumptions are justified would require further exploration of issues related to induction and variable selection, though I do not pursue these issues here.

In chapter 7 of the dissertation, I will argue that there are legitimate extrapolative inferences that are both not transportable and that do not require parametric assumptions. As we saw in discussing transportability, the reason that it is possible to extrapolate without making parametric assumptions is that variation in a variable in one part of a causal model does not lead to variation in all parts of a model. In other words, certain parts of a model are *invariant* to changes in other parts of the model. In chapter 7, I will show that there are invariance properties in a model that are missed by selection diagrams and that these enable one to make cross-populations inferences regarding certain non-transportable quantities.

7. Population-to-Subpopulation Inferences and Induction

In addition to the two types of inferences that discussed in the previous section, there is one other type of extrapolation that goes beyond the transportability framework. Namely, knowing that one population is a subpopulation of another licenses certain inference even in cases where transportability fails.

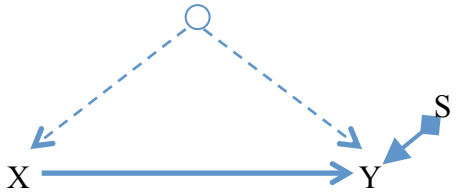

Figure 17

Suppose that one population is a subpopulation of another and that the two populations differ in the value of an unmeasured effect modifier (figure 17). For example, *X* might be general intelligence and *Y* might be one's SAT score. The effect of *X* on *Y* would be the average effect of intelligence on one's score in the population and the S-node may indicate sources of variation in one's test-taking ability. An individual is characterized as a subpopulation with a set of (unknown) effect modifiers whose distribution may differ from that of the population as a whole. In Weinberger (2015), I prove that the effect in a population is a weighted average of the effects in its subpopulations. I take this to show that the effect in the population provides some – possibly weak – evidence for the effect in a subpopulation. Although the effect in a given subpopulation may diverge greatly from that of the population as a whole, it places constraints on what sets of effect magnitudes may obtain across all subpopulations. The effect in the population is therefore evidentially relevant to the effect in a particular subpopulation.

The fact that effects in populations are evidentially relevant to propositions about effects in subpopulations might suggest a more general strategy for developing an inductive approach to extrapolation. When extrapolating from population P_s to population P_t, one can represent the two populations as being subpopulations of some third population P_{super}. Effect magnitudes in P_s provide evidence for effect magnitudes in P_{super}, which in turn provide evidence for the effects in P_t. The flaw in this strategy is that evidential relevance is not, in general, transitive, as we saw in our discussion of Cartwright and Hardie's theory of evidence. Just because facts about P_s provide evidence for facts about P_{super} and facts about P_{super} provide evidence for facts about P_t, it does not follow that facts about P_s provide evidence for facts about P_t.

I have been talking about whether the effect in A is evidence for the effect in B, but most philosophers assume that it only makes sense to talk about evidence relative to a set of background assumptions. If one is willing to make assumptions about features of the distribution of the effect of *X* on *Y* in P_s, P_t, and P_{super}, then there are straightforward evidential relations between the effects in the other two population. The rough idea is that the effect is .2 in P_{super} and one learns that it is .1 in P_s, this raises the probability that the effect will be greater than .2 in P_t. To develop this strategy, one would have to say more about how one is justified in one's beliefs about how the magnitudes of the effects are distributed across the populations. I am skeptical regarding whether it is possible to fill in these details. In non-parametric causal inference, facts about the strengths of causal relationships are not specified a priori, but are estimated from the probability distribution. In this case, however, one does not have frequency data for P_{super} and even if one has frequency data for P_t, the effect might not be identifiable without experiment.

One might reject the demand for evidence regarding one's beliefs about the magnitude of causal effects in different population. Griffiths and Tenenbaum (2009) have pioneered the use of

hierarchical Bayesian models for causal inference. Such models assign prior probabilities both to each model in one's hypothesis space and to the parameterizations of those models. As is typical of Bayesian approaches, one's priors do not require justification. The hierarchical Bayesian approach constitutes a step towards an inductive approach to extrapolation. The limitation of the hierarchical Bayesian approach is that since it requires one to assign probability distributions both to each possible causal model and to each parameterization of each model, it very quickly faces a combinatorial problem of there being more parameterizations than one could possibly evaluate. As a result, the approach is only practical given severe restrictions on the models and parameterizations that are considered. In the paper cited, for example, the authors both assume that one knows that causal ordering of the variables and that there is no effect heterogeneity. As a result, they bypass key steps in causal inference and avoid precisely the cases in which the question of extrapolation would arise.

8. Populations Revisited

In chapter 1, I raised a puzzle about how it is possible to understand populations in a way that allows the study and target populations to have differing causal effects. I will now explain how transportability resolves this puzzle. We can restate the puzzle as an inconsistent triad.

1. Populations can be represented by DAGs with associated probability distributions, and these probability distributions average over all factors not included as variables in the distribution

2. Two populations with the same causal structure differ in the magnitudes of their causal effects only if their background factors have different probability distributions.

3. Two populations with the same causal structure can differ in their average causal effects.

The first proposition relies on the assumption that in non-parametric causal inference, one does not consider populations that are homogenous with respect to all background factors, but rather,

the effects in populations are average effects across all background factors. The causal modeler provides the assumptions under which an average causal effect can be identified and then leaves it to the statistician to do the grunt work of estimating the values in the probabilistic expression that identifies the effect. The second proposition states that in order for two populations with the same structure to differ in their causal effects, they must differ in their background factors. Propositions 3 and 2 jointly entail that for two populations with different structures to differ in the magnitudes of their causal effects, they must have different distributions of background factors. But how is this compatible with the first assumption that causal models average over all background factors? If we discover that two populations with the same structure differ in their effect magnitudes, can't we just blame the statistician for not doing his job?

I take it that we need to accept the third proposition in order not to define the problem of extrapolation out of existence and that we cannot blame the statistician every time extrapolation fails. The second proposition might seem like an appealing target, since it leaves it vague what it means for two populations to have the same causal structure. When I say that two populations have the same causal structure, I mean that they are representable by a single DAG. As I noted in chapter 1, this is a substantive assumption. Yet, I see no way to use DAGs for extrapolation without this assumption, and once one assumes that the variables in the population are instantiated in the same type of system, any difference between them must be due to differences in background factors. If the differences in effects were not due to variation in background factors, in what sense are they representable by the same graph?

To resolve the puzzle, we need to somehow reject the first assumption. More specifically, it needs to be the cases that not all populations with the same structure are defined relative to the same distribution of background conditions. It might seem obvious that they are not. A

psychologist testing the effect of illegal drug use on productivity might be interested in the effect among college students, or among people of a certain socioeconomic status, or among all Americans. She might assume, for example, that the distribution of background factors is different among college students than among other groups and then try to estimate the distribution for college students from her sample. The question remains, however, how it is possible to make the assumption that the study and target population have different distributions without ruling out the possibility of extrapolation. Once one grants that the populations differ in their background factors, how can one infer that they will not correspondingly differ in their effects?

By now it should be clear how the transportability framework answers this question. Extrapolation is possible in the presence of varying background factors because not all variation in background factors leads to variation in an effect of interest. Additionally, even some factors that do lead to variation in the effect do so in a way that does not hinder one from identifying the effect in the target from its probability distribution. The key is that not all sources of variation influence all parts of the model and that by carefully specifying which parts of a model vary as a result of variation in background factors, one can determine which causal relationships will remain invariant across the populations.

9. A Brief Comparison of the Accounts

Pearl and Bareinboim's account differs from both Cartwright and Hardie's and Steel's in allowing for the extrapolation of quantitative causal claims and in providing a general solution to the problem of when one can extrapolate. (While I've argued that there are inferences that go beyond the transportability framework, this in no way diminishes Pearl and Bareinboim's achievement of providing a *general* framework for establishing transportability.) One sign of the

success of the account is that proponents of the alternative accounts have claimed it as a technical elaboration of their preferred account. In a recent chapter in an edited volume (Steel, 2013, p. 194) suggests that Pearl and Bareinboim's S-nodes are an alternative way of referring to what he called "disrupting factors" in his book. Alexandre Marcellesi (2015, pp. (8) #), portrays Pearl and Bareinboim and Cartwright and Hardie as independently developing the same approach to extrapolation. Here I will briefly compare the transportability framework to the others and show how it goes beyond either Cartwright and Hardie or Steel have achieved.

It is not difficult to point to ways in which the transportability framework goes beyond Cartwright and Hardie's. Carwright and Hardie appeal only to parametric structural equations and do not utilize the tools of non-parametric causal inference. They never consider ways of deriving an effect in one population by adjustment. And, if my analysis in chapter 2 is correct, they never provide an example in which one's assumptions actually license an extrapolation. Marcellesi's treatment of the two accounts as if they were the same obscures these limitations of Cartwright and Hardie's framework.

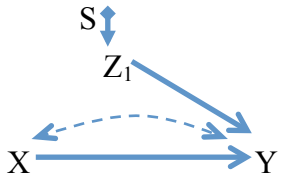

Figure 18

We can think of Cartwright and Hardie's account as considering the selection diagram presented in figure 18. Z_1 is what they would call a support factor for the effect of X on Y, and we can imagine there being other support factors $Z_2,\ldots,Z_n$ with additional S-nodes. Their account says that if we know that the support factors in the study are also present in the target, then we can extrapolate. This would be akin to denying the presence of S-nodes into the different support

factors (and into *Y*). Their account misses the fact that it is possible to extrapolate in the presence of S-nodes (see the appendix for the adjustment formula). Their account also does not generalize in any clear way to causal models more complicated than that in figure 18.

Steel anticipates some of the ideas in Pearl and Bareinboim's framework. His approach uses DAGs and relies on the fact that whether a difference between populations hinders extrapolation depends on where in the model the difference is. While Steel's account primarily concerns qualitative causal claims, in an appendix he gives an example of a quantitative extrapolation that works in the same way that adjustment formulas do.

Despite his anticipations of the transportability framework, this framework is not just a more precise and general way of developing his own. Consider, for example, his claim that S-nodes are the same thing as what he refers to as "disrupting factors":

> In Steel (2008 , pp. 58–62) [S-nodes] are called disrupting factors, while in Pearl and Bareinboim (2011 , p. 6) they are called selection variables . I follow Pearl and Bareinboim's terminology here, as the term "disruption" suggests factors that entirely block a causal relationship, while the differences between model and target could come in other forms. (2013, p. 194)

S-nodes and disrupting factors do not, however, refer to the same thing. A factor *Z* is a disruption factor for *X* just in case *Z* is a cause of *X* such that for some values of *Z, X* no longer depends on its direct causes in a mechanism. It is part of the *definition* of a 'disruption factor' that it entirely blocks a causal relationship. The fact that a disrupting factor plays this role is essential for several of Steel's proofs (e.g. his proof of the extrapolation theorem). Finally, while disruption factors are causes, S-nodes do not need to be. If some variable *M* is an effect modifier of the effect of *X* on *Y* and *S* is an *effect* of *M*, the effect of *X* on *Y* may differ conditional on different values of *S*.

If one views the accounts discussed in chapters 2 and 3 as attempts at producing the transportability framework, they come out looking impoverished by comparison. By considering

the differences among the accounts, we can appreciate some features of the first two accounts that are not found in Pearl and Bareinboim's. The fact that Cartwright and Hardie's account relies on equations with specified parametric forms could be construed as a virtue, since adopting parametric assumptions enables one to make inferences that are beyond the scope of non-parametric causal inference. Additionally, they discuss variable-selection issues that are neglected in Pearl's account. Steel's account is different from the other two in that he provides the most detailed discussion of the nature of the problem of extrapolation. While I have raised concerns about this part of his account, it is a good starting point for further philosophical discussion of extrapolation.

10. Conclusions

Pearl and Bareinboim's present an account of extrapolation that is both precise and general. No matter how complicated a selection diagram is, their methods allow one to determine which quantities are transportable and how to identify the transportable quantities in the target population. In cases where the desired quantity is transportable, I have nothing to add to their account. In cases where a desired quantity is *not* transportable, we require an alternative approach. In what follows, I explore the possibility of using causal mediation techniques to extrapolate non-transportable causal quantities.

Technical Appendix for Chapter 4: Adjustment Formulas

In my informal explication of Pearl and Bareinboim's account, I did not explain how to derive adjustment formulas for transporting causal relationships across populations. Here I present a more precise explication of S-nodes and then give examples of how to derive adjustment formulas.

Let P(y|x) be the probability of *y* given *x* in the study population and let P^*(y|x) be the probability of *y* given *x* in the target population. Given a selection diagram with S-nodes S_1,S_2, …, S_n, the causal relationship P(y|do(x), z) is related across the populations as follows:

$$(1) P^*(y|do(x), z) = P(y|do(x), s_1^*, \dots, s_n^*)$$

In other words, the effect in the target population is the effect in the study population conditional on the values (or distribution) of the S-nodes in the target population.

Equation (1) is correct whether or not $P(y|do(x), z)$ is transportable. This quantity is transportable if and only if it is possible to use the rules of probability and the selection diagram in order to transform $P(y|do(x), s_1^*, \dots, s_n^*)$ into a probabilistic expression in which there is no term that contains both *S* and a do-operator. If it is possible to do so, then the resulting expression is the adjustment formula for transporting the effect to the target population.

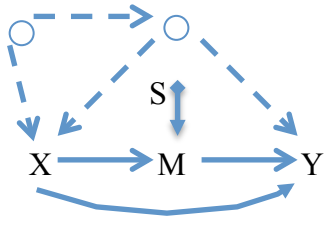

Figure 1
(Figure 9 in Chapter 4)

I will now provide the derivation of the adjustment formula for the effect of *X* on *Y* in the selection diagram in figure 9 from the last chapter. This selection diagram will be important for what follows in the dissertation, since it involves a mediator. The equation for the effect in the target population is

$$(2) P^*\big(y\big|do(x)\big) = P(y|do(x), s^*)$$

It follows from the axioms of probability that the right hand side equals

$$(3) \sum_m P(y|do(x), s^*, m)P(\,m|do(x), s^*)$$

From the graph, it is clear that the causal influence of *S* on *Y* is only transmitted through *M*. *M* therefore screens off S^* from *Y*, allowing us to simplify the first term like so:

$$(4) \sum_m P(y|do(x), m)P(\,m|do(x), s^*)$$

The second term is the effect of *X* on *M*. Since this is unconfounded in the diagram, we can remove the do-operator.

$$(5) \sum_m P(y|do(x), m)P(\,m|x, s^*)$$

Which equals

$$(6) \sum_m P(y|do(x), m)P^*(\,m|x)$$

Equation (6) contains no expressions with both s-nodes and do-operators, revealing that the effect of *X* on *Y* is transportable across populations. (6) is the adjustment formula for identifying the probability of *Y* given do(*x*) in the target population.

Pearl and Bareinboim provide procedures for using selection diagrams in order to determine whether a particular causal quantity is transportable. For further information, see Pearl

and Bareinboim (2013) and Bareinboim and Pearl (2012). The procedures in the latter paper are complete, which means that for any transportable relationship, it is possible to prove that it transportable using the inference rules provided.

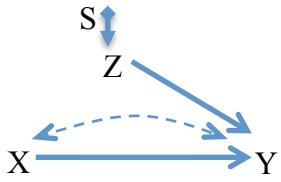

Figure 2

By way of comparison, it will be helpful to compare the adjustment formula derived for the selection diagram in figure 1 with that for another selection diagram. The diagram in figure 2 contains a potential effect modifier, *Z*, for the relationship between *X* and *Y*. The derivation of the adjustment formula for the effect of *X* on *Y* in figure 2 is straightforward.

$$(7)P^*\big(y\big|do(x)\big) = P(y|do(x),s^*)$$

$$= \sum_z P(y|do(x),s^*,z)P(\,z|do(x),s^*)$$

$$= \sum_z P(y|do(x),z)P(\,z|s^*)$$

$$= \sum_z P(y|do(x),z)P^*(\,z)$$

Since *X* and *Z* are uncorrelated according to the in diagram in figure 2, one can remove *X* from the antecedent of the second expression.

In comparing the two adjustment formulas, we see that when adjusting for a modifier *Z*, one can simply adjust for the probability of *Z* in the target population, but in adjusting for mediator *M*, one must adjust for the conditional probability of *M* given *X* in the target population. Later in the dissertation, I will seek to provide an account of the relative advantages of

extrapolating based on mediators as opposed to modifiers. There is a clear sense in which extrapolating using modifiers is simpler. Since extrapolating using mediators requires one to know P(M|X) in the target population, one will not be able to use Pearl and Bareinboim's framework to transport using mediators in cases where this quantity is not identifiable in the target population. On the other hand, figure 2 only allows for transportability because, by hypothesis, one has measured all of the modifiers that differ between the populations. In many scenarios it will be more plausible that one would have knowledge that two populations differ as a result of a difference in a mediator along a particular path than it would be to think that one has knowledge of all the background factors that may differ.

Conclusion to Chapters 2-4

In chapters 2-4, I critically evaluated three accounts of extrapolation. Here I further clarify how these accounts relate to one another and which questions remain open.

I. What Assumptions Should One Make While Extrapolating?

Figure 1 is a selection diagram in which there is an S-node going into every variable and a bidirected arc between every pair of variables. The three accounts considered all agree that given these assumptions extrapolation is not possible. The accounts differ regarding which assumptions must be added in order to enable extrapolations.

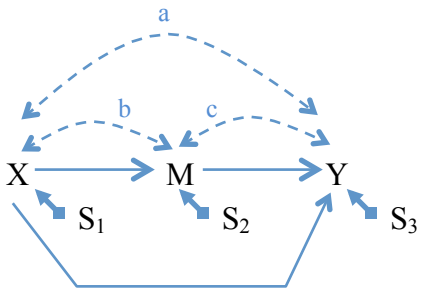

Figure 1

For both Steel and Pearl and Bareinboim, one extrapolates by assuming that at least *some* of the variables in the diagram do not have S-nodes. This assumption is evident in Steel's solution to the extrapolator's circle, which is that one can avoid looking at the complete mechanism (i.e causal path) in the target population by assuming that there are parts of mechanism that do not differ between the study and target populations. Within Pearl and Bareinboim's framework, one denotes this assumption by omitting an S-node from the variables in the mechanism that are presumed not to differ.

To be clear, there *are* cases in which a causal quantity is transportable even though every variable has an S-node. For example, if there were no bidirected arcs, every causal relationship

would be identifiable in the target population from the probability distribution alone and would therefore by trivially transportable. For quantities that are not trivially transportable, one needs to assume that at least some S-nodes are absent in order to achieve transportability.

Cartwright and Hardie's approach does not invoke the assumption that certain S-nodes are missing. Rather, it makes assumptions about the source of variation that is captured by specific S-nodes. Consider their example of an educational program that succeeded in Tennessee, but failed in California. The problem was that California did not have enough teachers for the policy to succeed. In figure 1, the program being implemented is *X,* the success of the program is *Y* and one of the factors corresponding to the variation denoted by S_3 is the number of available teachers. Given the plausible assumption that having enough teachers to implement the program is a necessary condition for it to succeed, it follows that one cannot extrapolate to populations in which this condition is not present.

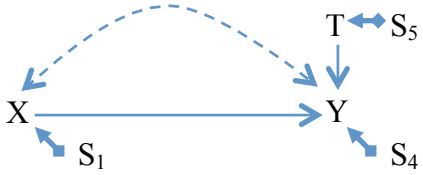

Figure 2

In figure 2, I give a selection diagram that explicitly includes a variable for the number of teachers. There is still an S-node (S_4) going into *Y*, since not all cross-population variation in *Y* is due to the number of teachers. The assumption that having a certain number of teachers is necessary for the policy to succeed is a parametric assumption about the functional form of the relationships between *X*, *Y* and *T*. In contrast, the transportability framework does not rely on any parametric assumptions. The effect of *X* on *Y* is not transportable in either figure 1 or figure 2.

Given the parametric assumption that there are values of *T* for which *X* never has an effect on *Y*, one can infer additional information regarding the conditions under which extrapolation fails.

As we saw, while Cartwright and Hardie provide conditions under which extrapolation fails, the account is much less useful for determining when an extrapolation will succeed. Recall the third premise of the effectiveness argument:

> The support factors necessary for *x* to play a positive causal role here [the target population] are present for at least some individuals here post-implementation. (Cartwright and Hardie, 2013 p. 45)

This is a much stronger assumption than the claim that there are some values of *T* for which *X* never causes *Y.* To establish it, one would need to know about all the factors corresponding to S_4 that are responsible for cross-population variation in *Y* and also about which combinations of factors are sufficient for *X* being positively relevant to *Y.* Moreover, even given such knowledge, one could not establish that the magnitude of the effect of *X* on *Y* will be similarly large in the study and target populations.

Despite the limitations of Cartwright and Hardie's approach, the difference between their approach and the Pearl/Bareinboim approach is noteworthy. The reason that extrapolation is ever possible in non-parametric causal inference is that given the assumption that certain mechanisms are invariant across the populations, one does not need to re-measure every part of the model in the target population. In cases where an effect does vary across populations and one cannot measure this effect in the target population – and where one cannot break it down into further parts that are themselves invariant across the populations or measurable in the target – the effect is not transportable. In principle, if one were justified in making parametric assumptions about how the effect varies across the populations, one would be able to make further inferences about when an effect generalizes. For example, one's parametric assumptions might state that a certain background factor influences the effect of *X* on *Y* for some of its values, but not others.

The following is an open question about extrapolation:

> **Open Question 1:** Are there non-parametric extrapolation-licensing assumptions that are stronger than the assumption denoted by the *presence* of an S-node, but weaker than the assumption denoted by the *absence* of an S-node?

The absence of an S-node is a very strong assumption. It indicates that there are no cross-population differences in the factors responsible for the value of a particular variable. In contrast, the presence of an S-node indicates that for all we know, all of the unmeasured factors influencing a variable vary greatly across populations. Parametric assumptions about how a causal effect depends on unmeasured variables are stronger than the assumption that there is an S-node into the effect variable and weaker than the assumption that there is no S-node into the effect variable. It is an open question whether one can develop a version of transportability that incorporates non-parametric assumptions that are stronger than the assumption that there is no S-node into a variable.

There are several reasons for seeking out specifically non-parametric assumptions. One is that parametric assumptions are often selected based on considerations of computational convenience rather than justified based on beliefs about the causal mechanisms captured by the equations. While there may be cases in which one has causal knowledge that *does* justify parametric assumptions – for example, one might have a theory that entails that two causes of an effect do not interact in producing the effect – one often does not have such knowledge. In such cases, making assumptions about the parametric form of the causal relationships in a model amounts to making a priori stipulations about the probability distribution for a population. Although non-parametric causal inference requires one to make assumptions about the relationship between a causal model and a probability distribution (the most common such assumption being the causal Markov condition), it does not require one to make any further

assumptions about the probabilistic relationships among variables in a distribution. If a causal relationship is identifiable, its magnitude is uniquely determined by the probability distribution.

The second reason for my interest in non-parametric assumptions relates to my goals in the dissertation. As I discuss in the following section, a central aim in the dissertation is to understand the relationship the transportability framework and causal mediation techniques. Both of these methods rely exclusively upon non-parametric assumptions. Since I want to understand the general relationship between these methods, I seek to establish the role the each plays in extrapolation when one does not supplement them with any parametric assumptions.

I take the transportability framework to be the starting point for developing my own account. In any case where a quantity is transportable, it is possible to extrapolate. The question remains as to whether there are cases where one can extrapolate non-transportable effects. The transportability framework is sound and complete, so given the assumptions embedded in a selection diagram it enables one to find all transportable quantities. If there are types of extrapolative inference that involve non-transportable quantities, they must rely on assumptions other than those embedded in selection diagrams. In chapter 7, I present non-parametric assumptions that enable one to go beyond the transportability framework.

II. Learning How Causes Bring About Their Effects

Steel writes:

> The mechanisms approach rests on the intuition that knowing how a cause produces its effect provides a basis for extrapolation. (5)

This dissertation follows Steel in attempting to unpack this intuition. In his extrapolation theorem, Steel gives a condition under which some cause raises the probability of its effect through each mechanism and combination of mechanisms. Given this condition, one can infer

that as long as not *all* of the mechanisms are disrupted, the cause will still raise the probability of its effect.

Here I will not reiterate my concerns about whether the extrapolation theorem is useful for extrapolation. There is nevertheless a feature of the theorem – and of Steel's account more generally – that is worth highlighting. In considering particular mechanisms, Steel only discusses cases in which a mechanism is disrupted and cases in which it functions normally. But these are not the only possibilities. In addition to the ideal interventions that Steel (following Woodward) considers, which fully disrupt the relationship between the variable that one intervenes upon and its direct causes, there are also 'soft' interventions (Korb et al., 2004; Eberhardt and Scheines, 2007) that alter this relationship without destroying it entirely. A full account of the relationship between individual mechanisms and the total effect going through all mechanisms should consider not just cases in which a mechanism is fully disrupted, but also cases in which it is modified by soft interventions.

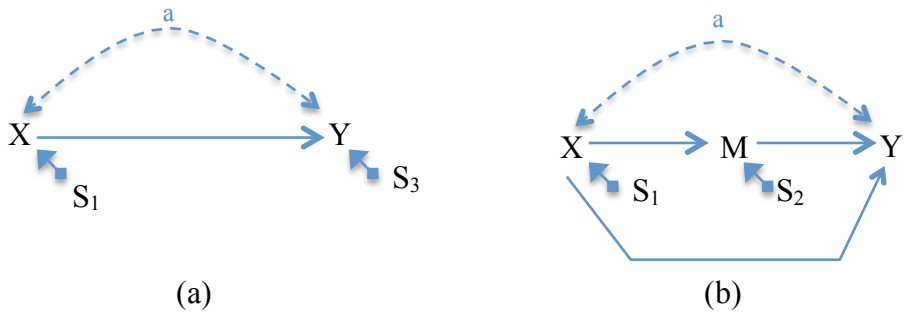

Figure 3

Pearl and Bareinboim's account provides a clear example in which learning how a cause brings about its effect facilitates extrapolation. In figure 3a, the effect of *X* on *Y* is not transportable. The effect *is* transportable in figure 3b. One can replace 3a with 3b given the assumptions that 1) *M* is a mediator between *X* and *Y*, 2) the variation indicated by S_3 in 3a is

entirely due to S_2 in 3b, and 3) bidirected arcs b and c from figure 1 are absent. Later on in the dissertation, I consider the question of whether extrapolation is possible when these assumptions are weakened. Since S_2 corresponds to any arbitrary way that P(M|do(X)) might vary, the selection diagram in figure 3b allows one to transport the effect of *X* on *Y* across populations that differ as the result of soft interventions on *M*.

Cartwright and Hardie provide compelling examples – such as the one with the nurse who can detect a disease – in which knowledge of how a cause brings about its effect helps one extrapolate. They are silent regarding how one gains such knowledge and I argued that in order to do so one needs to use causal mediation techniques. In a case such as the one depicted in figure 3b, these techniques enable one to measure the portion of the effect of *X* on *Y* that goes though *M* as well as the portion that does not. Advocates of these techniques – including Pearl and Bareinboim – have argued that they facilitate extrapolation, but there has been little explicit discussion of how they are supposed to do so.

It might appear that the case in figure 3 validates the claim that mediation techniques aid extrapolation, since it is a clear example in which measuring a mediator enables one to extrapolate. Yet it is possible to show that the effect of *X* on *Y* in figure 3b is transportable without appealing to key mediation concepts such as direct and indirect effects. This brings us to our second open question.

> **Open Question 2:** Do causal mediation techniques facilitate extrapolation and, if so, what is their relationship to the transportability framework?

This will be the central question considered in chapters 5-7.

III. Extrapolation and Induction

It should be clear that the three approaches considered are all deductive approaches. They seek to find premises that guarantee that a causal relation can be generalized rather than considering the general question of when the presence of a causal relationship in one population counts as evidence that the relationship will be similar in another. This is somewhat strange. One would have thought that at least some extrapolative inferences involve induction.

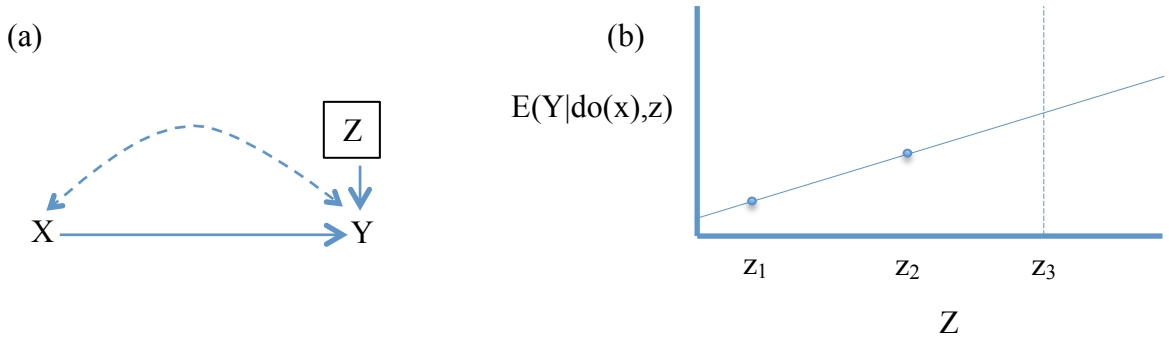

Figure 5

In the previous chapter I proposed a way to represent inductive extrapolative inferences within the causal modeling framework. First, instead of using S-nodes, one must explicitly represent variables that may make a difference in the effect of *X* on *Y*. In figure 5a, *Z* is such a variable. Second, one must represent *Z* in the study population as if it has been conditioned upon. The reason for this is that in problems of extrapolation one cannot presuppose that the joint distribution for the study population is also the joint distribution for the target population. But a standard assumption in causal modeling is that for a given DAG one does know the joint distribution of the variables for all possible combinations of the variables.[22] By conditioning on *Z*, one represents the situation in which one's distribution is known to be correct only for certain values of *Z*. Extrapolation involves inferring the effect of *X* on *Y* for unobserved values of *Z* given observed ones (figure 5b). Figure 5b contains a solid line corresponding to simple way that

[22] In practice, this assumption (which is called 'positivity') means that one cannot do causal inference when not all combinations of variables actually appear in one's sample.

one might infer the value of E(Y|do(x),z_3) but the figure should not be taken as an endorsement of this inference.

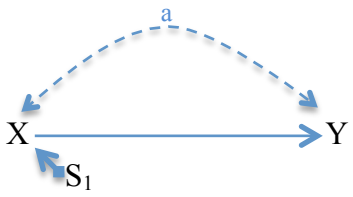

Figure 6

My analysis of extrapolation as the problem of inferring the effect of *X* on *Y* given *Z* for populations with unobserved values of *Z* makes precise the sense in which the problem of extrapolation has both deductive and inductive aspects. In figure 5b, the simple inference from the effect given z_1 and z_2 to the effect given z_3 resembles a statistical inference. Yet, the only reason that one can treat variation in P(Y|do(x), Z) as variation in the causal effect of *X* on *Y* is that *Z* is a cause of *Y* that can make a difference in the effect of *X* on *Y*. One does not treat the effect magnitudes in different populations as balls that are being randomly sampled out of an urn. The effect magnitudes are hypothesized to vary as a result of some variable playing a specified causal role and one seeks to learn the function indicating how they vary. The deductive accounts considered in the previous chapters are correct that extrapolation is only possible given a set of causal assumptions about how the populations differ.

While I have suggested that there is a way to represent extrapolative inferences that have inductive aspects, I have not said much about what would justify such inferences. This remains an open question:

> **Open Question 3:** Are inductive extrapolative inferences ever warranted? If so, what assumptions license such inferences?

Unfortunately, this question will still remain open at the end of the dissertation.

IV. Where This is All Going

Although Steel considers his approach to extrapolation to be a hybrid between structural causal models and contemporary philosophical accounts of mechanisms, his results do not require him to adopt any of the characteristic commitments of mechanistic accounts. Recent mechanistic accounts analyze the way that the components of a mechanism come together to bring about a phenomenon and presuppose that the relationship between a mechanism and its components calls for a non-causal form of explanation. Since I am similarly concerned with the means by which an effect comes about and my approach only invokes causal relations, in the following chapter I consider the objection that my account omits some important non-causal feature of mechanistic explanation that is important for extrapolation. I argue that the features of mechanisms that allegedly require non-causal forms of explanation can be adequately accounted for using causal mediation techniques.

After a somewhat informal introduction to mediation techniques in chapter 5, chapter 6 provides a thorough formal introduction to them. Chapter 7 then addresses the question of how causal mediation techniques relate to transportability, thus answering Open Question 2. Chapter 7 also answers the first open question. It turns out that direct and indirect effects can be extrapolated given certain non-parametric assumptions about how populations vary, and these assumptions cannot be represented in selection diagrams.

Chapter 5: Do Mechanisms Call for Non-Causal Explanation?

Mechanisms are currently a hot topic in philosophy of science. An example of a mechanistic phenomenon is a neuron's firing. Neuroscientists studying this phenomenon seek to explain it by describing the way that the components of a neuron are functionally organized to produce the firing. Proponents of theories of mechanisms (henceforth, "mechanists") have argued that existing theories of explanation cannot account for what scientists are doing when they invoke a mechanism to explain a phenomenon. While mechanists often contrast their accounts with law-based accounts of explanation, here I consider the relationship between mechanistic and *causal* explanations. Since the components of a mechanism are causally related to one another, mechanistic explanations are in part causal. But mechanists aim to provide a novel form of explanation that is distinct from causal explanation. This chapter raises doubts regarding the thesis that mechanisms call for a non-causal form of explanation.

Mechanists would likely be sympathetic to the mediation approach to extrapolation that I develop in the dissertation. Mechanists commonly claim that knowledge of a mechanism enables one to determine the conditions across which a particular phenomenon will continue (or fail) to be produced. The mediation approach provides a promising way of spelling out how knowledge of underlying mechanisms allows one to do this. Yet, the approach here is entirely grounded in facts about the causal relationships among variables. If mechanists were right that mechanistic explanations require one to appeal to some non-causal relationship, my account would be vulnerable to the criticism that it neglects this important relationship, which may be essential for understanding how mechanisms aid extrapolation. In this chapter, I justify my decision to provide an account of extrapolation that only appeals to causal relationships. Mechanists have failed to demonstrate that mechanistic phenomena require non-causal explanation and the cases

they use as a basis for believing that they do require such explanation are better understood using the causal mediation techniques that I sketch here and defend in subsequent chapters.

1. Overview of the Argument

Carl Craver's account (2007) contains the best-developed attempt to identify and explicate the non-causal element in mechanistic explanations. According to Craver, while the relations among individual components are causal, the relationship between the components and the mechanism as a whole is *constitutive*. He argues that this constitutive relationship is an explanatory one; the behavior of a mechanism both explains and is explained by the behaviors of its components. Craver explicates this symmetric relationship as follows. It is possible both to change the behavior of the mechanism by manipulating its components and to change the behavior of the components by manipulating the mechanism as a whole. By "manipulations" Craver means the ideal experimental interventions that Woodward (2003) defines. In developing his account, Craver is guided by the "inter-level" experiments that neuroscientists use to determine whether an entity is a component in a mechanism. In top-down experiments one observes the behavior of the entity in an undisrupted mechanism. In bottom-up experiments one alters the behavior of the entity to see how doing so influences the behavior of the mechanism.

Craver provides two arguments for the claim that the constitutive relationship is non-causal. First, since components are spatiotemporal parts of a mechanism, the components cannot be causally related to the mechanism. Second, since causal relations are asymmetric and the explanatory relation between the mechanism and its components is symmetric, constitutive explanations must be distinct from causal explanations. I reject the first argument on the grounds that, contrary to appearances, the fact that components are parts of a mechanism plays no explanatory role in Craver's account. I reject the second one on the grounds that the

manipulations in top-down and bottom-up experiments appear symmetric only because Craver does not properly specify the variable that is intervened upon in top-down experiments.

Although Craver fails to show that mechanistic explanations are non-causal, the inter-level experiments he discusses do suggest a possible limitation of causal explanation. In top-down experiments, one learns how a mechanism behaves when it is undisrupted. Bottom-up experiments reveal the causal relations among components, but only at the cost of disrupting the mechanism. The fact that bottom-up experiments are insufficient for establishing whether an entity is a component may suggest that causal relations alone are unable to account for the behavior of the undisrupted mechanism, and that they must therefore be supplemented with a top-down explanatory relation. But this limitation on causal explanation is only apparent. *Causal mediation techniques* enable one to use sets of joint interventions to determine the relationship between contributions of variables acting in disrupted systems and the role that these variables play in functioning systems. Given multiple causal paths between a cause and effect, these techniques allow one to determine, for example, the probability that the cause would still be sufficient for the effect were one of the paths to be fully disrupted. These techniques appeal only to causal relations among variables and therefore block the proposed threat to the sufficiency of causal explanations.

The rest of the chapter is organized as follows. Section 2 provides a brief summary of the recent literature on mechanisms. Section 3 outlines Craver's account of constitutive relevance and argues that the condition that a component must be a part of a mechanism is redundant. Section 4 argues that Craver's mutual manipulability condition does not imply a symmetric explanatory relation. Section 5 introduces causal mediation techniques and shows how they

undermine a possible argument that inter-level experiments call for non-causal explanation. Section 6 considers the relevance of mediation techniques to explanation. Section 7 concludes.

2. Mechanistic Accounts of Explanation

In "Thinking about Mechanisms," Machamer, Darden and Craver define mechanisms as:

> [E]ntities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. (2000, 3)

An example of a mechanism is a neuron's firing. When a neuron fires, it increases and then decreases in voltage. These voltage changes result from sodium and potassium ions moving across the cell membrane, thereby changing the proportions of sodium and potassium inside and outside the cell. To explain this process, one must identify the properties of the ion-channels – the entities that regulate the movement of ions across the membrane – and determine how they perform their functions (their "activities").

The entities that are organized to bring about the activity of a mechanism are its components, and one mechanism can be a component in a larger one. This suggests a hierarchical ordering of the world in which a whole mechanism counts as one level and its components are at a lower level. The concept of a mechanism level is distinct from other level-concepts in the literature such as levels of size (macro/micro), levels of abstraction and levels of properties (first-order/second-order etc.). While it is controversial whether second-order properties can have causal powers other than those of their first-order realizers (Kim, 2000), mechanisms can have effects that do not reduce to those of their components.

Craver and Bechtel (2007) argue that the relationship between mechanism levels is non-causal on the grounds that causes and effects must be spatially and temporally distinct. Since components and mechanisms stand in a part/whole relation, they cannot be causally related. The relationship between a mechanism and its components is *constitutive* rather than causal. There is

more to being a component of a mechanism than being a part of it. As Craver notes (2007, 4), a car's hubcaps are a part of it, but they are not components of its mechanism.[23] An account of constitutive relevance should specify how a part must contribute to the activity of a mechanism in order to count as a component.

Craver (2007) seeks to provide a precise account of mechanistic explanations in neuroscience. He refers to the mechanism as *S* and its components as $X_1, X_2 \ldots X_n$. *S*'s activity is denoted by Ψ ("psi") and the activities of *X*'s are denoted by $\varphi_1, \varphi_2 \ldots \varphi_n$ ("phi-1" etc). A neuron firing is an *S* that Ψs. A sodium-ion gate opening is an *X* that φs. In figure 1, the relationships among the φing *X*'s are causal and the relationship between the φs and Ψ is constitutive.

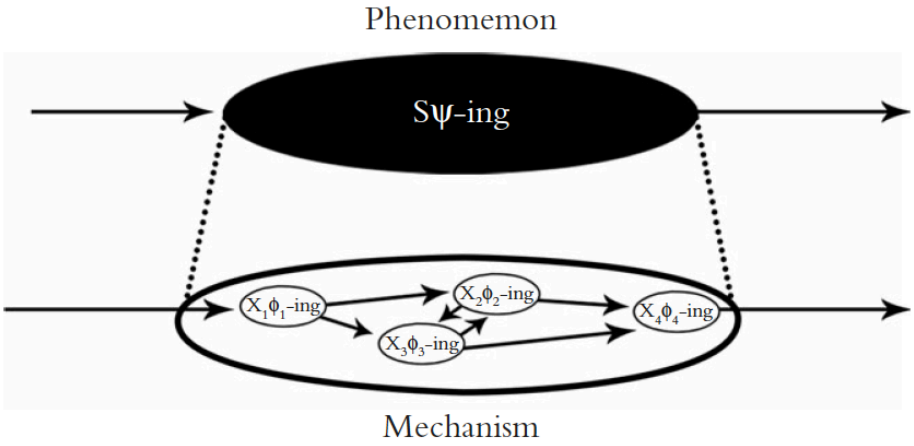

Figure 1 (from Craver, 2007, p. 7)

It is important to distinguish between the analytic truth that components are parts of a mechanism and the substantive claim that there is an explanatory relationship between them. Rather than asserting merely that there is a part-whole relation between an entity and its components, Craver maintains that components explain – and are explained by – the behavior of the mechanism.

[23] Below I raise concerns about this example.

3. Craver's Account of Constitutive Explanation

In this section, I outline Craver's (2007) account of constitutive relevance and argue that the condition that a component must be a part of the mechanism does no explanatory work in the account. Before I do so, it will help to sketch some of the experiments that motivate the account.

The following example comes from Craver (2002) and concerns the mechanism of spatial memory. The preliminary evidence that the hippocampus plays a role in spatial memory was that it activates in rats as they navigate a maze. These preliminary experiments are "top-down" – one intervenes on the rat's activity by placing it in a maze without intervening on any of the neural components involved in its navigating the maze. By themselves, these experiments do not establish that the hippocampus contributes to spatial memory. It may be that the hippocampus is activated whenever the rat runs the maze, but that it in no way facilitates the rat's navigation. To eliminate this possibility, one must intervene on the hippocampus to see if doing so affects the rat's maze-navigating ability.

Craver describes two types of experiments on the hippocampus: interference and stimulation. Interference experiments create lesions in the rat's hippocampus, with the result that the rat's ability to complete the maze is impaired. These experiments provide inconclusive evidence that the hippocampus contributes to spatial memory. It is possible that in creating the lesion, one impaired other parts of the brain and that these parts of the brain are responsible for spatial memory. Experiments in which one stimulates the hippocampus help eliminate this possibility. If administering an electric shock to the hippocampus alters the rat's maze-running ability, this bolsters the hypothesis that the hippocampus is a component in the mechanism for spatial memory. Both interference and stimulation experiments are "bottom-up." One intervenes on a potential component to see how doing so affects the rat's maze-running ability.

In developing his account of constitutive relevance, Craver pays careful attention to the details of top-down and bottom-up experiments, which he refers to as *inter-level experiments.* The methodology of these experiments is embodied in his *mutual manipulability criterion,* which says, roughly, that if it is possible to change a mechanism's behavior by intervening on an entity and to change the behavior of that entity by intervening on the mechanism, then that entity is a component in the mechanism. I consider this criterion in the following section. In addition to the mutual manipulability criterion, Craver stipulates that in order for some entity x to be a component of mechanism S, x must be a part of S. I will refer to this as the *part-hood criterion.*

The part-hood criterion appears to be redundant, since it is unclear how an entity could meet all other conditions for being a component of a mechanism, yet fail to be a part of it. One might think that in stipulating that components must be parts of the mechanism, Craver is requiring that they be spatially contiguous with the other components in the mechanism. He denies, however, that one can draw the boundaries of a mechanism using spatial criteria (Craver, 141 ff.). In the absence of some such criterion, however, the requirement that an entity be a part of a mechanism does no work in determining what counts as a component.

In fact, contra Craver, it does not make sense to distinguish between component and non-component parts of a mechanism. Consider his go-to example for illustrating this distinction:

> The hubcaps, mud-flaps, and the windshield are all parts of the automobile, but they are not part of the mechanism that makes it run. (140)

Upon inspection, this example does *not* show that it is possible for an entity to be part of a mechanism without being a component. What it shows is that it is possible for an entity to be a part of a *machine* without being a component in any of its mechanisms. Accordingly, a hubcap is

a part of a car, but is not a part of the mechanism for acceleration.[24] An entity can be part of a machine without being a component in any mechanism, since it might not contribute to the functioning of the machine. If, however, an entity is a part of a mechanism, then it is a component in that mechanism.

Craver repeatedly distinguishes between causal and constitutive explanations on the grounds that components are parts of a mechanism and there cannot be a causal relation between entities are not spatially distinct. If the part-whole relation played a role in explaining mechanistic phenomena, it would follow that such phenomena cannot be given an exclusively causal explanation. In Craver's account, however, the requirement that a component be a part of the mechanism is redundant, so he cannot use it to motivate the need for a non-causal form of explanation.

4. Inter-level Experiments and Mutual Manipulability

Since the part-hood criterion is redundant, Craver's account of constitutive relevance rises or falls with the mutual manipulability criterion. This criterion attempts to formally represent the interventions in inter-level experiments. In order to model these interventions Craver (2007) relies on Woodward's (2003) concept of an ideal intervention.

We already encountered Woodward's account in our discussion of Steel in chapter 3, but here is a brief review. Woodward (2003) presents an account of causation on which there is a causal relationship between *X* and *Y* just in case it is possible to change the value of *Y* by changing the value of *X* in some "appropriate" way. Woodward defines the concept of an ideal intervention in order to clarify which changes to *X* are appropriate for evaluating whether *X*

[24] One could argue that hubcaps *are* components in one of the car's mechanisms, namely the mechanism for keeping the end of the axle clean. This ambiguity in what counts as part of a car's mechanism presents a further reason for analyzing mechanisms rather than machines.

causes *Y*. To see that not all ways of changing *X* are appropriate, imagine that *X* is being drowsy and *Y* is having a headache. Suppose that I give you a pill that is a common cause of your being drowsy and of having a headache. Clearly, such an intervention would not show that there is a causal relationship between being drowsy and having a headache. To evaluate whether there is, I would have to find a way to make you drowsy without also influencing whether you have a headache (except, perhaps, *via* making you drowsy). More generally, an *ideal intervention* on *X* with respect to *Y* sets the value of *X* in such a way that any effect of the intervention on *Y* is transmitted through *X*.

An important feature of Woodward's account is that ideal interventions on *X* are *surgical.* That is, the only variable they directly influence is *X*. If an intervention is ideal, then the value of *X* is determined entirely by the intervention and all other changes to the model result from this intervention. Many actual interventions do not determine the value of a variable, but alter its probability distribution. Such interventions are "soft" interventions (Korb et al., 2004; Eberhardt et al., 2007). Unlike Woodward's "hard" interventions, soft interventions do not break all of the arrows going into a variable. As far as I can tell, one can always model a soft intervention on *X* as a hard intervention on a cause of *X* (though doing so might require adding variables to the model). Both hard and soft interventions are surgical and it is therefore straightforward to expand Woodward's account to allow for soft interventions.

According to Woodward, *X* is causally relevant to *Y* if there are some interventions on *X* that change *Y*. It is not required that *every* intervention on *X* changes *Y*. Craver accepts Woodward's account of causal relevance, but proposes an alternative account for constitutive relevance. According to Craver, x is constitutively relevant to S if (1) x is a part of S and (2) x's φ-ing and S's Ψ-ing meet the following conditions:

(CR1) When φ is set to the value φ_1 in an ideal intervention, then ψ takes on the value $f(\varphi_1)$. (155)

(CR2): if ψ is set to the value ψ_1 in an ideal intervention, then φ takes on the value $f(\psi_1)$. (159)

$f(\varphi_1)$ and $f(\Psi_1)$ are, of course, different functions. CR1 and CR2 are what I have been calling the mutual manipulability criterion. CR1 corresponds to bottom-up experiments and CR2 corresponds to top-down experiments. Regarding spatial memory, CR1 says that if the hippocampus is a component in spatial memory, then intervening on the hippocampus will influence the rat's maze-navigating ability. CR2 says that intervening on spatial memory will alter the activity of the hippocampus. According to Craver, neither principle individually is necessary or sufficient for φ being a component of Ψ. They are jointly sufficient. Where only one condition is met Craver offers no general guidelines, but says that we need to look at the details of the case (159).

The major problem with CR1 and CR2 is that Ψ does not refer to the same variable in each.[25] It is clear enough what an intervention on Ψ is. An intervention on Ψ sets the mechanism in motion. For example, placing the rat in the maze. One can think of an intervention on Ψ as an intervention on an input into Ψ (or a cause of the earliest φ in the mechanism). Since Ψ refers to the activity of the mechanism, one might be inclined to think of an intervention on Ψ as an intervention on the whole mechanism, but this is just to speak loosely. In order for a top-down experiment to do its job, one must *not* intervene on any intermediate components of the mechanism. Intervening on components of the mechanism would prevent the researcher from seeing how the mechanism behaves when it is undisrupted – that is, when one does not intervene upon its components.

While CR2 refers to an *intervention* on Ψ, CR1 refers to the *value* of Ψ. In the spatial memory example, the value of Ψ presumably denotes either whether the rat completes the maze

[25] Several philosophers have also noted that the meaning of Ψ is ambiguous (Fagan (2012); Menzies (forthcoming); Franklin-Hall (unpublished)).

or how far through the maze the rat gets. This is a different variable from the one that is intervened upon. Recall that an ideal intervention determines the value of a variable. An intervention that triggers a process does not determine whether the process runs to completion. Whether the process runs to completion depends not just on the triggering intervention, but also on the proper functioning of the mechanism's intermediate components. The variable denoting whether the mechanism runs successfully cannot be the same as the variable upon which one intervenes, since the latter variable does not determine the value of the former variable. Nor would it help to go beyond Craver's treatment and consider "soft" interventions that only set the probability of a variable. A soft intervention that triggers the mechanism will not determine the probability that it will run to completion, since this probability depends on whether there are interventions on the mechanism's components.

We can denote the variable that one intervenes upon as Ψ_{input} and the variable corresponding to the success of the intervention as Ψ_{output}. CR1, properly understood, says that intervening on φ alters the value of Ψ_{output}. If an entity is a component, then intervening on it influences whether the mechanism runs to completion. CR2, properly understood, says that intervening on Ψ_{input} alters the components of the mechanism. If one does not distinguish between Ψ_{input} and Ψ_{output}, it appears that one can both change Ψ by intervening on φ and change φ by intervening on Ψ. Given Woodward's account of causation, CR1 and CR2 entail that Ψ_{input} is a cause of $\varphi^{(26)}$ and φ is a cause of Ψ_{output}, but each of these relations is asymmetric. Craver's mutual manipulability criterion only *appears* to explicate a symmetric explanatory relation.

One might object that Ψ *can't* be decomposed into Ψ_{input} and Ψ_{output}, since Craver's account requires that *x*'s φ-ing be a part of *S*'s Ψ-ing. In other words, components must be *parts*

[26] Or at least of those φ that are distinct from Ψ_{input}.

of the mechanism, but a mechanism's components are *not* a part of its input or output conditions. But it is irrelevant whether components of Ψ are parts of Ψ_{input}. Interventions on "Ψ" are interventions on inputs into the mechanism, not on the operation of the mechanism. If Ψ denotes the whole mechanism, then CR1 and CR2 consider variables other than Ψ.

Craver faces a dilemma. If Ψ denotes the whole mechanism, it is not a variable that one can intervene upon, and he has failed to clarify the sense in which it is possible to change a mechanism's components by intervening on the whole. If Ψ denotes a variable upon which one can intervene, then Craver's allegedly symmetric non-causal relationship dissolves into two asymmetric causal ones. Either way, he has failed to provide an account of the symmetric relation that allegedly distinguishes mechanistic explanations from purely causal ones.

Bert Leuridan (2012) similarly argues that Craver's allegedly non-causal explanatory relation is in fact a causal one. His conclusion, however, is not that there is no symmetric explanatory relation, but that the symmetric explanatory relation that Craver describes is in fact causal (he is thinking of cases of bidirectional causation). I am drawing a stronger conclusion. In my view Craver fails to explicate any symmetric relation and any non-causal relation.

The purpose of a top-down experiment is to learn how a mechanism functions when it is undisrupted. Doing so requires an intervention. A scientist needs to have control over when the rat enters the maze in order for her to conclude that the correlation between the rat's running the maze and the observed activity in the hippocampus is not due to a common cause. But this is not enough. In addition to an ideal intervention on the input of the mechanism, top-down experiments require that the scientist *not* perform any further intervention that would prevent the mechanism from running to completion. In contrast, in bottom-up experiments one intervenes on a component in order to make the mechanism behave abnormally. This way of understanding

inter-level experiments is compatible with much of what Craver says when speaking informally. It also helps explain why one would appeal to a symmetric explanatory relation. If experiments that disrupt the behavior of candidate components cannot account for the behavior of the undisrupted mechanism, this suggests that a purely bottom-up explanatory approach will not succeed. In the following section I explain why this line of reasoning fails. The relationship between a functioning whole and the behavior of its parts can be analyzed using models that only invoke causal relations.

5. Causal Mediation Techniques

An ideal intervention on a component disrupts any previously existing relationship between the component and its causes. The results of such an intervention might not correspond to the effects that the manipulated component would have had in an undisrupted mechanism. This is why bottom-up experiments cannot by themselves reveal the behavior of a mechanism. At first glance, this points to a limitation of causal explanation – at least if we follow Woodward in explicating causal claims in terms of ideal interventions. If causal relations are explicated using ideal interventions and ideal interventions are unable to uncover the behavior of a mechanism, then mechanistic phenomena appear to require some non-causal form of explanation.

Melinda Bonnie Fagan (2012), a proponent of non-causal theories of mechanistic explanation, presents an argument similar to the one I just suggested. She faults Craver's account for being *too* causal. Although he denies that the constitutive relation is causal, he nevertheless explicates it in terms of interventions. Yet the interventionist account allegedly fails to explicate the following feature of mechanisms:

> [T]he behavior of isolated components is not a good guide to their behavior together, and their behavior in one context is not a good guide to their behavior in others. (462)

Fagan proposes an alternative account, on which mechanistic explanations work in part by specifying the properties in virtue of which two interacting components bring about an effect. For my purposes here, the details of her account are less important than its motivation, which is her belief that interventionist frameworks are unable to account for the contributions of components to the whole.

Contrary to appearances, one *can* use ideal interventions to model the relationship between the behavior of components in isolation and their behavior in an undisrupted system. One can do so using Judea Pearl's (2001, 2012) causal mediation techniques. These techniques provide conceptual resources for measuring the contributions of distinct causal paths between two variables. Pearl's treatment of mediation differs from earlier attempts (e.g. Barron and Kenny 1986) in that it allows one to model systems in which causes do not contribute additively to their effects. His techniques enable one to consider causal relations not in isolation, but in terms of to their contribution to a broader system. Causal mediation techniques rely on the DAG framework developed by Spirtes, Glymour and Scheines (2000) and Pearl (2009), which I introduced in chapters 2-4.

Given the importance of interaction in mediation techniques, I would like to remind the reader that the presence of two causal arrows from variables *X* and *Y* into an effect *Z* does not denote that the effects of *X* and *Y* on *Z* are additive and separable. The value of *Z* is a *function* of its direct causes, and this function can have any form, including one in which the effect of either cause on *Z* depends on the value of the other cause. In such a case *X* and *Y interact.* The distinct arrows do not indicate the independence of the *causal contributions* of *X* and *Y* to *Z*, but rather the possibility of separately *intervening* on *X* and *Y*.

We are now ready to discuss causal mediation techniques. Consider the following example. You tell me to wake up at 7 a.m. tomorrow, and this causes me to set an alarm. My setting an alarm is a cause of my waking up at 7. Furthermore, suppose that your command makes me stressed and that my being stressed makes it more likely that my alarm will wake me up when it goes off. If I don't set an alarm I will not wake up at 7, regardless of whether I am stressed. Given that I do set the alarm, my being stressed makes it more likely that I will wake up at 7. We can model this case using variables corresponding to your *command,* my setting an *alarm* and whether I *wake up* (figure 2). Although there is presumably some physical mechanism that explains how your command causes me to wake up by making me stressed, there is no variable in the model corresponding to any part of this mechanism. (It would, of course, be possible to formulate a more complicated model that contains a stress variable.)

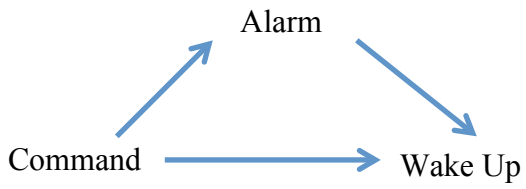

Figure 2

In our example, the *command* and *wake up* variables are referred to as the *treatment* and *outcome*, respectively. *Alarm* is the mediator. Here I will only consider one-mediator models. The effect of the treatment on the outcome going through all paths is the *total effect.* The effect of the treatment on the outcome not going through any specified mediator is the *direct effect;* the effect of the treatment on the outcome going through the mediator alone is the *indirect effect*. These concepts will require further clarification. Direct and indirect effects are model-relative. As the direct effect is the effect of the treatment on the outcome not going through any specified mediator, models with different mediators will have different direct effects.

In our example, your command increases the probability of my waking up via a path not going through the mediator. Nevertheless, this path only makes a difference when I set my alarm. When I do not set my alarm, my being stressed has no influence on whether I wake up. Whether your command directly influences my waking up therefore depends on the value of the mediator. If one did not already know the correct causal structure for this case, it is unclear how one could figure out whether there is an arrow from *command* to *wake up* using interventions. An intervention in which you tell me to wake up and I do wake up is compatible with there being only one path from *command* to *alarm* to *wake* up. If you then intervene by disabling my alarm, I won't wake up. This would suggest that there is no direct path. In this example, your "top-down" and "bottom-up" experiments do not suffice to find a causal relation that, by stipulation, exists.

In our example, causal mediation techniques enable us to answer the following questions:

1) If there were no path going through the mediator, what is the probability that your command would be *sufficient* for my waking up?

2) If there were no path going through the mediator, what is the probability that your command would be *necessary* for my waking up?

3) If there were *only* the path going through the mediator, what is the probability that your command would be *sufficient* for my waking up?

4) If there were *only* the path going through the mediator, what is the probability that your command would be *necessary* for my waking up?

It is important to distinguish between 1 and 2. In this example, the direct path is never sufficient for my waking up, but in some cases it is necessary. Since my being stressed raises the probability that I wake up, over a sufficiently large number of trials there will be cases where I would not have awakened in the absence of the stress. If one knows the *total effect* of your command on my waking up, one can derive the answers to 3 and 4 from the answers to 1 and 2. For example, suppose that the answer to 1 is that the direct path is never sufficient and that the

total effect of *command* is to raise the probability of *wake up* by 80%. The answer to 4 is that the indirect path is necessary for my waking up in 80% of all cases, since in 80% of all cases it is the case that I wake up and that I would not have woken up if your command did not cause me to set an alarm. By dividing this by the total effect (80/80) we get the result that in 100% of the cases where your command caused me to wake up, the indirect path was necessary for this to occur. Note that the *sufficiency* of one path is inversely related to the *necessity* of the other. The claim that the indirect path was necessary in 100% of the cases where your command caused me to wake up entails that the direct path was sufficient in 0% of these cases.

The first accomplishment of causal mediation techniques is to distinguish between the counterfactuals just considered and to map the logical relations among them. The direct effect (DE) indicates whether the direct path is *sufficient* for bringing about an outcome; the indirect effect (IE) indicates whether the *indirect* path is sufficient. As I will further explain in chapter 6, the portion of the total effect (TE) for which the direct effect is sufficient is given by DE/TE and the portion for which IE is sufficient is given by IE/TE. The portion of the total effect for which the direct effect is *necessary* is given by 1 – IE/TE, and the portion for which the *indirect* effect is necessary is given by 1 – DE/TE. In non-additive systems, the total effect is *not* the sum of the direct and indirect effects, but rather is divided between a portion for which one path is sufficient and a portion for which the other is necessary.

For reasons that I cannot address until I introduce the appropriate notation in chapter 6, the total effect is equivalent to the direct effect of introducing the treatment *minus* the indirect effect of removing the treatment. While the total effect does not decompose into the sum of DE and IE, it is possible to decompose it into the contributions of the paths.

I still need to present the interventions corresponding to DE and IE, but a few points should be clear already. First, causal mediation techniques do not work by analyzing the contribution of each causal arrow in isolation and then combining them to yield the total effect. The direct and indirect effects are defined to give them a clear-cut relation to the total effect. Second, in evaluating the contribution of a path to the total effect, it is important to distinguish between the questions of whether it is necessary and whether it is sufficient. Your command is directly relevant to my waking up, even though the direct path is never sufficient for my waking up, but only (in some cases) necessary.

It might seem trivial to measure the direct effect. Simply intervene on the treatment while simultaneously intervening on the mediator to disrupt the indirect path. This approach is not adequate, since in systems with interaction, the direct effect of the treatment on the outcome may depend on the value of the mediator. There will therefore be as many direct effects as there are values of the mediator. In our example, there are only two values of *alarm* (on, off), so we now need to move to a more complicated case to make this problem more salient.

Imagine that scientists develop a drug to reduce cholesterol. The drug has the intended effect, but unfortunately it also increases blood pressure. Worse, the scientists suspect that the drug is more effective at reducing cholesterol in people with higher blood pressure. They consider developing an auxiliary drug that blocks the effect of the cholesterol drug on blood pressure. Such a drug would only be worthwhile, however, if the cholesterol drug could still sufficiently reduce cholesterol without increasing blood pressure. In other words, the auxiliary drug is only worthwhile if the cholesterol drug has a non-negligible direct effect (figure 3).

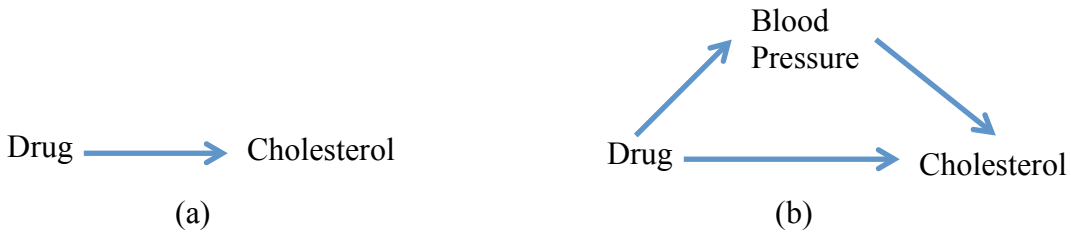

Figure 3 – in (a), the arrow represents the total effect of the drug on cholesterol. In (b), the arrow between *drug* and *cholesterol* indicates that *drug* influences *cholesterol* through a path not going through *blood pressure.*

Suppose that members of the drug trial would all have high blood pressure if they take the drug and medium blood pressure if they do not. Were the scientists to set everyone's blood pressure to *low* while intervening to give trial members the cholesterol drug, this would tell them the direct effect of the drug on cholesterol for low-blood-pressure individuals. This, however, is not the quantity of interest. It is not the direct effect for *the participants in the original study.* The direct effect for *these* individuals corresponds to the effect that the drug would have on their cholesterol were they to have the blood pressure that they *would* have had were they not to receive the treatment. Two interventions are required to measure this. First, one must assign some participants to take the drug. Second, one must intervene on each individual's blood pressure and set it to the value it would have had were they to not take the drug. In this example, one would set each individual's blood pressure to medium. The quantity calculated in this manner is called the *natural direct effect*.

The concept of the natural direct effect provides the solution to a problem that might otherwise appear insoluble. The scientists in this case desire to know about the contribution of the direct path in the case where the indirect path is not disrupted. Yet the only way to isolate the direct effect is to disrupt the indirect path by intervening on the mediator. Pearl's (2001) solution is to break the indirect path by intervening on the mediator, but to use an intervention that

mimics the behavior of the mediator in the case where the path is not broken and the individuals do not receive the treatment. By setting the mediator as a function of the treatment, one makes the mediator behave as if it were still causally dependent on the treatment. By setting the mediator to the value that it would have had in the case where one did *not* receive the treatment, one renders the indirect path inactive.

Why do all the work necessary to measure the natural direct effect? Only the natural direct effect identifies the contribution of the direct path in the system that one is studying. If one were to intervene on the mediator to set it some arbitrary value, one would destroy the indirect path, but there would be no interesting relationship between the effect of the treatment on the outcome in the resulting system and the role played by the direct path in the original undisrupted system. The natural direct effect provides the answer to the following question: Given that the treatment causes the outcome to take a certain value when none of the paths are disrupted, what is the probability that the treatment would still be sufficient for the outcome were the indirect path to be disrupted? It is only through joint interventions of the sort required for measuring the natural direct effect that one can make the relationship between the contributions of paths in a disrupted system and the effect of the treatment on the outcome in an undisrupted system precise.

The definition for the indirect effect is more complicated than that for the natural direct effect. It is straightforward to see that in order to measure the direct effect one has to intervene on the mediator. For the indirect effect, there is no measured variable on the direct path upon which one can intervene. This obstacle may be surmounted as follows. To measure the indirect effect, one must perform two joint interventions on the treatment and the mediator:

1. a) Assign each participant to not take the drug
   b) Set each participant's blood pressure to the value it would have had *had* she taken the drug

2. a) Assign each participant to not take the drug
   b) Set each participant's blood pressure to the value it would have had had she *not* taken the drug

The indirect effect is given by the difference in cholesterol levels resulting from joint interventions 1 and 2. By intervening on the indirect path and changing the mediator in the way that it would have changed had one intervened on the treatment, one simulates the effect of the treatment on the mediator. In this respect, the path behaves as if it were not broken. Nevertheless, by assigning everyone to the control group, one eliminates the contribution of the direct path. Using this method, one can identify the sufficiency of the indirect path without any disruption to the direct one.

The definition of the indirect effect is perhaps the most significant success of causal mediation techniques. Moreover, it is difficult to see how one could develop an adequate account of mechanistic explanation without appealing to indirect effects. Craver's discussion of inter-level experiments clarifies how one can establish that the rat's entering the maze activates the hippocampus, which in turn leads to the rat's successfully completing it. The existence of this causal path containing the hippocampus leaves open many important questions about the role of the hippocampus in spatial memory. For example, are there distinct causal paths that influence spatial memory? If so, to what degree does the rat's maze-navigating ability depend on the hippocampus, rather than on these other paths? Only by measuring the indirect effect can one determine the degree to which the hippocampus contributes to spatial memory independent of factors on other causal paths, if there are any. I can afford to be non-committal about whether there are other paths, since my point here generalizes beyond this particular case. As we will see in the following chapter, in identifying the indirect effect, one can determine the contribution of

the indirect path without having knowledge of unmeasured variables on paths not going through the mediator, or even knowing whether there are such variables or paths.

At long last, we can see why the insufficiency of bottom-up experiments for uncovering the behavior of a mechanism in no way reveals that mechanisms call for non-causal explanation. To start, we can now precisely identify why singular ideal interventions on components are inadequate for evaluating their role in a functioning mechanism. The ability of a mechanism to bring about a phenomenon may depend on causal paths not going through the component. Moreover, the effect of the component on the outcome may be sensitive to the activity of these causal paths. Consequently, one cannot evaluate the contribution of a component by intervening on it unless one *also* intervenes to make the other paths behave in whatever ways are necessary to answer the relevant counterfactual question.

Our discussion of causal mediation techniques reveals that questions about the "contribution" of a component are ambiguous between questions about its necessity and its sufficiency. Answering these questions requires joint interventions that set the treatment and mediator to specific values. Craver would refer to interventions on the treatment as top-down experiments and to interventions on the mediator as bottom-up experiments. But the necessity of using joint interventions to determine the contribution of a component is a result of interactions among variables and in no way points towards the existence of a non-causal form of explanation.

While it was important to specify how one can use ideal interventions to identify direct and indirect effects, the crucial point for our purposes is to understand what these effects are. The direct and indirect effects indicate the contribution of one path in the case where the other path behaves as it "naturally" would in the absence of an intervention on the treatment. Finding these contributions presents the challenge that in order to isolate a path one must disrupt the other path,

but in complex systems the contributions of the paths are not independent. This challenge may seem insurmountable, and I suggested that this appearance motivates the belief that causal explanation is limited to the relations among components in isolation and not applicable to their contributions to the whole. Causal mediation techniques surmount this apparent barrier. If mechanisms elude causal explanation, it is *not* because causal inference is limited to the synthesis of effects measured in isolation. Just as mechanists insist that we should decompose a mechanism into its component parts, causal modelers can extol the virtues of decomposing a total effect into direct and indirect effects.

The relationship between the total effect and the direct and indirect effects is not causal. Does this show that mediation techniques rely on a non-causal form of explanation? Of course not. The fact that the direct effect is not a cause of the total effect is irrelevant to the question of whether the contribution of the direct path to the outcome is causal. Clearly, it is. Similarly, we can grant to the mechanists that there is a non-causal relationship between the behaviors of components and the behavior of the mechanism without granting that there is a non-causal form of explanation.

6. Mediation and Explanation

Throughout the mechanist literature, one finds the idea that the phenomena produced by mechanistic systems are too complex to be explained solely by the "lower-level" behaviors of their parts (e.g. Bechtel and Richardson, 1993; Machamer *et al*. 2000). By examining Craver's account, I have made this idea precise. The "lower-level" refers to the causal relations among components. The feature of complex systems that these allegedly cannot explain is the undisrupted behavior of the mechanism. I have answered this challenge to the sufficiency of causal explanation by appealing to causal mediation techniques. In this section I briefly consider

the usefulness of these techniques for developing an account of explanation that addresses the questions with which the mechanists have been concerned.

It is easy to see why measuring mediators contributes to one's understanding of a phenomenon. According to Woodward (2003, 191), successful explanations are able to answer "w-questions" of the form: What if things had been different? Mediation techniques allow one to answer a wider range of w-questions than one would be able to answer given only knowledge of the total effect. Woodward is not idiosyncratic in judging explanations by their ability to answer counterfactual questions. Glennan echoes a common mechanist claim when he writes:

> Understanding the nature, structure, and functional organization of the parts that make up that mechanism will allow one to determine the range of counterfactual circumstances under which the dependency between X and Y would be maintained—roughly those circumstances in which the mechanism will not break down. (2012, 288)

Mediation techniques enable one to do more than determine when a dependency between *X* and *Y* will break down. They allow one to quantify the ways that the magnitude of the dependency is sensitive to the activity of particular paths.[27]

Mediators often correspond to the entities that scientists call components. In fact, on the interpretation of Craver's mutual manipulability criterion as saying that if φ is causally between the input and output of the mechanism, then it is a component, *all* mediators are mechanism components.[28] If so, then mediation techniques have a role to play in mechanistic explanation. This is *not* to say that mediators explain *in virtue of* being (properties of) components. The burden of proof remains on the mechanists to show that the mechanism-component relation has any explanatory significance.

Mechanists might disparage explanations that reduce physical mechanisms to DAGs as being anemic. Scrapings and pushings and dryings and carryings are uniformly replaced with

---

[27] Gebharter (2014) offers an account of multi-level mechanisms that allows for quantitative predictions, but he does show how to quantify the contributions of paths.

[28] Menzies (forthcoming) similarly argues that components are mediators.

arrows and variables. To develop this into an objection, mechanists would have to say more about what it is that DAGs allegedly omit. DAGs are able to represent the counterfactual dependence of variables on other variables and of a total effect on direct and indirect effects. If there are physical features of a system on which the mechanism does not counterfactually depend, in what sense are they explanatory? And what, precisely, do they explain?

One apparent limitation of DAGs is that they do not represent the spatiotemporal organization of a mechanism. Mechanists have yet to produce a compelling account of "spatiotemporal explanation". Craver's account is opaque regarding what explanatory role spatiotemporal organization is supposed to play independent of the mutual manipulability criterion. Fagan invokes spatiotemporal organization to explain the interdependent behaviors of components. But this interdependence can be adequately explained using mediation techniques.

As textbooks are replete with diagrams detailing the components of mechanisms, it may seem obvious that the tracing of a mechanism's components plays an integral part in explaining phenomena in the higher-level sciences. Causal mediation techniques suggest a way of understanding the role of mechanisms in science without needing to address the question of what counts as a component. I have not proven that mediation techniques are adequate for explaining all mechanistic phenomena, but I have, I hope, refuted the main argument to the contrary.

7. Conclusion

Scientists often pejoratively refer to "black-box explanations". These explanations describe a phenomenon at too coarse-grained a level and therefore fail to explain it. What makes such explanations so bad? For mechanists, the problem is that they omit the physical entities and activities that are responsible for the phenomenon. To open the black box, one must examine the

relationship between the mechanism producing the phenomenon and its parts. Since this relationship is not causal, a new form of explanation is required.

Causal mediation techniques suggest a different answer for why black-box explanations are deficient. Black-box explanations are uninformative regarding the counterfactual conditions under which a functional relationship will break down or be modified. Mediation techniques shed light on why one can better explain a phenomenon by uncovering mechanism components. One can therefore grant that scientists studying mechanisms are engaging in a task of great explanatory importance without granting that whether something counts as a component matters for explanation.

The question of specifying the conditions under which a mechanism breaks down and when it doesn't is a question about extrapolation. If one knows that a mechanism breaks down under a particular set of circumstances, one knows that the causal relationship between that mechanism's input and output does not generalize to those circumstances. In the following chapter, I provide a more thorough introduction to causal mediation techniques and explain how and when they license inferences about the robustness of a causal relationship across interventions on particular causal paths. I then further explore the relationship between mediation, extrapolation and transportability in chapter 7.

Chapter 6: Causal Mediation Techniques

In my discussion of mechanistic explanations in chapter 5, I informally introduced causal mediation techniques. In chapter 7, I will evaluate the usefulness of mediation techniques for extrapolation. To do so, I need to provide a more extensive introduction to the details of mediation techniques. That is the aim of this chapter.

I present four ways of expressing the definitions of the direct and indirect effects in mediation models. First, one can give general definitions using Rubin's (1974) potential outcomes. Second, in contexts where there are no confounding variables, one can define direct and indirect effects using conditional probabilities. Third, in contexts with possible confounding, one can express the effects using the conditional probabilities and do-operators. As we will see, this way of expressing direct and indirect effects is problematic. Fourth, in models with their structural equations specified, it is possible to express direct and indirect effects in terms of the structural parameters in the equations.

After presenting an overview of mediation techniques, I elucidate one way that they facilitate extrapolation. A clear sense in which mediation techniques relate to extrapolation is that the direct effect is what the total effect would be in a population where the indirect path is entirely disrupted, and the indirect effect is what the total effect would be in a population where the direct effect is entirely disrupted. I explore the possibility of using mediation techniques to extrapolate not only to cases where one of the paths is fully disrupted, but also to cases where the paths are only partially disrupted. I show that given parametric assumptions, it is possible to find the highest reduction potential of a policy that seeks to disrupt the indirect path. By "highest reduction potential" I mean the maximum reduction in the total effect that could result from any version of the policy that is at least partially successful in disrupting the indirect path. The idea

that mediation techniques enable one to find the highest reduction potential of policies that seek to disrupt particular paths is an elaboration on a point made by Pearl (2012a). I clarify the parametric conditions under which mediation techniques identify the highest reduction potential. I also argue that one can only specify such conditions when considering policies that disrupt the indirect path. There is no way to use parametric models to non-trivially specify the possible degrees of success of a policy that seeks to disrupt the direct path.

The formatting of this chapter differs from others in that each section is divided into subsections. Section 1 introduces the potential outcomes notation (1.1), defines natural and controlled direct effects (1.2), defines the indirect effect (1.3), and then compares these definitions to definitions relying on probabilistic expressions and do-operators (1.4). Section 2 differentiates between different versions of direct and indirect effects (2.1) and shows how to decompose the total effect into direct and indirect effects (2.2). Section 3 reviews the distinction between parametric and non-parametric causal inference (3.1), gives parametric definitions of the direct, indirect effect and total effects in a model assuming additivity (3.2), and then provides the corresponding definitions for a model allowing for interaction (3.3). Section 4 presents the parametric conditions under which mediation techniques enable one to identify the highest reduction potential for a policy. Section 5 concludes.

1. Defining Direct and Indirect Effects

1.1. Potential Outcomes

Within Pearl's causal modeling framework, the standard way to represent causal effects is in terms of interventions. The effect of X on Y is given by the probabilistic expression $P(Y|do(X))$, where $do(X=x)$ indicates that X is set to x by an arrow-breaking intervention. For reasons that will become clear, expressions consisting of probabilistic expressions and do-operators are ill

suited for representing direct and indirect effects. The standard definitions for path specific-effect are given in terms of potential outcomes, which I will now introduce.

The potential outcomes notation was developed by Donald Rubin (1974). Potential outcomes are counterfactuals concerning how an individual would respond to a treatment. The potential outcome of receiving treatment $X=x$ on outcome $Y=$y for individual i is denoted as follows:

$$(1)\ Y_x^i = y$$

(1) is a deterministic counterfactual saying that if i receives treatment level x, her outcome with respect to Y will be y. For example, if $X=x$ is taking Excedrin and $Y=y$ is not having a headache, (1) says that were i to take Excedrin she would not have a headache. One can also consider the probability that taking Excedrin will cure a headache for a randomly selected individual in non-homogeneous population:

$$(2)\ P(Y_x = y)$$

In most cases we will be considering potential outcomes in non-homogeneous populations and I will therefore usually omit superscripts referring to individuals. Proponents of the potential outcomes framework emphasize the deterministic nature of the counterfactuals, but in practice one can typically identify only the average outcome for a treatment.

The potential outcome in (2) is equivalent to Pearl's P(Y=y|do(x)). The difference between Pearl's approach and Rubin's is largely notational. In Pearl's framework, one represents one's causal knowledge using a DAG and then derives facts about potential outcomes from a DAG. One derives the potential outcome of x on y by breaking all of the arrows into X and setting its value to x. In Rubin's framework, one does not use graphs. Rubin takes $Y_x^i = y$ to be an undefined primitive specifying how an individual would respond to treatment level $X=x$. He

then provides additional assumptions that, when met, allow one to derive causal effects from potential outcomes. For example, the total effect on *Y* of changing the value of treatment *X* from *x* to *x'* is defined as follows.

$$(3)\ TE_{x,x'}(Y) = Y_{x'} - Y_x$$

In (3), Y_x indicates the value of *y* given *x*, so the total effect is the difference in the value of the outcome given two distinct values of the treatment. The assumptions under which one can infer the average $TE_{x,x'}$ for a population are roughly those that are met by an experimental population in which individuals have been randomly assigned *x* or *x'*. Rubin's particular assumptions do not matter for our purposes here. Pearl proves that his and Rubin's frameworks are interchangeable.

We have already seen that identifying direct and indirect effects requires one to intervene on both the treatment and the mediator. Here is the potential outcome expression for the result of the joint intervention on the variables *X* and *M*:

$$(4)\ Y_{x,m} = y$$

Additionally, it is also possible to nest potential outcomes. That is, the treatment variable in a potential outcome expression can itself be given as a potential outcome, as follows:

$$(5)\ Y_{Mx} = y$$

(5) states that when one sets the value of *M* to the value that it would have been had *X* been *x*, the outcome is *Y*=*y*. The expression Y_{Mx} does not specify a value *M*=m to which *M* is set, but rather allows the value of *M* to depend on the value of *X*.

While one can nest potential outcome expressions, one cannot nest do-expressions. The expression P(Y|do(M=m|do(X=x)) is not a well formed formula. To see this, it helps to compare this expression to two expressions that are well formed. P(Y|do(M=m)) denotes the probability of *Y* if one sets the value of *M* to *m* so that *M* no longer depends on its prior causes.

P(M=m|do(X=x)) denotes the probability that *M* would naturally take on the value of *m* when one does not intervene upon it, but one does intervene upon *X.* The expression P(Y|do(M=m|do(X=x)) is nonsensical, since it denotes the impossible case in which one both sets *M* to a particular value and to lets its value be determined by *X*. The reader may recall from chapter 5 that identifying the direct and indirect effects requires one to set the mediator in such a way that it behaves *as if* it were still responding to the treatment. In §1.4 we will address the question of whether such interventions can be represented with do-operators. Since one *can* nest potential outcomes, it is straightforward to use them to denote interventions that set the mediator to a value as a function of the treatment, as we will see presently.

1.2. Natural and Controlled Direct Effects

We can now use the potential outcomes notation in order to give precise definitions of the direct and indirect effects. In giving the definitions, I will continue to use the cholesterol drug example from the previous chapter. In the example, the treatment, *X*, is a drug that reduces cholesterol, the mediator, *M*, is blood pressure, and the outcome, *Y,* is cholesterol level. *X*=0 is taking a placebo and *X*=1 is taking the drug. We don't need to specify the possible values of the other variables (they could be dichotomous or not, discrete or continuous etc.).

The direct effect of the treatment on the outcome is the effect that is not due to the path going through the mediator. When one intervenes on the mediator to set it to the same value for every member of a population, one can identify the *controlled direct effect* for that population:

$$(6) CDE(m)_{0,1}(Y) = Y_{1,m} - Y_{0,m}$$

The controlled direct effect is the difference in outcome between the treatment and control cases when one holds the mediator fixed at *M*=*m*. The subscripts 0 and 1 on the left-hand-side denote the control and treatment values of *X.* The order of the subscripts matters: had I written

$CDE(m)_{1,0}(Y)$, this would correspond to $Y_{0,m} - Y_{1,m}$which equals *negative* $CDE(m)_{0,1}(Y)$. An example of a controlled direct effect would be the effect of the drug on cholesterol when one sets everyone's blood pressure to low. Since the drug can interact with blood pressure in changing one's cholesterol level, there are as many controlled direct effects as there are values of *M* and they could all have different values.

Hitchcock (2001b) distinguishes between total effects and component effects. He defines the component effect of *X* on *Y* as the effect of *X* on *Y* while holding all other variables (including the mediator) fixed. He does not specify to which value one should set the mediator, so his definition for direct effect coincides with that of the controlled direct effect. As I just noted, however, there are many controlled direct effects, so without specifying a particular value of the mediator one cannot say anything definite about the direct effect – not even whether it is positive or negative. Worse, if one sets the mediator to the value that it naturally would take on given the treatment (i.e M_1), one will not identity a component effect at all, since the controlled direct effect for that value of *M* will equal the total effect.

The *natural direct effect* is the effect that the treatment would have on the outcome were the mediator to take on the value that it naturally would were one not to receive the treatment:

$$(7) NDE_{0,1}(Y) = Y_{1,M(0)} - Y_{0,M(0)}$$

There are two related senses in which the natural direct effect is natural. First, one sets the mediator as a function of the treatment, since it takes on the value it would have had in the control scenario (*X*=0). Second, one sets the mediator to the value that it would have had in the control scenario *for every individual.* If the individuals of the population differ in the value of M_0, the NDE for the population is a weighted average of the NDE across all individuals. In this

sense, although NDE is calculated by an intervention on the mediator, it is not a population-level intervention, since it does not set every individual's mediator to the same value.

One cares about the *natural* direct effect because one wants to know about what the direct path is contributing in the population that one is studying. Suppose that the value of M_0 is less than that of M_1 and one set the mediator to a value that is less than that of M_0. In the drug case, this would involve setting blood pressure to low when no one in the population has low blood pressure in either the treatment or control scenarios. Learning that taking the pill lowers the average cholesterol in the population when everyone has low blood pressure does not tell one that there is a direct effect in the population when the pill is taken under ordinary circumstances. What one wants to know is whether the pill would have lowered cholesterol in the members of the population if they took the pill and it did not raise their blood pressure. To evaluate this, one cannot set their blood pressure to just any level; one must set it to the level that it would have had were they not to take the drug, but rather the placebo.

Note that $NDE_{0,1}$ does not equal negative $NDE_{1,0}$. I'll have more to say about this later. Here I'll just highlight that in evaluating $NDE_{0,1}$ one holds the mediator fixed to M_0 and in evaluating $-NDE_{1,0}$ one holds the mediator fixed to M_1. In the following, I will only be referring to the natural direct effect, so from here on I will usually refer to NDE as the direct effect. I will refer to $NDE_{0,1}$ as the *sufficient direct effect* and $-NDE_{1,0}$ as the *necessary direct effect*, for reasons I will make clear in §2.1. Whenever I refer to the 'direct effect' without further clarification, I mean the sufficient natural direct effect.

1.3 Indirect Effects

In a model with a direct and indirect path, there is no included variable that one can intervene upon in order to disable the direct path and evaluate the indirect effect. For this reason, in the

first edition of *Causality* (2000) Pearl argued that it is not possible to provide a meaningful interpretation of the indirect effect. Apparently undaunted by the impossibility of the task, Pearl (2001) provided a definition of the indirect effect.

To define the indirect effect one must find a way to make the mediator behave as if it were responding to a change in the treatment variable without actually changing the value of the treatment value. The first step to doing this is already present in the definition of the natural direct effect. In finding the natural direct effect, one does not intervene on the mediator in such as way as to fully sever its connection with the treatment. Rather, by setting it as a function of the control value of the treatment one makes it behave as if it were still responding to the treatment. The second insight necessary for defining the indirect effect is that one can make the mediator behave as if it were responding to a *change* in the value of the treatment. By comparing the mediator at both M_0 and M_1, we can see how the mediator would respond to a change in the treatment variable without actually changing the value of the treatment variable.

Within the potential outcomes framework, the indirect effect is given by the following equation:

$$(8)\ IE_{0,1}(y) = Y_{0,M(1)} - Y_{0,M(0)}$$

By comparing the values of the outcome for both M_1 and M_0, one simulates the behavior of the mediator in response to the treatment. Since in both terms the value of the treatment is 0, any difference between the terms is not a result of a change in the value of the treatment.

As with the direct effect, $IE_{0,1}$ does not equal negative $IE_{1,0}$. While $IE_{0,1}$ holds *X* at 0, $IE_{1,0}$ holds *X* at 1. I will refer to $IE_{0,1}$ as the *sufficient indirect effect* and to $-IE_{1,0}$ as the *necessary indirect effect* and I will be referring to the former whenever I don't explicitly say otherwise. Note the minus sign in front of the necessary indirect effect. The minus sign is there to correct

for the fact that when one switches from $IE_{0,1}$ to $IE_{1,0}$ one switches the order in which one subtracts one potential outcome from another. That is, as a result of the negative sign, the necessary indirect effect corresponds to $Y_{1,M(1)} - Y_{1,M(0)}$ rather than $Y_{1,M(0)} - Y_{1,M(1)}$. This allows for a more straightforward comparison between the necessary and sufficient indirect effects, since both are derived by subtracting the term with M_0 from the M_1. The important difference between the sufficient and necessary direct effects is that they are evaluated relative to different values of *X*.

Tellingly, there is no controlled version of the indirect effect. Since there is no variable that one can fix for all members of the population in order to derive IE, one can only evaluate it by making the mediator vary as it naturally would in response to the treatment.

Direct effects can only be defined relative to a model. The direct effect of *X* on *Y* is the effect not going through any mediator that is included in the model. Indirect effects are also model relative in the trivial sense that to specify an indirect effect, one must specify a mediator. Yet they differ from direct effects in that one can define them without reference to any other paths.[29] Of course, if one knows about the mediators along other paths, one can identify the indirect effect of interest by intervening to set those mediators to their "natural" values. But one does not have to do so. Holding the treatment fixed at $X=0$ ensures that any of these unmeasured variables will take on their "natural" values automatically.

I will postpone discussing the relevance of mediation to extrapolation until the next chapter, but the fact that one can define the indirect effect independently of the behavior of the other paths is clearly important. A central difficulty for addressing extrapolation is that the effect of one variable on another can depend on an indefinite number of background factors and even if

[29] In the case where there is only one path and that path has a mediator, the "indirect" effect equals the total effect.

one tries to isolate some of those factors, there could be others that vary across populations in ways that alter the magnitude of the effect. This is what I called "the problem of the unknown unknowns". When one identifies the indirect effect, one is able to learn something about the behavior of an indefinite number of unmeasured mediators along other paths. Namely, one is able to determine what would occur if none of them responded to the change in the treatment.

1.4 Potential Outcomes, Conditional Probabilities and do-Expressions

I will now explain why the definitions of NDE and IE are more easily expressed using potential outcomes notation than they are using conditional probabilities with do-operators, despite the fact that the two notations are inter-translatable. To do so, I will show how both NDE and IE can be represented using conditional probabilities in a population with no confounding and then reveal the difficulties with adding do-operators to these expressions in cases with confounding.

In populations where there is no confounding and the error terms are therefore independent (as follows from the Causal Markov Condition), we can replace expressions of the form P(*Y*|do(*X*)) with the simpler P(*Y*|*X*). The expression for NDE is as follows:

$$(9) NDE_{0,1} = \sum_{m} [E(Y|X = 1, M = m) - E(Y|X = 0, M = m)]P(M = m|X = 0)$$

In reading this expression, it helps to momentarily ignore the term in brackets in order to highlight the way that the function $\sum_m[\quad]P(M = m|X = 0)$ enables one to provide an average effect over different values of the mediator. Not every member of the population has the same value of the mediator when *X* equals 0. $P(M = m|X = 0)$ gives the distribution of the values of the mediator given that *X*=0. The assumption of no confounding guarantees that this distribution corresponds to the distribution of the mediator if one were to *intervene* to set the treatment for each individual to 0. We see that the function $\sum_m[\quad]P(M = m|X = 0)$ takes the term in

brackets and reweights it according to the distribution of the mediator in the case where *X*=0. Now let's look at the term in brackets. This term denotes the difference in the expected value of the outcome given the treatment value and its expected value given the control value for a particular value of the mediator. This is the expression that captures the intervention on the treatment that changes it from *X*=0 to X=1. This effect is different for different individuals in the population based on their value for the mediator. To calculate the average direct effect in the population, one finds the effect of changing the treatment from 0 to 1 for each value of the mediator, and then uses the weighting function to determine what the distribution of the mediator would be in a population where everyone received the control version of the treatment. Averaging in this manner yields the natural direct effect ($Y_{1,M(0)} - Y_{0,m(0)}$).

One cannot in general assume that there is no confounding. One might try and generalize the definition of NDE to scenarios where there is possible confounding by adding do-operators as follows:

$$(10) NDE_{0,1} = \sum_{m} \frac{\left[E(Y|do(X=1), do(M=m)) - E\left(Y\middle|do(X=0), do(M=m)\right)\right]}{P(M=m|do(X=0))}$$

The use of the do-operator in (10) is highly non-standard. The do-operator usually corresponds to an ideal experimental intervention in which one sets a variable to a particular value for every member of a population. Here one sets the value of the mediator to different values for different members of the populations in such a way that the percentage of individuals who receive a certain value of *M* is proportional to P(M=m|do(X=0)). Moreover, for (10) to provide an estimate of the average NDE in the population, one needs to assign values of the mediator in such a way that whether an individual *i* is assigned a particular value of the mediator is uncorrelated with how she would react to the mediator assignment (i.e her distribution for $P(Y^i_{x,m} = y)$). So although NDE can be expressed in do-notation, it is not clear whether it corresponds to a well-

defined intervention. To identify the NDE, it might be more promising to attempt to deconfound the *X*-*M* relationship (by conditioning on common causes), rather than physically intervening on *M*.

In a population with no confounding, the probabilistic expression for IE is as follows:

$$(11) IE_{0,1}(Y) = \sum_{m} E(Y|X = 0, M = m)\,[P(M = m|X = 1) - P(M = m|X = 0)]$$

The first term is expected value of the outcome for the control value of the treatment and different values of the mediator. The weighting of the mediator is given by the second term, which, for each value of the mediator, is the difference in its value given the treatment and control values of the treatment variable. Note that the variable *X* plays two distinct roles in the definition. In evaluating the expected value of the outcome, one holds the treatment fixed at the control value. In calculating the weighting for the value of the mediator, one varies the mediator to mimic the way it would vary were the treatment to change. This corresponds to the two roles of the treatment in the corresponding potential outcome expression $IE_{0,1}(Y) = Y_{0,M(1)} - Y_{0,M(0)}$.

The natural way to generalize (11) to cases with confounding is as follows. First, distribute the first term in the expression over the two terms in the brackets:

$$(12) IE_{0,1}(Y) =$$

$$\sum_{m} E(Y|X = 0, M = m)P(M = m|X = 1) - E(Y|X = 0, M = m)P(M = m|X = 0)$$

Then add do-operators as follows:

$$(13) IE_{0,1}(Y) = \sum_{m} \begin{matrix} E(Y|do(X = 0), do(M = m))P(M = m|do(X = 1)) - \\ E\big(Y\big|do(X = 0), do(M = m)\big)P(M = m|do(X = 0)) \end{matrix}$$

Yielding what I believe to be the most ugly equation in the dissertation.

The problems with (13) are not merely aesthetic. There does not appear to be any way to identify the indirect effect using an experiment. This is clearest if we consider some individual *i*. To know the indirect effect for *i,* we need to know the following four things:

A) *i*'s value of the mediator for X=0 (M_0)

B) *i*'s value of the mediator for X=1 (M_1)

C) *i*'s value for the outcome in the case where she has the value of the mediator she would have if X=0 ($Y_{0,M(0)}$)

D) *i*'s value for the outcome in the case where she has the value of the mediator she would have if X=1, but she did not receive the treatment ($Y_{0,M(1)}$)

It is a familiar problem in causal inference that to measure the magnitude of an effect, one must know how a subject would respond to both the control and the treatment; but every individual receives only X=0 or X=1. The standard solution is to randomize, so that the behavior of individuals in the control condition corresponds to that of how those who took the treatment *would* have responded had they been in the control condition. This solution does not work in the case of mediation. To find the values of both C and D, one would have to randomize the mediator. But randomizing the mediator would break the arrow between the treatment and the mediator and thus make it impossible to identify both A and B by randomizing the treatment.

The upshot of the present discussion is that although it is possible to express direct and indirect effects using do-operators, doing so does not grant one the characteristic benefits of using do-expressions. While such expressions generally provide a guide to how to identify a causal relationship using controlled experiments, the do-expressions for NDE and IE do not.

Although the details of this section have been fairly technical, the explanation for why do-expressions are ill suited for discussing direct and indirect effects is simple. To find the direct or indirect effects, one must intervene on the mediator. Interventions on the mediator break its connection with the treatment. Yet, one must set the mediator so that it behaves as if it were

responding to the treatment. In the absence of confounding, the response of the mediator to the treatment corresponds to P(M|X). While Pearl's notation can easily evaluate P(M|X) and also expressions containing do(M), the evaluation of one expression in a population is incompatible with the evaluation of the other.

2. The Interpretation of Direct and Indirect Effects

2.1 Necessary and Sufficient Effects

The total effect $TE_{x,x'}(Y)$ always equals $-TE_{x',x}(Y)$. If moving from *x* to *x'* increases *Y* by a certain amount, then moving from *x'* to *x* decreases *Y* by that same amount. With the total effect, the order of the subscripts affects whether TE is positive, but nothing else. In contrast, for NDE and IE, changing the order or the subscripts changes the quantity that one is evaluating. When evaluating $TE_{x,x'}$, $NDE_{x,x'}$ is the sufficient direct effect, $-NDE_{x',x}$ is the necessary direct effect, $IE_{x,x'}$ is the sufficient indirect effect, and $-IE_{x',x}$ is the necessary indirect effect. These ways of labeling the effects are mine, though the interpretation I am about to provide for these effects is standard.

In thinking about the interpretations of path-specific effects, it helps to bear the following identity in mind:

$$(14) Y_{x,M(x)} = Y_x$$

In other words, the result of setting the treatment to *X*=*x* and the mediator to the value it takes when *X*=*x* is the same as just setting the treatment to *X*=*x* and not intervening on the mediator. Given this identity, we can rewrite the sufficient and necessary direct effects as follows:

$$(15) NDE_{0,1}(Y) = Y_{1,M(0)} - Y_0$$

$$(16) -NDE_{1,0}(Y) = Y_1 - Y_{0,M(1)}$$

These expressions make salient the counterfactuals that each quantity represents. The second term in (15) is the value the outcome variable would have in the control condition. The first term in (15) indicates the value the outcome variable would have were one to take the treatment while rendering the mediator unable to respond to this change in the treatment. The difference between these terms is the increase in the value (or expected value) of the outcome that would result from taking the treatment rather than the control were there to be no indirect path. It is the benefit (or harm) of the direct path as compared to the scenario where there is no treatment.

While the second term in (15) refers to the outcome when one does not receive the treatment, the *first* term in (16) refers to the outcome when one does. The second term then subtracts the value of the outcome when there is no treatment, but the mediator acts as if there were one. It is the harm (or benefit) of lacking the direct path as opposed to having the treatment act through both paths.

In a deterministic system where *Y* has only two values, 1 and 0, corresponding to whether the outcome does or does not occur, the sufficient direct effect will be 1 just in case the outcome would not occur in the absence of the treatment and would occur in the presence of the treatment even if there were no indirect path. We would then say that the direct path is sufficient for the outcome. The necessary direct effect is 1 just in case the outcome would occur given the treatment, but would not occur if the direct were path not active. We would then say that the direct path is necessary for the outcome.

In mixed populations, the sufficient and necessary direct effects will indicate changes in the expected value of the outcome and may have values between 0 and 1. The sufficient direct effect is then the increase over $E(Y_0)$ in the expected value of the outcome that would occur if

there were a direct effect. The necessary direct effect is the amount by which the expected value of the outcome in the case where there were no direct effect would be less than $E(Y_1)$.

When the outcome variable is non-dichotomous, one cannot claim that an effect is necessary or sufficient for the effect without further specification. In order to say that the pill was either necessary or sufficient for reducing cholesterol, one needs to specify a level of cholesterol such that the pill is considered to be successful when it brings cholesterol below that level. It only makes sense to talk about the treatment being necessary or sufficient for an outcome once one has specified the level of the outcome variable for which it is necessary or sufficient.

We can explicate the sufficient and necessary indirect effects in a similar manner. Using the equivalence in (16) we get the following definitions:

$$(17)\ IE_{0,1} = Y_{0,M(1)} - Y_0$$

$$(18)\ -IE_{1,0} = Y_1 - Y_{1,M(0)}$$

The sufficient direct effect is the effect of the treatment acting only through indirect path as compared to the scenario in which it acts through neither of them. The necessary direct effect is the effect of the treatment not acting through the indirect path, as compared to its acting through both. These are different, because they are evaluated relative to different values of the treatment.

2.2 Decomposing the Total Effect Into DE and IE

The total effect decomposes into direct and indirect effects in one of two ways:

$$(19) TE_{0,1} = NDE_{0,1} - IE_{1,0}$$

$$(20)\ TE_{0,1} = IE_{0,1} - NDE_{1,0}$$

Note that in each equation, the subscripts for the direct effect and the indirect effect are reversed. The proofs of (19) and (20) trivially follow from definitions 15-18. For example, we can explicate the right hand side of (19) with using (15) and (18):

$$(21) NDE_{0,1} - IE_{1,0} = Y_{1,M(0)} - Y_0 + Y_1 - Y_{1,M(0)} =$$

$$Y_1 - Y_0 = TE_{0,1}$$

It may appear strange that to get the total effect one must *subtract* one path-specific effect from another, but this is just an artifact of the negative sign in the necessary direct and indirect effect. Using my definitions, the total effect decomposes into either the sum of sufficient direct effect and the necessary indirect effect, or the sum of the sufficient indirect effect and the necessary direct effect.[30]

It is important that the total effect is *not* the sum of the sufficient direct effect and the sufficient indirect effect. The reason for this is that when the contributions of the paths are non-additive, the total effect is not the effect of two independent contributions. To dramatize the point, if there are two paths that are individually sufficient, then both the sufficient direct effect and the necessary direct effect would be one and the total effect would also be one. One can think of decompositions (19) and (20) as follows. Suppose that the total effect is that *X* raises the expected value of *Y* by .8 and the direct effect is .5. Since the direct path cannot account for the total effect in the absence of the indirect path, the indirect path is needed to get the total effect. The necessary direct effect picks up slack, so to speak, for the failure of the sufficient direct effect to bring about the total effect by itself.

The sum of the sufficient direct and indirect effects *does* equal the total effect when the contributions of the paths are additive. When the paths are additive, the direct effect of the

[30] I am grateful to Malcolm Forster for pointing out that there was a way to represent the total effect as a sum.

treatment on the outcome does not depend on the value of the treatment, so $NDE_{0,1} = -NDE_{1,0}$. Plugging this equivalence into (19) yields $TE_{0,1} = IE_{0,1} + DE_{0,1}$. The additivity of the direct and indirect effects in models without interaction is a special case and should not be assumed in general.

In section 2.1, I explicated direct and indirect effects in terms of whether they were necessary or sufficient for the outcome to take on a certain value. Another way to explicate direct and indirect effects is in terms of how the magnitude of path-specific effects compares to the total effect. Pearl (2012) distinguishes between the portion of an effect that is explained by a path and the portion of an effect that is owed to a path. By dividing the direct effect by the total effect, one gets the portion of the total effect that is *explained* by the direct effect. That is, it is the portion of the total effect for which the direct path is sufficient. The rest of the total effect is *owed* to the indirect path. 1 – DE/TE gives the portion that is owed to the indirect effect. 1 – DE/TE is equivalent to the necessary indirect effect divided by the total effect, as can be seen by dividing all the terms in (19) by $TE_{0,1}$.

Correspondingly, IE/TE gives the portion of the effect for which the indirect path is sufficient and 1 – IE/TE is the portion for which the direct path is necessary. 1 – IE/TE is equal to the necessary direct effect divided by the total effect.

3. Parametric Versions of DE and IE

3.1 Parametric and non-Parametric Inference

The definitions that I have provided so far for path-specific effects are entirely non-parametric. Most crucially, they allow for any degree of interaction between the treatment and the mediator. In this section, I consider the properties of direct and indirect effects in models with particular parametric forms. Although these are only special cases of the more general definitions just

provided, they will be helpful for getting a more concrete picture of what the different effects represent. Moreover, in section 4 I argue that one of Pearl's suggestions for how mediation techniques facilitate extrapolation only works given certain parametric assumptions.

The two types of models I will consider are linear models and models that contain an interaction term, but are otherwise linear. Although causal models may have an indefinite number of functional forms, the difference between models with and without interaction is more important for understanding mediation than whether, for example, a relationship is linear or quadratic.

3.2 TE, DE and IE in Additive Models

A nice example of a mediation model in which the direct and indirect paths make additive contributions comes from consumer choice theory.[31] Consider an agent Renée who spends all her money on coffee and cigarettes. For a given level of income, Renée buys a bundle of coffee and cigarettes that maximizes her utility. Renée is willing to substitute coffee for cigarettes, and the more coffee she has, the fewer cigarettes are needed in order to compensate for the loss of a cup of coffee (and vice versa). This is represented by the fact that she has a convex indifference curve (see figure 1a). Every bundle on an indifference curve is equally desirable to Renée. In the figure, the line from the y-axis to the x-axis is a budget constraint that is determined by her income and the prices of coffee and cigarettes. Renée maximizes utility when she buys the package of goods corresponding to the point where the budget constraint is tangent to the indifference curve ('A' in the figure).[32]

[31] I am grateful to Arik Roginsky for suggesting a possible connection between mediation techniques and consumer choice theory.

[32] A well-known exception to the claim that the budget constraint is tangent to the indifference curve is cases where there is a "corner solution" – i.e. cases where an agent is unwilling to trade good *X* for any amount of good *Y*.

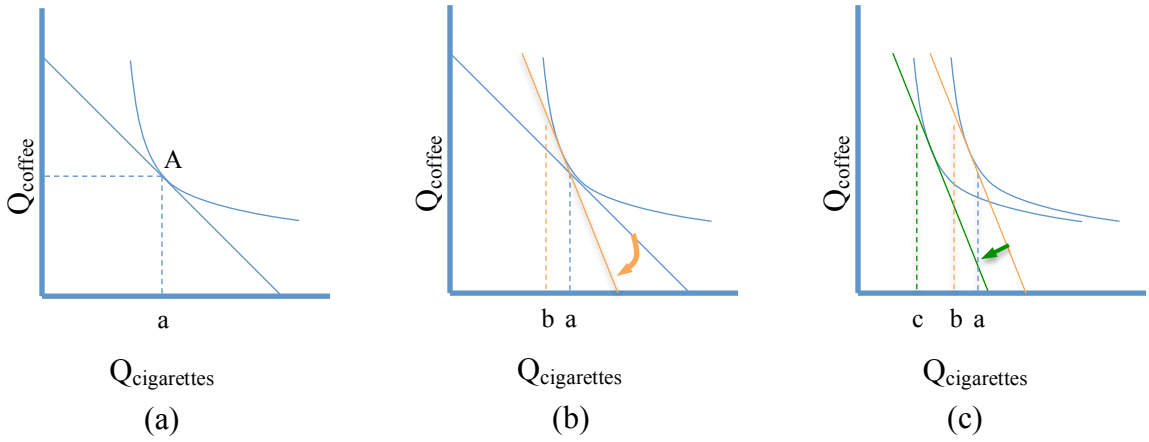

Figure 1 – Not Drawn to Scale

(a) The preferred bundle of goods corresponds to the point 'A' at which the budget constraint meets the indifference curve.

(b) The substitution effect (a – b) is the effect that results from the shift in the relative desirability of the goods while remaining on the same indifference curve.

(c) The income effect (b – c) is the effect of a decrease in buying power while keeping preferences fixed. It orthogonally shifts the budget constraint towards the origin.

The *price effect* for cigarettes is the change in the quantity of cigarettes consumed due to a change in the price of cigarettes. The price effect can be decomposed into two additive effects, the substitution effect and the income effect. Suppose the price of cigarettes increases. The *substitution effect* is the effect of the price change on quantity that results from the fact that cigarettes are now less desirable to Renée relative to coffee. Even if the price change did not reduce Renée's total buying power (so that she could remain on the same indifference curve), she would still buy a package of goods that had fewer cigarettes and more coffee (fig. 1(b)). But the price change *does* reduce her buying power, since prices have risen and her income has not. This is represented by orthogonally shifting the budget constraint closer to the origin (fig 1(c)). She cannot buy a bundle of goods on the same indifference curve, but must buy a bundle of goods on an indifference curve that is tangent to the shifted budget constraint. The *income effect* is the

effect of the price change on the quantity of cigarettes consumed that is due to Renée's decreased buying power. Alternatively, it is the decrease in cigarette consumption that is *not* due to the shift in the relative desirability of the two goods. In figure 1, the substitution effect shifts the quantity of cigarettes consumed from c to b and the income effect shifts it from b to a. The total effect is given by a – c.

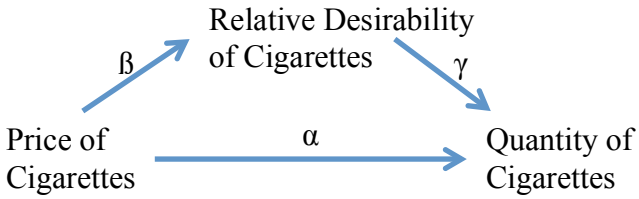

Figure 2

The relationship between the price, substitution and income effects is easily represented using a causal mediation model. Figure 2 is a model with three variables: the price of cigarettes, the relative desirability of cigarettes and the quantity of cigarettes consumed. We will continue to denote the treatment as *X,* the mediator as *M,* and the outcome as *Y.* We will ignore the parameters for a moment.

Relative to the model, the total effect corresponds to the price effect, the direct effect corresponds to the income effect, and the indirect effect corresponds to the substitution effect. The direct effect corresponds to the income effect, since the income effect is the effect the price change would have on the quantity consumed were the value that Renée places on cigarettes relative to other goods not to change. The indirect effect corresponds to the substitution effect, since it is the effect the price change would have on the quantity consumed were Renée's relative preferences to change while her buying power remained the same. The total effect is the price effect.

Let's suppose that the structural equations for the model in figure 2 are linear, as follows:[33]

$$(22)\ x = a_0 + \boldsymbol{\epsilon}_x$$

$$(23)\ m = b_0 + ßx + \boldsymbol{\epsilon}_m$$

$$(24)\ y = c_0 + \alpha x + \boldsymbol{\gamma} m + \boldsymbol{\epsilon}_y$$

There is no reason to suppose that the equations are in fact linear, but the discussion here will easily generalize to any additive model. In additive models, the decomposition of the total effect into the direct and indirect effects is simple, since it is just the sum of the other two effects. Additionally, and relatedly, in additive models there is no need to distinguish between necessary and sufficient effects. The way that these ideas are related is that the total effect is always the sum of the sufficient direct effect and the necessary indirect effect (eq. (19) above). It follows that the sufficient direct effect and the *sufficient* indirect effect make additive contributions to the total effect if and only if the sufficient direct (indirect) effect equals the necessary direct (indirect) effect.

In linear models, the parametric versions of the total, direct, and indirect effects are as follows:

$$(25) TE_{0,1}(Y) = -TE_{1,0}(Y) = \alpha + \beta\gamma$$

$$(26) DE_{0,1}(Y) = -DE_{1,0}(Y) = \alpha$$

$$(27) IE_{0,1}(Y) = -IE_{1,0}(Y) = \beta\gamma$$

The reason that the sufficient and necessary direct effects are not generally equivalent is that they are evaluated relative to different values of the mediator. In models without interaction, however, the activity of the direct path does not depend on the value of the mediator. According to

[33] Note that in equation (22) a_0 and $\boldsymbol{\epsilon}_x$ cannot be independently estimated unless one makes an assumption about the distribution of the error term.

standard economic theory, the magnitude of the income effect does not depend on an individual's relative preferences for goods. So the income effect of moving from price X=x to price $X=x'$ has the same absolute value as the income effect of moving from $X=x'$ to $X=x$.

The additivity of the decomposition of the price effect into the income and substitution effects is a consequence of two standard assumptions. First, the budget constraint for an agent is treated as exogenous, and thus does not depend on the values of other variables in the model. Second, one assumes that the shape of the indifference curve is independent of the budget constraint. For non-economists, it is not difficult to imagine cases in which the second assumption fails. It would fail, for example, if when Renée has less buying power, she is more reluctant to trade cigarettes for coffee. In such a case, the sufficient indirect effect (the substitution effect) would not equal the necessary indirect effect (which has no name in economics). Suppose that the total effect of a decrease in the price of cigarettes from $9 to $8 a pack is to change her consumption from 6 cigarettes a week to 10. The sufficient indirect effect of the price decrease is the increase in cigarettes (over the 6) that would result from the price decrease were there to be no income effect. The necessary indirect effect of decreasing the price is the amount by which the quantity would fall short (of 10) were there to be *only* the income effect.[34] Yet, 10-cigarette Renée feels richer than 6-cigarette Renée. So the answer to the question of how many fewer cigarettes 10-cigarette Renée would smoke in the absence of the substitution effect need not correspond to the number of cigarettes that 6-cigarette Renée would add in the presence of the substitution effect.

---

[34] This claim might initially seem counterintuitive. If the indirect effect effectively holds the values along the direct path corresponding to the income effect fixed, then shouldn't there no be an income effect whether one is evaluating the sufficient or the necessary direct effects? While it is true that in evaluating the sufficient and necessary direct effects one holds the treatment fixed, in evaluating the necessary direct effect, one holds the treatment fixed to its post-price-increase value. The sense in which the necessary indirect effect assumes that there *is* an income effect is that it evaluates the policy relative to a scenario in which it has already influenced the value of the treatment.

From the perspective of mediation, the most important consequence of the assumption that the substitution and income effects make additive contributions is that if one only knows one of them and the price effect, one can calculate the other. For example, the income effect is calculated by subtracting the substitution effect from the price effect. We have already seen that this is not true in general. If the component effects did not make additive contributions, then subtracting the substitution effect from the price effect would not yield the income effect (i.e. the sufficient direct effect), but rather the necessary direct effect.

3.3 TE, DE and IE in Linear Models with Interaction

As we consider a more complex parametric form, it will help to switch to a simpler example. A university discovers that when students meet with their professors before submitting a paper, they get a better grade on the paper. In other words, the total effect of meeting with the professor (as opposed to not doing so) on one's grade is positive. Does this reveal that professors are helping their students write better papers? Not necessarily. Suppose that when a student meets with a professor, this leads the professor to grade his paper more charitably and that this fully accounts for why students who meet with the professor get better grades. If so, then although the policy would raise student grades, it would not do so by making students write better papers.

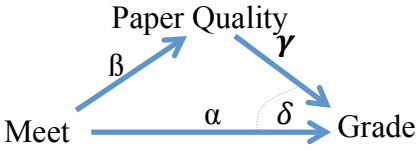

Figure 3

We can represent this case using a mediation model in which the treatment is meeting with the professor, the mediator is paper quality (independent of the grade) and the outcome is the grade (figure 3). The structural equations corresponding to figure 3 are as follows.

(28) $x = a_0 + \boldsymbol{\epsilon}_x$

(29) $m = b_0 + ßx + \boldsymbol{\epsilon}_m$

(30) $y = c_0 + \alpha x + \boldsymbol{\gamma} m + \delta mx + \boldsymbol{\epsilon}_y$

The structural equations are similar to those in the linear case, with an added interaction term, δmx, corresponding to the interaction between the treatment and the mediator.

The sufficient direct effect ($Y_{1,M(0)} - Y_{0,M(0)}$) is the effect that meeting with the professor and writing a paper of the same quality one would have written without the meeting. The parametric version of the sufficient direct effect is as follows:

(31) $DE_{0,1} = \alpha + b_0\delta$

It is unsurprising that the direct effect depends on α. The second term requires some explanation. Since the direct effect is the effect of meeting in the case where one would have written a paper of non-meeting quality, and the values of the treatment and the mediator interact in causing the outcome, the value of $DE_{0,1}$ depends on the non-meeting quality of the paper. For example, suppose that the worse a paper is, the more room there is for the grade to benefit from the teachers disposition. The teacher grades a group of freshmen papers and a group of sophomore papers. In the absence of the meeting, the freshman would have written worse papers. Since the direct effect is the effect of meeting on the grade given the non-meeting-quality paper and the teacher grades worse papers more charitably, the direct effect will be greater for the freshmen.

It is easy to see why the direct effect depends on the value of b_0 by considering its potential outcomes representation, $Y_{1,M(0)} - Y_{0,M(0)}$. The coefficient b_0 corresponds to M_0. The

dependence of the direct effect on b_0 is a consequence of the fact that it depends on M_0. The dependence of the direct effect on M_0 reveals a subtlety regarding the sense in which the direct effect is independent of the behavior or the path going through the mediator. The direct effect is independent of the behavior of the indirect path in that it does not depend on the way that the variables along the path *change* in response to the treatment. That is, it does not depend on the values of ß and $\boldsymbol{\gamma}$. It nevertheless depends on the mediator in the sense that it is sensitive to the value that the mediator would have in the absence of the treatment. This sensitivity is captured by the interaction term δ, which is multiplied by b_0.

The parametric version of the necessary direct effect is as follows:

$$(32) - DE_{1,0} = \alpha + (b_0 + \beta)\delta$$

The key difference between the sufficient and necessary direct effects is that they are evaluated relative to different values of the mediator. The necessary direct effect is evaluated relative to M_1, which in this example equals $(b_0 + \beta)$. If we distribute the interaction term, we see that the difference between (31) and (32) is given by the term $\beta\delta$.

In plain English, the difference between the sufficient and necessary direct effects is that in the former we are asking how much one would *gain* by the professor's charitability and in the latter we are asking how much one would *lose* if one were to not have it. When we ask how much one would gain we compare it to the case where one writes the non-meeting-quality paper and meets with the professor. When we ask how much one would lose, we compare it to the case in which one wrote the meeting-quality paper, but didn't get the advantage of the professor's charitability. These are different, because the two effects are being compared relative to papers of different quality and the quality of the paper interacts with the direct effect of the meeting on the grade. The difference in the quality of the papers is given by ß and the interaction of the

direct effect with paper quality is δ, so the difference between the necessary and the sufficient direct effect is the product of these coefficients.

Here's a story according to which the sufficient direct effect would be higher than the necessary direct effect. Imagine that when a paper is very good, the teacher will give it an A whether or not the student meets with her, but if the paper is average, the she will give it a B with the meeting and a B- without. Further suppose that the student would write a very good paper with the meeting and an average paper without. Then there will be no necessary direct effect, since from the perspective of the student who met and wrote the very good paper, there would have been no harm from losing the professor's charitability. Had the professor forgotten about having met, this student would have been as well off. There would, however, be a sufficient direct effect. A student considering whether to meet would be well advised to meet even if the meeting doesn't benefit the paper. If the student meets and then writes a non-meeting quality paper, he still benefits.

The indirect effect is the effect of meeting with the professor and only benefiting insofar as the meeting improves the paper quality (i.e the professor grades the paper as if she had never met with the student). The indirect effect would be of interest to a student who is considering meeting with a tutor who is just as good an improving the paper as the professor. The lower the indirect effect, the more incentive there is to meet with the professor over the tutor.

The parametric version of the sufficient indirect effect in this case is straightforward:

$$(33)\ IE_{0,1} = \beta\gamma$$

The indirect effect is just the change in the mediator resulting from the treatment multiplied by the change in the outcome resulting from the change in the mediator. The necessary direct effect is as follows:

$$(34)\ -IE_{1,0} = \beta\gamma + \beta\delta$$

The reason for the difference between (33) and (34) is that these effects are evaluated relative to two values of the treatment and the treatment interacts with the mediator. So changes in the value of the mediator corresponding to ß will have different effects proportional to the interaction term δ.

The reason that $-IE_{1,0}$, but not $IE_{0,1}$, depends on δ is parallel to the reason that $-NDE_{1,0}$, but not $NDE_{0,1}$ depends on ß. Just as with the two types of direct effects correspond to two values of the mediator, the two types of indirect effects correspond to two versions of the treatment. In both cases, the difference between the sufficient and necessary versions is the result of treatment mediator interaction, and is given by ßδ.

A comparison of the parametric versions of the direct and indirect effects to the total effect is illuminating.

$$(35)\ TE_{0,1} = \alpha + (b_0 + \beta)\delta + \beta\gamma$$

Since b_0 is just the value of the mediator when X=0 and we haven't said anything about the values of the mediator, we can stipulate that the mediator is defined in such a way that b_0=0. Given this assumption, the total effect is:

$$(36) TE_{0,1} = \alpha + \beta\delta + \beta\gamma$$

The first and third terms are the same as they are in the total effect for the additive model ((25)) and in the case where b_0=0, they correspond to the direct and indirect effects, respectively. The middle term corresponds to the degree of interaction and what is responsible for the non-equivalence of the necessary and sufficient versions of the direct and indirect effects. The middle term is the reason that the total effect cannot be neatly divided up into path specific effects. The reason that it is nevertheless possible to define $DE_{0,1}(Y)$ and $IE_{0,1}(Y)$ is that when one is

evaluating the effect of either the treatment or the mediator and one holds the other fixed, one does not need to take the magnitude of the interaction term into account. Counterfactually, the treatment and mediator always interact – the effect of one always depends on the value of the other. Yet, when evaluating a case in which one of them is held fixed, the degree of interaction δ does not matter, since the value of the fixed variable doesn't vary.

4. Mediation and Extrapolation

4.1 DE and IE as guides to Highest Reduction Potential

The next chapter is devoted to the question of how the quantities measured using mediation techniques are useful for extrapolation. Here I will begin by discussing the most obvious sense in which direct and indirect effects are relevant to extrapolation. Learning about direct effects enables one to extrapolate to scenarios in which the indirect path is disabled, and learning about indirect effects enables on to extrapolate to scenarios in which the direct path is disabled. The very definitions of direct and indirect effects entail facts about the relationship between populations with different causal structures (i.e those in which paths are and are not disabled). This type of extrapolation is of limited usefulness. First, given that measuring direct and indirect effects typically involves intervening on variables in a population, it is not clear that one is ever able to extrapolate from the total effect to the path-specific effects or vice versa. If discovering a path-specific effect involves transforming that population into one in which there is only the path-specific effect and then measuring it, then which unknown causal quantities remain to be inferred? It is true that there are cases in which one can identify DE and IE without intervention – i.e. cases with limited or on confounding – but in such cases one will presumably be able to identify these effects in the target population without experiment, so extrapolation appears to be

unnecessary. Second, there is a sense in which such effects only allow extrapolations across two states of the *same* population. Third, even where we are only concerned with the relationship between the total effect and the path-specific effects in a population, we will want to know not merely what happens when a path is fully disrupted, but also when its behavior is altered.

Pearl's (2012a) discussion of mediation contains an answer to the third concern. His proposal is that if one knows what would happen were the direct (or indirect) path entirely eliminated, this would also have implications for what would happen were the influence of the path merely attenuated. He briefly notes that mediation techniques enable one to identify (what I will refer to as) the *Highest Reduction Potential* (HRP) of certain policies. For example, if consuming a large amount of carbohydrates causes weight gain directly by increasing the amount of calories that the body does not use and indirectly by making one less likely to exercise, mediation techniques allow one to find the upper bound on how successful a policy encouraging people to exercise could be at reducing the effect of carbohydrates on weight. The proposal here is that the HRP of the policy corresponds to the necessary indirect effect, which corresponds to the decrease in the total effect that would result from disabling the indirect path (but not the direct one). While policies encouraging exercise might not fully offset the reduction in exercise that results from carb consumption, the benefit that would result in the case where it fully offset this reduction informs us of how successful the policy would be if it were to be maximally effective.

Here is the passage in which Pearl links mediation to extrapolation:

> Scientifically, mediation tells us "how nature works" and, practically, it enables us to predict behavior under a rich variety of conditions and interventions. For example, an investigator interested in preventing Y may wish to assess the extent to which Y could be prevented by changing an intermediate variable, Z, standing between X and Y, or modifying some intermediate process between X and Z. (Pearl 2012, p. #)

Later on, in the context of discussing linear models with interaction, he shows how to identify the "highest prevention potential" for policies seeking to reduce the influence of a path.

The phrase "highest prevention potential" is apt in situations where one *wants* to reduce the probability of the effect. Here I will use the more general term "highest reduction potential" (HRP) to encompass both cases in which one wants to prevent an effect and cases in which one desires the effect, but seeks to cut corners to reduce expense. Pearl gives a corner-cutting example when he discusses a hypothetical drug company that considers replacing a drug that works in part by producing a catalyzing enzyme with a drug that does not produce the enzyme and wants to evaluate the effectiveness of the replacement drug. Even the term HRP is too narrow, since in cases where the contribution of a path is negative, disrupting it will *increase* the effect. In this section I will focus on cases in which DE and IE are positive, though the results of this section easily generalize. The purpose of this section is to make the point that necessary direct and necessary indirect effects only correspond to HRPs given parametric assumptions.

The causal quantity that plausibly corresponds to the HRP of a policy that seeks to disrupt the indirect path is the necessary indirect effect. This is the decrease in the total effect that would result from rendering the indirect path inactive (as compared to the case in which both paths are active). Since the policy aims to block the influence of the indirect path, if the policy is fully successful, then the decrease in the magnitude of the total effect will equal the magnitude of the necessary indirect effect. To say that the necessary indirect effect corresponds to the *highest* reduction potential, one must add that in cases where the policy is not fully successful in blocking the indirect path, the total effect will not decrease by the same amount. The claim that the necessary indirect effect corresponds to the HRP becomes especially plausible when one recalls that the total effect is equal to the sufficient direct effect plus the necessary indirect effect.

Suppose that the sufficient direct effect of a treatment is to raise the expected value of a dichotomous outcome by .6 and the total effect of the treatment is 1. Intuitively, the most that a policy to block the indirect path could do would be to reduce the total effect by .4.

To make this more concrete, let's return to the grading example. The treatment is whether a student meets with the professor, the mediator is the quality of the paper he writes and the outcome is the grade she gives the paper. The indirect effect is the effect of the meeting on the grade for which the improvement in paper quality is sufficient. The direct effect is the portion of the total effect for which the professor's increase in charitability as a result of the meeting is sufficient.

Suppose our paper-writing student is considering an action that will reduce the effectiveness of his meeting with the professor on the quality of his paper. For example, perhaps he is being initiated into a fraternity and as part of the process he is required to consume three shots of vodka right after meeting with any professor. Doing so would reduce the student's recall of the meeting and make him less able to benefit from the professor's advice. Taking the shots will reduce the contribution of the indirect path without necessarily eliminating it. Since the indirect path will still plausibly have some influence, the direct effect will not identify the effectiveness of the meeting for the student. If we assume, however, that the scenario in which his grade improves least is the one where his paper quality does not improve at all based on the meeting, then the direct effect identifies the effectiveness of the meeting on the grade in the worst-case scenario. The HRP of the action is then given by the necessary indirect effect. If X=0 indicates not meeting and X=1 indicates meeting, $HRP_{vodka} = TE_{0,1} - DE_{0,1} = -IE_{1,0}$.

While the necessary indirect effect plausibly corresponds to the HRP of actions that block the indirect path, similar reasoning yields that the necessary direct effect corresponds to the HRP

of a policy that blocks the direct path. Suppose the professor is considering a policy in which the students write their ID numbers rather than their names on their papers. Although she might still be able to identify some of the students based on their writing styles, such a policy would make her less likely to know the identity of the paper writer and would therefore decrease the average bump that students who meet with her get as a result of her increased charitability. If the policy were to entirely eliminate the charitability bump, the effectiveness of meeting with the professor would be the indirect effect and the decrease in student grades would correspond to the necessary direct effect. Under the assumption that student grades are reduced the most in the case where there is no direct effect $HRP_{policy}=TE_{0,1} - IE_{0,1} = -DE_{1,0}$.

One assumption that one must make in order for the necessary effects to correspond to the HRPs of the different types of policies has to do with the type of policy or actions that one considers. Suppose that the student's drinking vodka after meeting with the professor would not merely hinder him from benefiting from the meeting via writing a better paper, but would actually make him write a worse paper than he would have written in the absence of the meeting. If so, then the highest reduction potential of drinking vodka could be even greater than the necessary indirect effect, since this only measures the amount he would lose if he wrote a paper as bad as the one he would have written without the meeting, but his paper could be even worse. So in considering policies that disrupt indirect effect, we need to be thinking of policies that might make the mediator unreceptive to the change in the treatment, not those that lower (raise) the value of the mediator below (above) the value it would have had in the control scenario. Similarly, in thinking about policies that disrupt the direct path, we need to be thinking about policies that at worse (or at best) disrupt the transmission of the effect through the direct path. It

is plausible that the professor's anonymity policy would at best reduce the direct effect to 0, and that it would not render the direct effect negative.

It is clear when Pearl talks about the necessary effects as having the highest reduction potential, he is only thinking of policies that block the transmission of the paths, rather than reversing the direction of their effects. While it is important to make explicit that the HRP proposal only applies to certain types of policies, this is a clarification of the proposal rather than one of its limitations.

Here's an example in which the HRP of drinking vodka would not correspond to the necessary indirect effect. Imagine that the paper the student would write in the case where he drinks vodka after meeting is slightly better than the one he would have written had he not met with the professor, but not as good as the one he would have written had he not taken the shots. Further imagine that the slightly better paper would receive a *worse* grade. Perhaps the improvement in writing makes it more salient to the professor exactly how confused the student is. If so, then the vodka-drinking student would lose more points in the case where the meeting is slightly effective at improving the paper than he would in the case where it has no influence on the paper quality. That is, it is better for the student for there to be *no* indirect effect than it would be for there to be a small indirect effect. It follows that the HRP for drinking is not given by the necessary indirect effect. Rather, the HRP of drinking is the decrease in the grade in the case where the meeting leads to a slight improvement in the paper.

If it is the case that some decreases in the value of the mediator increase the value of the outcome and others decrease it, one cannot assume that the necessary indirect effect corresponds to the HRP of policies that disrupt the indirect path. In the example just given, some improvements in paper quality lead to a higher grade and some do not. Consequently, it is

possible for the vodka-drinking policy to not have its full effect of obliterating the student's memory of the meeting, but lead to more of a decrease in the student's grade than what would have resulted from the policy's having its full effect.

In short, the HRPs for path-disrupting policies do not always equal the necessary direct or necessary indirect effects. Given common parametric assumptions, however, they will be equal. For example, suppose that the model for this case contains linear parameters plus an interaction term, as I suggested above. Here they are again for reference:

(37) $x = a_0 + \epsilon_x$
(38) $m = b_0 + ßx + \epsilon_m$
(39) $y = c_0 + \alpha x + \gamma m + \delta mx + \epsilon_y$

The necessary indirect effect is $ß(\gamma+\delta)$. $(\gamma+\delta)$ is a constant. Suppose for a moment that both ß and $(\gamma+\delta)$ are positive. Policies that seek to disrupt the indirect path by intervening on the mediator will change the value of ß. A maximally effective policy will make ß equal zero. The reduction in the total effect due to that policy is $ß(\gamma+\delta) - 0(\gamma+\delta) = ß(\gamma+\delta)$. A non-maximally effective policy will change the value of ß to some lower positive value ß'. The reduction in effect due to such a policy will be $ß(\gamma+\delta) - ß'(\gamma+\delta)$, which is less than $ß(\gamma+\delta)$. So given the assumptions, the necessary indirect effect corresponds to the HRP of a policy that disrupts the indirect path.

If one of ß or $(\gamma+\delta)$ is negative, then disrupting the indirect path will increase the total effect rather than decreasing it. In such cases, $ß(\gamma+\delta)$ corresponds to the highest potential for increase rather than the HRP. For example, if ß is positive and $(\gamma+\delta)$ is negative, then $ß(\gamma+\delta)$ will be negative. A policy that reduces $ß(\gamma+\delta)$ to 0 will increase the total effect by $ß(\gamma+\delta)$. Using reasoning similar to that given in the last paragraph, the increase that results from

successfully setting ß to 0 will be higher than the increase from any policy that changes ß to some value ß' that is between 0 and ß.

Here's a set of equations for which the necessary indirect effect does not correspond to the HRP of a policy that disrupts the indirect path. Suppose that the possible values for paper quality are M={0,1,2,3} that the paper is graded out of 10 and that the equations for the mediator and the outcome are as follows (the treatment exogenously set to *X*=0 or *X*=1):

$$(40)\ m = 3x$$

$$(41)\ y = x + (m-1)^2 + 5$$

M_0=0 and M_1=3. In evaluating the natural indirect effect, one holds the treatment at *X*=1. In calculating it, Equation (41) can therefore be simplified to $y = (m-1)^2 + 6$. The necessary indirect effect is calculated like so:

$$(42) - IE_{1,0} = Y_{1,M(1)} - Y_{1,M(0)} = [(3-1)^2 + 6] - [(0-1)^2 + 6] = 10 - 7 = 3$$

If the student were to lose the full indirect benefit of meeting with the professor, this would hurt his grade by 30%. Now imagine that the student's drinking reduces the benefit of meeting on the quality of his paper so that the meeting improves his paper quality by 1 point as opposed to 3. Then his paper grade would be $[(1-1)^2 + 6] = 6$. The student's action reduces his grade by 40%, instead of 30%, so the necessary indirect effect does not correspond to the HRP in this case.

Thus far I've primarily discussed policies that seek to disrupt the indirect path. What about policies that seek to disrupt the direct path? It is straightforward to check that in the parametric model presented the HRP of such a policy is the necessary direct effect. The necessary direct effect is $\alpha + (b_0 + \beta)\delta$. Policies aiming to disrupt the direct path will influence neither the value of the mediator nor the way that the value of the treatment interacts with the

value of the mediator, so in this case such policies will only influence α. Clearly, the policy will reduce the effect by the most when it sets α to 0.

Does the correspondence between the necessary direct effect and policies that disrupt the direct path depend on parametric assumptions? The answer to this question is somewhat complicated. When we considered policies that disrupt the indirect path, it was easy to treat the success of the policy as corresponding to how much it is able to move the value of ß towards 0. That is, the necessary indirect effect identifies the HRP of a policy that seeks to disrupt the indirect path just in case it reduces the total effect by more than any policy that reduces the value of ß without getting it all the way down to 0. When we ask the corresponding question about the necessary direct effect, how do we characterize the cases in which the policy is not fully successful at disrupting the direct path? To say that the necessary direct effect gives the *highest* reduction potential of policies that try to disrupt the direct path we need to specify what the reduction potential is in cases where the policy is less successful. Without measuring a mediator along the direct path, I see no way to specify this in a way that is both non-arbitrary and non-trivial. Consider, for example, the simplest case where the necessary direct effect is given by some positive parameter κ. It is trivial that among policies that do not render κ's value negative, the policy with the HRP will set it to 0. Moreover, even if one were to give some model with a more complicated necessary direct effect, it is not clear how one could say anything more informative than that the case in which the policy has the HRP is the one in which the equation corresponding to the necessary direct effect has its minimal value.

Presumably, it is in principle possible to replace the direct path with a set of indirect paths going through mediators that were not included in the model. If one designed a policy that disrupts the direct path by influencing one of these variables, the scenario would be entirely

parallel to the one in which one seeks to disrupt the indirect path. It might therefore seem trivial to show that the equivalence of the necessary direct effect and the HRP of policies that disrupt the direct path depends on the same parametric assumptions upon which the equivalence of the necessary indirect effect and the HRP of policies that disrupt the indirect path depends. But in order to talk about parameters, we need to be talking about particular models. That fact that it is possible to distinguish between the levels of success for policies disrupting a path in a model containing mediators along every path does mean that one can do so in a model with a direct path that, by definition, contains no mediators.

The problem with determining whether the magnitude of the necessary direct effect has the highest reduction potential of any policy that seeks to reduce the direct path is that the set of policies being compared is insufficiently specified. I would like to suggest, however, that there is a way to think about the necessary direct effect such that the claim that it corresponds to the HRP of policies that disrupt the indirect path is intuitively plausible, if not precise. In considering whether there is a direct effect, whether of the necessary or sufficient variety, one is asking whether the outcome *responds* to the treatment in a way that does not depend on the mediator. For example, the way that I suggested one might evaluate the case in which the direct path is disabled was by making the teacher entirely unaware of the identity of the students. Given the story I told about the direct effect corresponding to the teacher's charitability, it makes sense that if the teacher were totally unaware of which student wrote which paper, the direct effect would disappear. A rough way to think about the claim that the necessary direct effect gives the HRP is that it has a higher reduction potential than any policy that diminishes the receptivity of the outcome to the treatment without eliminating it. This cannot be spelled out parametrically, but it

does provide us with grounds for evaluating whether the necessary direct effect corresponds to the HRP in a particular case.

In short, mediation techniques do enable one to discover the HRP of certain policies, but only given certain parametric assumptions. It is straightforward to provide the parametric assumption under which the necessary indirect effect corresponds to the HRP of a policy that disrupts the indirect path. It does not appear possible to provide parametric conditions under which the necessary direct effect gives the HRP of a policy that seeks to disrupt the direct path, though I have presented a way that we can try to evaluate the claim that the direct effect does correspond of the HRP of such policies in particular cases.

5. Conclusion

In this chapter, I have provided a technical introduction to causal mediation techniques by providing different ways of representing direct and indirect effects and mapping the relationships among the representations. In the next chapter, I will be primarily concerned with the non-parametric definitions given in terms of potential outcomes, but it will be useful at times to move between the different representations for the sake of illustration. I will also continue to highlight cases in which one can license extrapolations by making parametric assumptions. My reason for emphasizing non-parametric causal inference is not because I believe that parametric assumptions are never legitimate, but rather because I seek to determine whether it is possible to say something general about the relationship between mediation and extrapolation without relying on such assumptions.

Now that I have provided an overview of both mediation techniques and the transportability framework, I can now turn to questions regarding how they relate to one another.

Do mediation techniques have a role to play in enabling extrapolative inferences? How does this role relate to the transportability framework? These questions are the focus of chapter 7.

Chapter 7: Mediation, Transportability and Extrapolation

I began the dissertation with an informal characterization of the problem of extrapolation, which I referred to as the problem of the unknown unknowns. The problem is that a causal effect may diverge across populations as a result of variation in an indefinite number of unknown background factors. Given that in most contexts one will never be able to know about all of these background factors, how is extrapolation ever justified? The discussion of transportability in chapter 4 suggests a two-part answer to this question. First, the probability distribution for a population is defined relative to a set of background factors, such that the effects in the population are average effects across those factors. Through non-parametric causal inference we can identify these average effects without knowing what these background factors are. Second, when two populations diverge in their probability distribution, it is nevertheless sometimes possible to transport some of one's knowledge of the study population to the target population, provided that one has some knowledge of the points at which the populations differ. The reason for this is that variation in one part of a model does not lead to variation in other parts of a model, so learning that two populations differ in certain respects does not undermine one's ability to transfer some of one's knowledge from one population to another. While transportability is a type of extrapolation and general non-parametric causal inference is not, both of these methods enable one to gain causal knowledge without knowing all of the causally relevant background factors.

Not only does the transportability framework enable one to make cross population inferences, but it also helps explain why learning how a cause brings about its effect helps one extrapolate. As we saw in chapter 4, there are cases in where measuring a mediator transforms a non-transportable quantity into a transportable one. As I show in section 2, in such cases the

transportability framework enables one to infer the total, direct and indirect effects in the target population. This might make it seem like mediation techniques have nothing to contribute to extrapolation over and above transportability techniques. The aim of this chapter is to show that this is not the case. Mediation techniques both expand the scope of non-parametric identification and license cross-population inferences involving non-transportable quantities. In order to show how mediation techniques license such inferences, one must introduce a way of individuating populations that is more fine-grained than S-nodes.

I describe three ways that mediation techniques contribute to extrapolation. First, there are cases in which one can extrapolate direct and indirect effects even where the total effect in not transportable. Second, mediation techniques enable us to develop Steel's proposal that when there is just one chain between a cause and its effect, learning about links in the chain is helpful for extrapolating. Steel used comparative process tracing as a means for extrapolating in the case where there is just a single path between a cause and its effect. He was unable to generalize his account to cases in which there are multiple paths. I argue that by identifying the indirect effect, one is in certain respects able to treat the indirect path as if it were the only path, and that this enables one to use comparative process tracing for extrapolating the indirect effect. Third, I argue that the indirect effect is the average effect across both unmeasured background factors and unmeasured mediators along the direct path. While this feature of mediation techniques arguably is more relevant to standard non-parametric causal inference than to extrapolation proper, it does have implications for extrapolating path-specific effects from populations to subpopulations.

This chapter is organized as follows. Section 1 briefly reviews selection diagrams. Section 2 shows how in contexts with limited confounding, the problem of identifying the direct

and indirect effects can be represented as a special case of transportability. Section 3 argues that the conditions across which the natural direct effect is invariant cannot be represented in a selection diagram and introduces new notation for representing these conditions. Section 4 considers the more difficult question of the invariance conditions of the indirect effect. Section 5 shows how one can use mediation techniques to generalize Steel's insights about extrapolation in the single-path case. Section 6 shows that the indirect effect in a population is the average of the effects in all subpopulations. Section 7 offers a speculative proposal for determining the robustness of the indirect effect across populations by comparing its magnitude to that of the total effect. Section 8 concludes.

1. Review of Selection Diagrams

Pearl and Bareinboim represent the difference between populations using selection diagrams (e.g. figure 1). Selection diagrams contain S-nodes indicating cross-population variation in the structural equation that determines the value of a variable. The absence of an S-node into some variables implies that there is no variation among populations in the frequency of the causal factors responsible for that variable. A causal quantity is transportable just in case one can identify it in the target population based on experiments on the first population and the probability distribution for the second. Two simple cases of transportable quantities are given in figures 1a and 1b. In 1a, the variation due to the S-node makes no difference for the effect of X on Y, so the quantity is *directly transportable.* In 1b, the relationship between X and Y is unconfounded and can therefore be identified in both populations without experiment. It is *trivially transportable.*

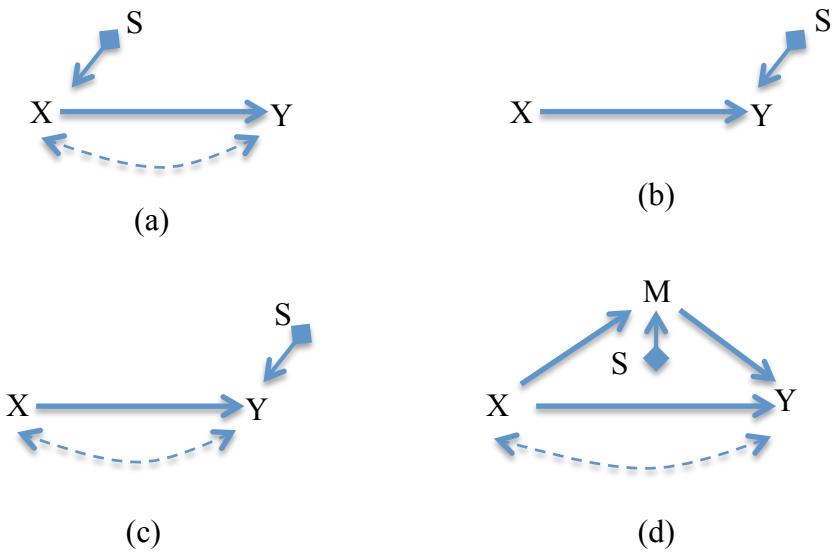

Figure 1

The simplest example of a non-transportable quantity is given in 1c. The S-node indicates that the relevant populations differ in the distribution of unknown causes of *Y*, but the effect of *X* on *Y* is not identifiable in non-experimental populations (due to confounding). Fortunately, in cases where 1c may be replaced with 1d, the effect of *X* on *Y* is transportable. 1c may be replaced with 1d just in case *M* is a mediator between *X* and *Y* and the differences between the populations captured by the S-node in 1c are entirely due to differences in the mediator due to the S-node in 1d. The adjustment formula identifying the effect in the target population is given by the following equation.[35]

$$(1)P^*\big(Y\big|do(x)\big) = \sum_m P(Y|do(x),m)P^*(m|x)$$

The probabilities without asterisks are those from the study population. Those with asterisks are from the target population. "do(x)" indicates that in estimating the conditional probability given in the first term, one must consider the probability distribution that results from intervening on

[35] See appendix to chapter 4.

$X=x$. Note that there is no term on the right hand side with both an asterisk and a do-operator. This reveals that equation (1) is identifiable without any experiments on the target population and that the quantity is therefore transportable.

Figure 1d is a mediation model. The fact that it is sometimes possible to replace a non-transportable quantity with a transportable one by measuring a mediator bodes well for the thesis that mediation techniques facilitate extrapolation. What remains to be seen is whether direct and indirect effects have anything to do with extrapolation. After all, it is possible to demonstrate that the effect of X on Y in figure 1d is transportable without appeal to the terminology drawn from the mediation literature. What do mediation techniques contribute to extrapolation that we could not get using transportability methods?

2. Mediation and Transportability in Contexts with Limited Confounding

Any attempt to relate direct and indirect effects to transportability faces a preliminary notational problem. Transportability concerns whether an expression with do-operators can be identified in the target population without doing experiments on that population. We have seen, however, that the general definitions for direct and indirect effects in the mediation literature cannot be adequately given using formulas with do-operators (Chapter 6, §1.4). So it is not possible to write out the general expressions for DE and IE and use the rules of the transportability framework to prove that they are or are not transportable relative to a selection diagram. Nevertheless, one can sometimes determine that DE and IE are identifiable in the target population without experiment. In such cases, we can speak loosely and refer to DE and IE as being transportable.

A simple example of a case in which DE and IE are "transportable" is the selection diagram in figure 1(d). Equation (1) is valid for all values of *X*, so the total effect of *X* on *Y* ($Y_1 - Y_0$) is transportable. We can represent the direct effect in the target population as follows:

$$(2)\ DE^*_{0,1} = Y^*_{1,M^*_0} - Y^*_{0,M^*_0}$$

Note that there are asterisks next to every potential outcome, including those relating the mediator to the treatment. Since there is no selection node into *Y*, the relationship between the outcome and its direct causes will not vary between the populations and we can replace (2) with (3).

$$(3)\ DE^*_{0,1} = Y_{1,M^*_0} - Y_{0,M^*_0}$$

Equation (3) reveals that any difference between the direct effects across the populations is a result of a difference in the value of M_0. Since there is no treatment mediator confounding in the selection diagram, M^*_0 is identified by P*(M|X=0). As the direct effect in the target population only differs from the direct effect in the study population by a term that is identifiable in the target population (M_0), it is possible to identify the direct effect in the target population without experiment. Moreover, since one can identify the value of the mediator for any value of the treatment, one can also identify the indirect effect ($Y_{0,M^*_1} - Y_{0,M^*_0}$).

Interestingly, given the selection diagram in figure 1(d) one can treat the derivation of the direct and indirect effects within a *single* population as a special case of transportability. Consider again the probabilistic expression for the natural direct effect in a population with no confounding ((9) in chapter 6):

$$(4)NDE_{0,1} = \sum_m [E(Y|X=1, M=m) - E(Y|X=0, M=m)]P(M=m|X=0)$$

This equation takes the difference in the value of the outcome given each value of the treatment and weights it according to the different values the mediator takes on when *X*=0. This weighting enables one to determine the effect of *X* on *Y* when the mediator has the distribution it would have given *X*=0*,* rather than the distribution that it has in the actual population. The weighting term in (4) plays exactly the same role as the weighting term $P^*(m|x)$ in (1). Equation (4) considers the difference between two values of the treatment variable, rather than the whole distribution of P(Y|do(X),M), but otherwise (4) is just a special case of (1).

While (4) is the equation for a population with no confounding, in 1(d) there is treatment-outcome confounding. This is easily corrected for by replacing (4) with the following.

$$(5) NDE_{0,1} = \sum_{m} [E(Y|X = 1, M = m) - E(Y|X = 0, M = m)]P(M = m|do(X = 0))$$

Equation (5) makes precise the sense in which in figure 1(d), the derivation of the direct effect is a special case of the transportability of the total effect across populations. One transports the total effect to a population in which the indirect path is blocked.

The derivation of the indirect effect is similarly a special case of transportability.

$$(6) IE_{0,1} = \sum_{m} \begin{matrix} E(Y|X = 0, M = m)P(M = m|do(X = 1)) - \\ E(Y|X = 0, M = m)P(M = m|do(X = 0)) \end{matrix}$$

The equation is more complicated, but it is fundamentally the same in that it requires one to reweight the mediator (twice) in order to find the magnitude of the total effect in a population where the direct effect is blocked.

This discussion reveals that given the selection diagram in 1(d), mediation techniques have nothing to contribute to extrapolation over and above transportability methods. It is possible to transport the total, direct and indirect effects across populations. It is impressive that this is

possible, but the fact that the transported quantities are the direct and indirect effects plays no role in enabling one to extrapolate.

One should not read too much in to the fact that the derivations of the direct and indirect effects are special cases of transportability in cases with limited confounding. Recall that it is straightforward to provide probabilistic expressions for the direct and indirect effects in cases with no confounding and that difficulties only arise in contexts where one needs to add do-operators to avoid confounding. Although there is treatment-outcome confounding in 1(d), there is no *X*-*M* or *M*-*Y* confounding, and it is these types of confounding that make it difficult to identify the direct and indirect effects. The transportability of direct and indirect effects across the populations represented in selection diagram 1(d) tells us more about the relative ease of identifying direct and indirect effects in cases of limited confounding than about an important conceptual relationship between mediation and transportability.

3. Invariance Properties of the Direct Effect

In section 2, we considered a selection diagram in which the total effect (as well as DE and IE) was transportable. Here we will consider a case in which the total effect is not transportable, and consider the conditions under which the direct effect can be inferred across populations. This case is given in the selection diagram figure 2. The only difference between figure 2 and figure 1(d) is the addition of a bidirected arc between the treatment and the mediator. This bidirected arc makes it impossible to identify M_0^* without intervening on *X*. One needs to know this quantity in order to identify the total, direct and indirect effects in the target population. Therefore, the total effect is not transportable, and, speaking loosely, neither are the direct nor indirect effects.

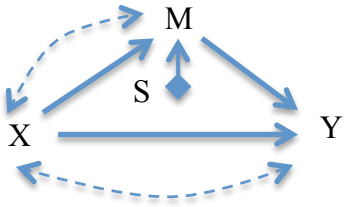

Figure 2

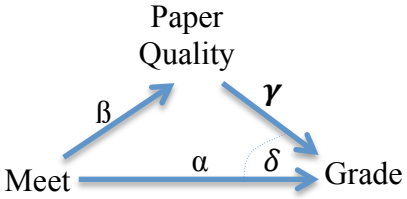

Figure 3

While none of the effects are transportable, it is possible to specify ways that populations might differ in their value of the mediator without differing in their direct effect. To illustrate this, let's return to one of the parametric examples from the last chapter. Figure 3, gives the DAG for the effect of meeting with the professor on one's grade, with a mediator corresponding to the quality of the paper. The Greek letters indicate the parameters for the study population. The selection diagram in figure 2 gives the difference between the study population and various study populations. The structural equations are:

$$(7)\ x = a_0 + \epsilon_x$$
$$(8)\ m = b_0 + \beta x + \epsilon_m$$
$$(9)\ y = c_0 + \alpha x + \gamma m + \delta mx + \epsilon_y$$

The natural direct effect in this case is the effect that meeting with the professor would have on one's grade if one were to write a paper of the same quality as the paper one would have written without the meeting.

Let's now consider two ways that a target population might differ from the study population. First, suppose that in the target population, the paper that the students would have written without meeting is of a different quality than that of the non-meeting paper for the students in the study population. This might occur if the study population consists of freshmen and the target population consists of sophomores. This difference is reflected in a different value of b_0, which is the value of the mediator when X=0. Second, suppose that the two populations are

identical in their value for b_0, but differ in how much their paper would improve based on the meeting. This might occur if one of the groups is more obedient to authority and therefore more wiling to take the professor's comments seriously. The difference between these populations would be reflected by a difference in the value of ß.

The direct effect is invariant across populations that differ in the second way, but not invariant across populations that differ in the first way. That is, the direct effect is invariant to differences in ß, but sensitive to differences in b_0. This is clear from the parametric version of the direct effect ((31) in chapter 6):

$$(10)\ DE_{0,1} = \alpha + b_0\delta$$

The significance of this for extrapolation is that if one knows that two populations have a common value of b_0, the direct effect will be the same across the populations even if the distribution of the mediator differs as a result of a difference the value of ß. We can refer to a difference in b_0 as a *baseline difference* in the value of the mediator and a difference in ß as a *structural difference* in the treatment-mediator relationship.[36]

The invariance of the direct effect to structural differences in the treatment-mediator relationship is not limited to models with the parametric form just considered. We can specify the baseline value of the mediator non-parametrically as M_0 and the structural differences in the *X*-*M* relationship correspond to the quantity M_1–M_0. That the direct effect is sensitive to the former, but not the latter, is evident from its non-parametric definition ((7) in chapter 6):

$$(11) NDE_{0,1}(Y) = Y_{1,M(0)} - Y_{0,M(0)}$$

[36] This distinction resembles Morgan and Winship's (2007, pp. 46-48) distinction between *baseline bias* and *differential treatment bias,* which Xie et al.'s (2012) refer to as Type-I and Type-II bias, respectively. Here I am not referring to estimation and therefore not explicitly concerned with bias, but the distinctions are similar insofar as they both differentiate between baseline differences and structural differences. See Pearl (forthcoming) for further discussion.

Since M_0 appears in the definition and *M* interacts with *x*, the direct effect depends on M_0. Since M_1 does not appear in the definition, the direct effect does not depend on the difference between M_1 and M_0.

It is significant that the invariance conditions for the direct effect cannot be represented in a selection diagram. In a selection diagram, an S-node into *M* indicates that the factors determining the value of *M* can differ among populations in any arbitrary way. Direct effects are not invariant to arbitrary changes to the value of the mediator, but only to changes in the structural relationship between the treatment and the mediator. While transportability concerns the invariance of causal quantities to changes in variables, mediation techniques identify quantities that are invariant to changes in parameters. Or, more precisely, the effects identified by mediation are invariant to cross-population variation in the structural relationships that are represented by parameters.

In order to graphically represent the invariance conditions of the direct effect, we need to introduce a new notational device. Specifically, we need a way to specify that two populations do not differ in the baseline value of the mediator. This can be done with D-nodes (figure 4). D-nodes indicate that two populations do not differ in the distribution for a particular variable when all exogenous variables in the model are set to their *default* value. In figure 4, the only exogenous variable is *X* and the diagram indicates that the default value of *X* is 0. There is nothing in the model that privileges *X*=0; the default value of the exogenous variables must be supplied independently.

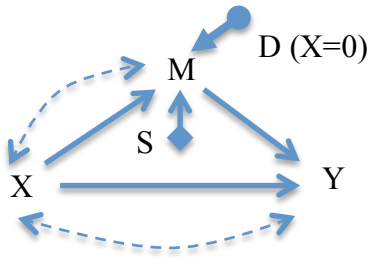

Figure 4

The direct effect is invariant across populations represented by the selection diagram in figure 4. The S-node alone corresponds to changes in the value of the mediator that may or may not lead to changes in the direct effect. The D-node indicates that the populations are similar in their value (or distribution) for $M_{0,}$ thereby eliminating the ways that the mediator could vary that would make a difference for the direct effect.

Here I only consider cases in which there is a D-node into the mediator and a single exogenous variable, but there is a straightforward generalization to other variables in models with many exogenous variables. Given a model with exogenous variables $X_1,\ldots,X_n$ and default values $x_1,\ldots,x_n$, a D-node into variable Y indicates that the potential outcome $Y_{x(1),\ldots,x(n)}=y$ (or the probabilistic distribution $P(Y_{x(1),\ldots,x(n)}=y)$) is invariant across the populations represented by a selection diagram.

The idea of specifying the default states of a variable in a model is inspired by Menzies (2007). Menzies, among others (Hall, 2007; Halpern, 2008; Halpern and Hitchcock, 2012), has argued that claims about actual causation – e.g. the claim that X happened and Y happened and X caused Y – must be relativized to default states of a system. Here I take no stand on how to explicate claims about actual causation. The concept of the default state of a variable is nevertheless useful for understanding direct effects. The direct effect is the portion of the total effect that is invariant to changes in the mediator that result from *changes* in the treatment. If we stipulate that in evaluating $TE_{0,1}(Y)$ we treat X=0 as the default state, it becomes easier to distinguish between populations that differ in the way the mediator changes in response to the treatment and populations that differ in other ways with respect to the mediator. Given this stipulation, the direct effect is invariant across all changes to the mediator that preserve its default state.

One might worry that talk of default states introduces an arbitrary element into the discussion (cf. Blanchard and Schaffer, Forthcoming). Since there is nothing in the model that privileges certain states as being the default state, it appears that the default state must be specified based on non-causal considerations. While I am in general sympathetic to this concern about default states, it does not undermine my appeal to them here. What matters for extrapolation is not the particular value that one selects as the default, but rather that the populations do not differ with respect to whatever that value is.

D-nodes differ from S-nodes in that while the presence of an S-node indicates the *presence* of cross-population variation, D-nodes indicate the *absence* of a specific form of cross-population variation. D-nodes are only useful when combined with S-nodes. For variables in which there are no S-nodes, one can place D-nodes corresponding to any possible default value of the exogenous variables without changing the content of the selection diagram. D-nodes limit the ways that a variable with an S-node may differ across populations.

The most important similarity between S-nodes and D-nodes is that both allow one to represent cross-population differences without making parametric assumptions. While the default value of a variable given particular settings of exogenous variables depends on the structural relationships among the variables, one can insert D-nodes into a selection diagram without making any claims about the parametric form of the structural equations. This reveals that it is possible to provide a non-parametric specification of population differences that is more fine-grained than the specification that could be provided with S-nodes alone. For our purposes, the crucial point is that D-nodes enable us to specify the invariance conditions for the direct effect.

More research needs to be done in order to determine the conditions under which the assumption that the mediator shares a common default across two populations is justified. Cases

in which there is both an S-node and a D-node into the mediator are those in which populations are similar in how they act in the absence of the treatment, but potentially vary in how they respond its presence. For example, if one is evaluating the total effect of influenza on death and the measured mediator is temperature, it is plausible that the average temperature in two populations among people without the flu is around 98.6°F, but that the populations differ in how much getting the flu increases body temperature. In such a case, the default value of the influenza variable would be that one does not have the flu and there would be a D-node into the variable for body temperature.

One way that knowledge of the systems studied by the mechanists discussed in chapter 5 might be relevant to extrapolation is that mechanisms often seem to have well-specified default states. For example, the default state of a neuron is to not fire in the absence of being triggered by the firing of another neuron. Suppose that most neurons are equally unlikely to fire in the absence of being triggered, but that there is widespread variation in whether neurons will fire when they *are* triggered. This variation might be due to the fact that there are different ways that neurons can malfunction and thus cease to fire in response to being triggered. If neurons tend to be relatively causally homogenous in their default states, but differ how they respond to a causal factor, then the representational framework developed here would be helpful for modeling their behavior. To the extent that it is common for mechanisms to have well-specified default states such that populations are relatively homogenous in how they behave when in the default state, mechanistic knowledge will be helpful in extrapolating direct effects.

4. Invariance Properties of Indirect Effects

Let's now turn to the question of what role indirect effects play in extrapolation. In this section I explore the possibility of giving indirect effects an analogous treatment to our treatment of direct

effects in the previous section. That is, just as the direct effect is invariant across certain population differences that cannot be represented within a selection diagram (without D-nodes), the indirect effect is also invariant across population differences not captured by selection diagrams.

While in evaluating the invariance conditions for the direct effect, the obvious proposal was that it is invariant across cross-population changes in the mediator, in thinking about the indirect effect no similarly appealing proposal jumps to mind. The indirect effect will certainly not be invariant to variation in the structural relationship between the treatment and the mediator. And if the populations vary in the equations determining the value of the outcome variable, this will potentially change both the direct and indirect effects. It follows that the indirect effect is neither invariant across populations whose differences are represented by an S-node into *M,* nor across populations with an S-node into *Y.* I can think of two other proposals. The first is that the indirect effect is invariant to variation in the default value of the mediator. The second is that the indirect effect is invariant to variation in unmeasured mediators along paths that are captured by the direct effect. I will consider these in turn.

In the linear model with interaction that we have been considering, the indirect effect is invariant to variation in the baseline value of the mediator. The parametric equation for the indirect effect is simply $\beta\gamma$, which does not contain b_0. There are two reasons to think that this invariance might hold more generally. First, the only way that the value of b_0 makes a difference in the total effect is in combination with the interaction term ($TE_{0,1}(Y) = \alpha + \boldsymbol{b_0\delta} + \beta\delta + \beta\gamma$). In evaluating the indirect effect, the value of the treatment is held fixed, so the magnitude of the interaction term should not matter. Second, we saw that the influence of b_0 on the outcome is already incorporated into the direct effect ($DE_{0,1}(Y) = \alpha + b_0\delta$).

These considerations suggest a more general proposal. While the direct effect is invariant to variation in $M_1 – M_0$, but not to variation in M_0, the indirect effect is invariant to variation in M_0, but not to variation in $M_1 – M_0$. If this proposal were correct, then any cross-population in the distribution of the mediator could be decomposed into two distinct parts: the part that influences the direct effect and the part that influences the indirect effect. This would facilitate extrapolation, because in cases where there was an S-node and no D-node into the mediator, and in which the total effect was non-transportable, one could infer that at least one of the direct and indirect effect will vary across the populations by less than the total effect. The direct effect would vary less than the total effect in cases where at least part of the cross-population variation in the total effect is due to variation in $M_1 – M_0$. The indirect effect would vary by less than the total effect in cases where at least part of the cross-population variation is due to variation in M_0.

Concretely, in our grading example the indirect effect is invariant to variation in b_0, since the impact of the student's meeting with the professor that results from writing a better paper (rather than from her charitability) does not depend on the quality of the paper that he would have written in the absence of the meeting. Meeting with the professor improves the paper by a certain amount above its baseline value, and the increase in the quality of the paper leads to a corresponding increase in the grade. This invariance property follows from the parametric assumptions that I made, but it is not difficult to see how it could fail to obtain. It is plausible that not all increases in the quality of paper will correspond to similarly large increases in the grade. When a paper is really problematic, a small improvement in quality will lead to a larger increase in the grade than it would if the paper were much better. Consequently, the indirect effect will be sensitive not just to the amount by which the paper improves in quality as a result of the meeting, but also to how good the paper would have been without the meeting.

While it is not in general true that the indirect effect is invariant to the baseline value of the mediator, it is invariant given the assumption that equal changes in the value of the mediator correspond to equal changes in the value of the outcome (given a fixed value of the treatment). This assumption is resembles the parametric assumption that we considered in discussing the conditions under which mediation techniques allow one to discover the highest reduction potential (HRP) for a policy (Chapter 6, §4.1). There we saw that the necessary indirect effect (negative $IE_{1,0}(Y)$) corresponds to the HRP of a policy that seeks to disrupt the indirect path only if it is not the case that some increases in the value of the mediator lead to increases in the value of the outcome and others lead to decreases. The assumption that all increases in the value of the mediator lead to identical changes in the value of the outcome is a much stronger assumption than the assumption that it is not the case that some changes in the mediator are positively relevant to the outcome and others are negatively relevant.

Given the stronger parametric assumption that increases in the mediator lead to similar increases in the outcome regardless of the initial value of the mediator, we can infer a lot about how populations might differ as a result of variation in the mediator. First, as we saw in the last chapter, we can infer that the highest reduction potential of a policy that seeks to disrupt the indirect path is given by the necessary indirect effect. Second, we can infer that at least one of the sufficient direct and the sufficient indirect effect will vary less than that of the total effect.

Moreover, since variation in either the direct or indirect effect entails variation in the total effect, the path specific effects cannot vary *more* than the total effect. If one measures just the direct or just the indirect effect, one cannot be sure that it will vary less than the total effect. It could be that all the variation in the total effect could be due to that effect that you measured. There is also no guarantee that the measured effect does not vary greatly across populations.

What one *does* know is that if one infers that a direct or indirect effect will be similar across populations, one's inference will be at least as reliable – and possibly more reliable – than a similar inference regarding the total effect.

Now that we have considered and refined the proposal that the indirect effect is invariant to changes in the baseline value of the mediator, let's turn to the second proposal, which is that the indirect effect is invariant to changes in the value of unmeasured mediators along paths whose influence is captured by the direct path. The idea here is that just as the direct effect is in some respects invariant to what happens to mediators along the indirect path, the indirect effect is in some respects invariant to what happens along the direct path. Of course, there are no measured variables along the direct path, so there is no variable into which we can place an S-node or D-node in order to represent the proposed invariance conditions for the indirect effect. However, there are, presumably, unmeasured mediators along paths not going through the measured mediator, so we can consider the invariance of the indirect effect to variation in these unmeasured mediators.

Before proceeding, it will help to introduce some new terminology. Let's refer to unmeasured mediators along paths not going through the measured mediator as *effect transmitters*. If one were to include an effect transmitter in the model, it would become a mediator along a path whose influence had formerly been captured by the direct path. The reason for not simply calling effect transmitters "mediators" is that whether something is a mediator is relative to a model and in models with a direct path there are, of course, no measured mediators along the direct path. The reason for not calling them "unmeasured mediators along the direct path" is that this might misleadingly suggest that there is some mediator such that were one to measure it, the direct path would be transformed into a single indirect path. But this might not be

the case. Adding an effect transmitter to a model will not always eliminate the direct path, but will often lead to a model in which there is still a direct path in addition to the added indirect path. The term "effect transmitters" is meant to convey that the denoted variables are the means by which the treatment influences the outcome via the direct effect.

In considering the way that the indirect effect is sensitive to cross-population variation in effect transmitters, we can use reasoning analogous to that used in finding the invariance properties of *direct* effects to cross-population variation in measured mediators. Effect transmitters can vary across populations in terms of their default values or in terms of how they respond to the change in the treatment (or both). The indirect effect is invariant to variation in the way that the effect transmitters respond to the treatment, provided that it is not accompanied by variation in their value given an intervention that sets X to 0. Moreover, this is true not just for a particular effect transmitter, but for the variation of any effect transmitter.

There is something strange about presenting the invariance conditions of the indirect effect in terms of hypothetical mediators that contribute to the direct path. As we saw in chapter 6, a distinctive feature of the indirect effect is that it is possible to identify it without having any knowledge of unmeasured mediators along other paths, or even knowing whether there are such paths. To try and extrapolate indirect effects by making stipulations about what happens along the direct path seems like a step backwards. In the next section I pursue a different approach. Instead of specifying ways that the indirect effect is invariant across cross-population variation in effect transmitters, I consider how the fact that one does not need to know about these effect transmitters in order to identify the indirect effect is relevant to extrapolation.

5. Extrapolation of Single Paths and the Indirect Effect

It will help to momentarily take a step back to consider the question of why we thought that mediation methods might be relevant to extrapolation in the first place. Steel was able to tell a plausible story about how measuring links in a causal chain facilitates extrapolate in cases where there is only a single path between a cause and its effect. By examining particular links in the chain across the populations, one can make a more reliable prediction about whether the total effect will extrapolate. He refers to this as comparative process tracing. While there are some ambiguities regarding how comparative process tracing is supposed to work – for example, it is unclear what basis one has for one's background knowledge about which parts of a mechanism are likely to differ across populations – he seems to be correct in maintaining that one can make better predictions about the mechanism in the target by examining parts of it in the study population, even if one's knowledge does not *guarantee* that the effect will generalize to the target. Yet Steel was unable to tell a compelling story about how the single-path case relates to the cases where there is more than one path between the treatment and the outcome. I will now use causal mediation techniques to fill in the necessary details.

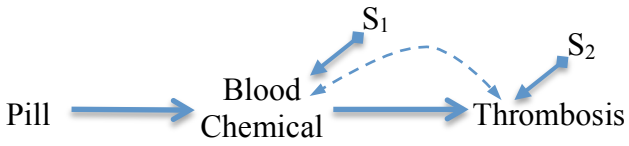

Figure 5

Let's begin by considering the single path case. Figure 5 depicts the effect of taking a particular pill on thrombosis and assumes that it is mediated exclusively via a blood chemical. The effect of the pill on the blood chemical is transportable, but the effect of the chemical on

thrombosis is not, due to the bi-directed arc. Consequently, the total effect of the pill on thrombosis is not transportable.

While the effect of the pill on thrombosis is not transportable in figure 5, there is a clear sense in which measuring the mediator facilitates extrapolation. If one only measured variables corresponding to the pill and thrombosis, there would be two possible sources of variation in the total effect: those corresponding to S_1 and those corresponding to S_2. In the selection diagram in figure 5, the effect of the treatment on the mediator is transportable. So one can divide up the cross-population variation into a part that one can measure and a part that one cannot. Since one does not know how much the populations differ in the effect of the mediator on the outcome, one cannot determine what the total effect will be in the target population. Nevertheless, by measuring the cross-population variation due to the effect of the treatment on the mediator, one can account for one possible source of cross-population variation.

As a special case, suppose that one considers the probability distributions for each population to determine the effect of taking the pill on the presence of the chemical in the blood in each, and one discovers that they are the same. This should increase one's confidence that the total effects will be similar in the populations. It may still turn out to be the case the populations differ in the way that the blood chemical influences thrombosis. But by establishing that one part of the causal chain does not vary between the populations, one eliminates one source of possible variation.

In cases where the total effect is not transportable, the advantage of measuring a mediator along a single path will be comparative: one reduces one's uncertainty regarding how much the populations vary. How much it reduces one's uncertainty depends on one's prior beliefs about how much the total effect varies across the populations and how much the relationship between

the treatment and the mediator varies. Additionally, in cases where one can examine the physical system in which the variables are instantiated, the physical location of the mediator in relation to the treatment and the outcome may serve as a proxy for how likely a causal relationship is to vary across populations. The greater the distance between the two variables, the more points there are at which the causal relationship may be interrupted. The farther the mediator is from the treatment, the more that learning that the treatment-mediator relationship does not vary across populations will reduce one's uncertainty that the total effect of the treatment on the outcome will differ across populations.

Here I have simplified matters by considering a path with three variables, but one could make similar claims about paths with multiple mediators. One can think of the path as a causal chain in which disrupting a particular link will disrupt the total effect. Evaluating particular links enables one to check for possible points of disruption.

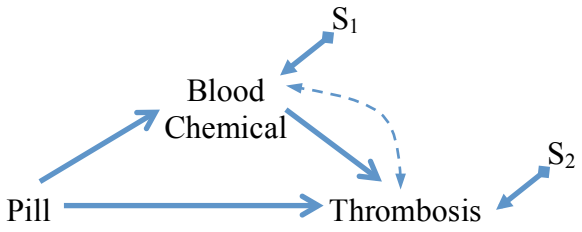

Figure 6

There is more to be said about how measuring more mediators along a single path facilitates extrapolation, but I now want to turn to the question of how the single-path case relates to the multi-path case. Let's suppose that the selection diagram in figure 5 is false and that there are paths from the pill to thrombosis not going through the blood chemical (figure 6). In a twist on Hesslow's (1974) well-known case, let's imagine that this pill turns out to prevent pregnancies and that pregnant women are more likely to develop thrombosis, though the

scientists studying the pill may not know about its usefulness as a contraceptive. In other words, pregnancy is an effect transmitter. So there is at least one way that the pill reduces one's chance of thrombosis via the direct path, though there could be others (figure 7). Moreover, since we are making no parametric assumptions, the effect of the pill on thrombosis on the *indirect path* may be sensitive to whether one has taken the pill and whether one gets pregnant, and also to any other unmeasured mediator that is not on the path going through the blood chemical. These variables could influence the indirect effect via interacting with *blood chemical* in causing thrombosis.

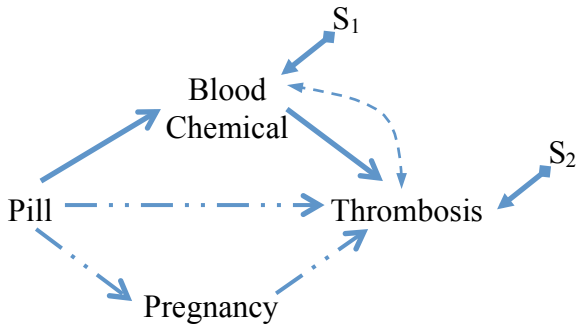

Figure 7 – Pregnancy is a hypothesized variable along a path not going through the mediator. The dashed lines indicate that the causal relations among variables on paths not going through the measured mediator are unknown.

I will now consider the way that variation in pregnancy leads to variation in the effect of the pill on thrombosis via the blood chemical (the indirect effect in figure 6), though what I say will generalize to cases in which there is cross-population variation in other effect transmitters. The values of the treatment variable X are X=0 for not taking the pill and X=1 for taking the drug and we will treat X=0 as the default state of the variable. Two women have the same default value for pregnancy iff they have an equal probability of getting pregnant if they do not take the pill. A common case in which two women would differ in their default value for pregnancy is when one is more sexually active than the other. In addition to differing in their default values,

two women can differ in how effective the pill would be in *lowering* their probability of getting pregnant. These two ways that individuals can differ in the effect of the pill on pregnancy correspond to four types of women. Table 1 presents an individual of each type, for reference.

	Pregnancy Risk if No Pill	Decrease in Pregnancy Risk Due to Pill
Andrea	High	High
Betty	Low	High
Carla	High	Low
Danielle	Low	Low

Table 1

Let's temporarily assume that the differences among the women in table 1 are the only ways that the women differ. That is, all variation between the women is due to differences in their effect transmitters. As I noted in the previous section, the indirect effect is invariant to changes in effect transmitters that preserve their default values. Accordingly, the indirect effect would be the same in Andrea and Carla and the same in Betty and Danielle. The differences in the ways that the women would react to the treatment – i.e. the differences in the right-hand column – do not make a difference in the indirect effect, since the indirect effect is evaluated by holding the treatment fixed at its default value.

So far I have emphasized the way in which the indirect effect is *invariant* across populations in which the default values of effect transmitters are constant. Here I want to emphasize the contrapositive of the claim that if the default values of effect transmitters don't vary, then the indirect effect does not vary either. Namely: if the indirect effect *does* vary across populations, this variation is due to variation in the default values of effect transmitters. Here it is crucial to keep in mind that I am temporarily assuming that all variation across the populations is due to variation in effect transmitters. In the thrombosis example, the only way that cross-population variation in pregnancy can influence the indirect effect is if it is variation in the

pregnancy status of the women in the default case where they do not take the pill. Moreover, this is true not just of pregnancy, but of any effect transmitter.

Let's now consider the way that the indirect effect changes as a result of cross-population variation in effect transmitters. As I have argued, the only type of variation in these variables that makes a difference for the indirect effect is variation in their default values. The only way that the default values of effect transmitters make a difference for the indirect effect is that these variables interact with the measured mediator. When one evaluates the indirect effect, the treatment is held fixed at its default value, and all effect transmitters also remain at their default values. The effect transmitters influence the indirect path not because they change their values (or distributions) in response to the treatment, but because the effect of the measured mediator on the outcome may depend on their fixed values (or distributions). To the degree that cross-population differences in the default values of effect transmitters make a difference in the indirect effect, it is by influencing the structural relationship between the mediator and the outcome. Accordingly, in the thrombosis case, all cross-population differences in effect transmitters influence the indirect effect only by influencing the magnitude of the causal relationship between *blood chemical* and *thrombosis.*

When one considers the way that the indirect effect varies as a result of cross-population differences in effect transmitters, one notices that it is not different from the way that the total effect varies across populations as a result of cross-population differences in background factors in the case where there is only a single path. That is, although cross-population differences in effect transmitters can make a difference in the indirect effect, the way that they do so is not different from the way that run-of-the-mill effect modifiers may make a difference for any effect. Just as variation in background factors influencing an effect variable can make a difference in the

relationship between that effect variable and its direct causes, variation in effect transmitters influencing the outcome can make a difference in the relationship between the outcome and the measured mediator. The fact that effect transmitters are effects of the treatment is irrelevant for considering their influence on the magnitude of the indirect effect.

In identifying the indirect effect within a population, one learns about how the treatment would influence the outcome if there were no direct path. The question we have been considering in this section is whether the way that the indirect effect varies across populations depends on the way that factors contributing to the direct path (i.e. effect transmitters) vary across populations. Since the indirect effect depends on the default values of the effect transmitters, the way that it varies across populations does depend on cross-population difference in what happens along the direct path. Yet, the indirect effect does *not* vary as a result of cross-population differences in the way that effect transmitters *respond* to the treatment. For this reason, we can treat cross-population differences in the values (or distributions) of effect transmitters as being no different than cross-population differences in effect modifiers.

Let's now return to the question with which we began. Assuming that one can use Steel's methods to extrapolate in the single-path case, is it possible to generalize these methods to the case with multiple-paths? The answer is yes. The way that the indirect effect varies across populations is no different from the way that the total effect varies across populations in the single-path case. Thus, to the degree that is possible to use comparative process tracing to extrapolate the total effect, it will also be possible to use it to extrapolate the indirect effect.

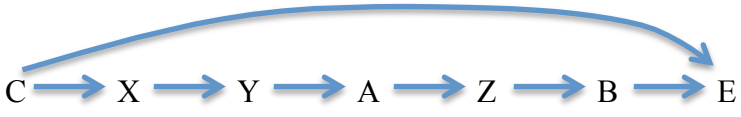

Figure 8

The DAG in figure 8 is almost identical to the DAG that Steel uses to describe comparative process tracing. The one difference is that it contains a direct path from *C* to *E*. To identify the indirect effect in this DAG, one only needs to intervene on one of the mediators between *C* and *E.* For example, the indirect effect $IE_{c,c'}(E)$ is identified by both $E_{cA(c')} - E_{cA(c)}$ and $E_{cZ(c')} - E_{cZ(c)}$. While the indirect effect can vary across population as a result of variation in effect transmitters, all such variation will only influence the structural relationship between *B* and *E.* Depending on which relationships in the models are confounded, one may or may not be able to extrapolate the indirect effect in such a case. Once one identifies the indirect effect, the question of whether it generalizes to a target population is not fundamentally different from the question of whether the effect generalizes in the single-path case that Steel gives. In both cases, one is considering an effect corresponding to a causal chain and trying to improve the reliability of one's extrapolation by comparing the study population and target population at various links in the chain.

6. Non-Parametric Inference and Average Component Effects

Given what I have said in this chapter about how direct and indirect effects depend on the default values of variables along other paths, it is straightforward to show that direct and indirect effects in a population are averages of the direct and indirect effects across subpopulations. In a sense, this result is more relevant to explaining the role of mediation techniques in general non-parametric causal inference than it is specifically to extrapolation. Yet the fact that a particular type of effect in a population is an average over the corresponding effect in subpopulations does have an implication for extrapolation. With regards to the total effect, the fact that the total effect in a population is the average effect across subpopulations entails that there cannot be a causal

effect in a population, but not in any of its subpopulations (Weinberger, 2015). Learning that path-specific effects are average effects over subpopulations would similarly enable one to make similar inferences about direct and indirect effects.

It is not difficult to see that the natural direct effect in a population is an average of the direct effect across subpopulations. For simplicity, let's consider cases in which there is an S-node into the mediator, but not into the outcome variable.[37] We saw above that the only way that an S-node into the mediator influences the direct effect is by influencing the value of M_0 – that is, the value it would take on given the default value of the treatment. The natural direct effect ($DE_{0,1}$) is *defined* as the effect of changing the treatment from X=0 to X=1 while holding M to the value of M_0 for every member of the population. So it is definitional that the direct effect is the average effect across different values of M_0 in the population. If one stratifies the population based on different values of M_0, the natural direct effect will be the average over the direct effects in the subpopulations.

It is not part of the definition of the indirect effect that it is an average effect across the value of some particular variable or potential outcome. Yet, it is like the direct effect in identifying the degree to which one path is invariant to changes in the other path that do not change the default values of variables along that path. As we have seen, the indirect effect is invariant across all changes in effect transmitters that do not change their default values. Moreover, since the *only* way that cross-population differences in effect transmitters changes the indirect effect is by changing the default value, if we were to stratify a populations based on their default values for their effect transmitters, the indirect effect in the whole population would be an average of the indirect effects in each of the resulting subpopulations, weighted by the size of

[37] If there are factors influencing the outcome, it is straightforward to show that the total effect of the treatment on the outcome is the average effect across such factors.

each subpopulation. Of course, since we do not know what the effect transmitters are, we would not be able to stratify the population in this way. But we don't have to. Since one holds the treatment to its default value, the effect transmitters also take on their default values automatically, and the indirect effect is the average over the different combinations values for the effect transmitters. In the same way that, in general, non-parametric causal models allow one to (in principle) measure average causal effects without knowing which background factors one is averaging over, mediation techniques allow one to identify the average indirect effect without knowing which effect transmitters one is averaging over.

The fact that the indirect effect in a population is the average effect of the indirect effect across subpopulations implies that it is not possible for there to be an indirect effect in a population, but not in any of its subpopulations. While this does not entail that the indirect effect in a subpopulation will be similar to that in the population, it does mean that the indirect effect in the population is evidentially relevant to the effect in subpopulations. Whether it is strongly or weakly relevant depends on one's background beliefs about how much the indirect effect varies across subpopulations.

7. A Robustness Test?

In this final section before the conclusion, I speculatively propose one way that the indirect effect might play an important role in an inductive account of extrapolation. As I've noted in several places, one of the problems with developing an inductive account of extrapolation is that it is hard to see how one might think about the magnitudes of a causal relationship in different populations as being sampled from a distribution. If anything, one expects an effect to vary across populations as a result of the local distribution of background factors. Yet, suppose that one tries a policy in several heterogeneous places and finds that it works similarly in each. For

example, the economist Guido Imbens (2010) mentions three studies that found an effect of smaller class size on educational outcomes. One was in Tennessee, one was in Israel, and one was in Connecticut. He argues that these studies intuitively provide evidence for the claim that reducing class sizes will be beneficial in California. The intuition here is clear. While we do not know about the various background factors that influence the magnitude of the effect in the distinct populations, the fact that the policy had similar outcomes in three very different locations suggests that it is robust across changes.[38]

Since I do not know how to provide an inductive account of extrapolation, I am unsure of how to think about robustness tests in a rigorous manner. Yet, I would like to suppose for a moment that inferences like Imbens' are sometimes legitimate in order to see how mediation techniques might be helpful for making such inferences. There appear to be two ways one could test how robust a causal effect is across changes in background conditions. One can test for variation in the effect across populations that differ in known factors and one can test for variation of the effect across populations that differ in unknown factors. An example of the first type of test would be a test of whether the effect of reduced class size on educational outcomes varies across populations with varying degrees of parental income. An example of the second is Imbens' comparison of the three studies. He does not seek to identify particular factors in virtue of which the effects in the various study populations might potentially differ. The strength of his conclusion is derived from the fact that the effect is similar across populations that differ in a variety of factors, both known and unknown. The advantage of the first type of test is that potentially enables one to gain knowledge about how an effect is sensitive to changes in

[38] Here I ignore difficulties with determining whether the *same* policy was implemented in all three places.

specified effect modifiers. The advantage of the second type of test is that it enables one to test the robustness of the effect to unknown sources of variation.

Here's how mediation techniques are relevant to this discussion. Suppose that one discovers that the total effect of smaller class sizes on educational outcomes is large and that the indirect effect of class size on outcomes via increasing student satisfaction with the class is also large (though maybe not as large as the total effect). This provides a robustness test for the indirect effect. By comparing the total effect to the indirect effect, one is able to evaluate the contribution of the indirect path relative to two sets of values of the effect transmitters. The indirect effect is evaluated relative to the default values of the effect transmitters. The total effect tells one about the behavior of the indirect path when the effect transmitters respond to the treatment. It may seem strange to talk about the total effect as telling one something about the indirect path. Yet the total effect tells one how the indirect path would act *in conjunction* with the activity of the direct one. Since the indirect effect and the total effect tell one about the contribution of the indirect path relative to different sets of values for the effect transmitters, learning that they have similar values tells one that the indirect effect is relatively invariant across changes in the values of effect transmitters.

8. Conclusion

In this chapter, I argued for several important claims about the role of causal mediation techniques in extrapolation. First, I explained why the cross-population invariance conditions of the direct and indirect effects cannot be expressed within the transportability framework. S-nodes are not sufficiently fine-grained to distinguish between cross-population sources of variation that do and do not change the default values of variables. Second, I argued that the way that the indirect effect varies across populations in response to differences in effect transmitters

resembles the way that total effects vary in response to differences in background conditions. This enables one to extend Steel's approach for extrapolating single-paths to cases in which there are multiple paths. Third, I showed that the direct and indirect effects in a population are the average of the direct and indirect effect in subpopulations. Finally, I suggested that by comparing the indirect effect and the total effect, one could determine how robust the contribution of the indirect path is to variation in effect transmitters.

Bibliography

Bareinboim, E., & Pearl, J. (2012). *Transportability of causal effects: Completeness results*. In M. Welling, Z. Ghahramani, C. Cortes, and N. Lawrence (eds.), *Advances of Neural Information Processing 27 (NIPS Proceedings)*, 280-288, 2014.

Bareinboim, E., & Pearl, J. (2013). A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference*, *1*(1), 107-134.

Bechtel, W., & Richardson, R. C. 1993. *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT Press.

Blanchard, T., & Schaffer, J. (2013). Cause without default. *Making a Difference*.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy: a practical guide to doing it better*. OUP USA.

Craver, C. and Bechtel,W. 2007. ‘‘Top-down Causation without Top-down Causes,’’ Biology and Philosophy.

Craver, C. 2002. ‘‘Interlevel Experiments and Multilevel Mechanisms in the Neuroscience of Memory,’’ Philosophy of Science, 69: S83 –97.

Craver, C. F. 2007. *Explaining the brain*. Oxford: Oxford University Press.

Dash, D., & Druzdzel, M. (2001). Caveats for causal reasoning with equilibrium models. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty* (pp. 192-203). Springer Berlin Heidelberg.

Eberhardt, F., & Scheines, R. 2007. “Interventions and causal inference”. *Philosophy of Science*, 74(5), 981-995.

Fagan, Melinda Bonnie. 2012. "The joint account of mechanistic explanation." *Philosophy of Science* 79.4: 448-472.

Fisher, R. A. (1935). *The design of experiments*.

Franklin-Hall, L. R. unpublished. "The Emperor’s New Mechanisms." Available at: https://files.nyu.edu/lrf217/public/scientific-explanation/franklin-hall----the-empero.pdf (accessed October 22, 2014)

Forster, M. R. (2007). A Philosopher’s Guide to Empirical Success. *Philosophy of Science*, *74*(5), 588-600.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*(1),

1-35.

Gebharter, Alexander. 2014. "A formal framework for representing mechanisms." *Philosophy of Science*. Vol. 81 no. 1

Glennan, S. 2010. Mechanisms. In *The Oxford Handbook of Causation.* Oxford: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, *116*(4), 661.

Hausman, D. M. (1998). *Causal asymmetries*. Cambridge University Press.

Hausman, D. M. (2005). Causal relata: Tokens, types, or variables?. *Erkenntnis*, *63*(1), 33-54.

Hausman, D. M., Stern, R., & Weinberger, N. (2014). Systems without a graphical causal representation. *Synthese*, *191*(8), 1925-1930.

Hesse, M. (1970). Theories and the Transitivity of Confirmation. *Philosophy of Science*, 50-63.

Hesslow, G. (1976). Two notes on the probabilistic approach to causality. *Philosophy of science*, 290-292.

Hitchcock, Christopher 2001a: "The Intransitivity of Causation Revealed in Equations and Graphs," Journal of Philosophy 98: 273 - 299.

Hitchcock, C. 2001b. A tale of two effects. *Philosophical Review*, 361-396.

Hitchcock, Christopher (1993). "A Generalized Theory of Causal Relevance." Synthese 97: 335–64.

Howick, J., Glasziou, P., & Aronson, J. K. (2013). Problems with using mechanisms to solve the problem of extrapolation. *Theoretical medicine and bioethics*, *34*(4), 275-291.

Imai, K., Keele, L., Tingley, D., & Yamamoto, T. (2011). Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review*, *105*(4), 765-789.

Imai, Kosuke, Dustin Tingley, and Teppei Yamamoto. (2013). "Experimental Designs for Identifying Causal Mechanisms." (with discussions) *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 176, No. 1 (January), pp. 5-51.

Imbens, G. W. (2009). *Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)* (No. w14896). National Bureau of Economic Research.

Iwasaki, Y., & Simon, H. A. (1994). Causality and model abstraction. *Artificial Intelligence*,

67(1), 143-194.

Kim, J. 2000. *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT press.

Korb, K. B., Hope, L. R., Nicholson, A. E., & Axnick, K. (2004). "Varieties of causal intervention". In PRICAI 2004: Trends in Artificial Intelligence (pp. 322-331). Springer Berlin Heidelberg.

LaFollette, Hugh, and Niall Shanks (1995). "Two Models of Models in Biomedical Research." Philosophical Quarterly 45: 141–60.

Leuridan, B. (2012). "Three problems for the mutual manipulability account of constitutive relevance in mechanisms". *The British Journal for the Philosophy of Science*, *63*(2), 399-427.

Machamer, P., Darden, L., & Craver, C. F. 2000. "Thinking about mechanisms". *Philosophy of science*, 1-25.

Mackie, J. L. (1980). *The cement of the universe*. Oxford: Clarendon Press.

Marcellesi, A. (forthcoming). "External Validity: Is There Still a Problem?", *Philosophy of Science*.

Menzies, P. (forthcoming) "The Causal Structure of Mechanisms", forthcoming in a special issue of *History & Philosophy of Science*.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference*. Cambridge University Press.

Pearl, J. 2000. *Causality: models, reasoning and inference* (Vol. 29). Cambridge: MIT press.

Pearl, J. 2001. Direct and indirect effects. In Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, CA, 411–420.

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Pearl, J. 2012a. The causal mediation formula: A guide to the assessment of pathways and mechanisms. Prevention Science 13 426-436, DOI: 10.1007/s11121-011-0270-1.

Pearl, J. 2012b. "The mediation formula: A guide to the assessment of causal pathways in nonlinear models". *Causality: Statistical Perspectives and Applications*: 151-179.

Pearl, J. Forthcoming. "Detecting Latent Heterogeneity". *Sociological Methods and Research*.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, *29*(4), 579-595.

Reichenbach, Hans. 1956. The Direction of Time. Berkeley, CA: University of California Press.

Reiss, J. (2013) *The Philosophy of Economics*. New York: Routledge

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Spirtes, P., Glymour, C. N., & Scheines, R. 2000. *Causation, prediction, and search* (Vol. 81). MIT press.

Steel, Daniel. 2008. *Across the Boundaries: Extrapolation in Biology and Social Science*. New York: Oxford University Press.

Steel, D. (2013). Mechanisms and extrapolation in the abortion-crime controversy. In *Mechanism and causality in biology and economics* (pp. 185-206). Springer Netherlands.

VanderWeele, T. J., & Robins, J. M. (2007). Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology*, *18*(5), 561-568.

Weinberger, N. (2015). If intelligence is a cause, it is a within-subjects cause. *Theory & Psychology*, 0959354315569832.

Whewell, W. (1840). *The philosophy of the inductive sciences: founded upon their history* (Vol. 1).

Wimsatt, W. C. (1998). Simple systems and phylogenetic diversity. Philosophy of Science, 267-275.

Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press

Worrall, J. (2007). Evidence in medicine and evidence‐based medicine. *Philosophy Compass*, *2*(6), 981-1022.

Xie, Y., Brand, J. E. and Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. Sociological Methodology **42** 314-347