Extrapolating Policy Effects - Hopes, Assumptions, and the Extrapolator's Bind

KHOSROWI, DONAL,DJEN,GHESCHLAGHI

Durham University
Department of Philosophy

Doctoral thesis submitted in fulfilment of the requirements for the degree 'Doctor of Philosophy' in Philosophy

Extrapolating Policy Effects -
Hopes, Assumptions, and the Extrapolator's Bind

submitted to:

Durham University

'Student Services Desk
Palatine Centre
Stockton Road
DH1 3LE, Durham

submitted by:

Donal Khosrowi

Department of Philosophy
Student No.: xxxxxxxxx
Date of submission: 14.08.2019

Address:
xxxxxxxxxxx
xxxxxxxxxxx

Contact:
xxxxxxxxx
d.khosrowi@gmail.com

ABSTRACT

Evidence-Based Policy is the movement according to which policy should be based on high-quality evidence for 'what works'. A major problem in using evidence to inform policy is concerned with *extrapolation*, i.e. using evidence of policy effectiveness from a study population to learn something about the effects of a policy in a novel target population. This thesis provides a critical discussion of extrapolation in Evidence-Based Policy and aims to make general contributions to improving both the theory and practice of extrapolation. It proceeds in three parts. *Part I* provides a comprehensive analysis of extrapolation, including what different kinds of extrapolation there are, what makes some of them highly challenging, and what *successful* extrapolation is. *Part II* critically examines existing strategies for extrapolation proposed by philosophers, econometricians, and computer scientists. Emphasis is put on the empirical assumptions about similarities and differences between populations that these strategies involve, and it is argued that supporting these assumptions is often over-demanding. In particular, the knowledge about the target population required to underwrite an extrapolation is often so extensive that we can learn the effect of interest in the target based on this knowledge alone. This is problematic, as it can render the evidence from which one extrapolates irrelevant to an envisioned conclusion, thus undermining the success of an extrapolation. Detailed investigations are provided to highlight the conditions under which existing strategies fall prey to this problem. Building on this critical investigation, *Part III* makes several positive proposals for how to improve the theory and practice of extrapolation in EBP and evade the central problems that it faces.

CONTENTS

STATEMENT OF COPYRIGHT

DECLARATION

This thesis is the result of my own work and has not been submitted for consideration in any other examination. Material from the published or unpublished work of others, which is referred to in the dissertation, is credited to the author(s) in questions in the text.

ACKNOWLEDGEMENTS

I would like to thank my PhD supervisors Wendy Parker, Nancy Cartwright, and Julian Reiss for their fantastic support in helping me write this thesis, as well as my former BA and MA supervisor, Roberto Fumagalli, for supporting my efforts to pursue a PhD in philosophy.

I would also like to thank Finola Finn for supporting me throughout the final year of my PhD, proposing central terminology, and proofreading the manuscript.

Further, I would like to thank, in no particular order, Erin Nash, Rune Nyrup, Sarah Wieten, Anthony Fernandez, Katherine Furman, Anna de Bruyckere, William Peden, Tamlyn Connell, Marzia Beltrami, Emanuele Belli, Giacomo Giannini, Lukas Beck, Tobi Grohmann, Richard Endörfer, Anna Alexandrova, Magdalena Małecka, Çağlar Dede, Joe Roussos, David Teira, Menno Rol, Derek Beach, Alison Wiley, Peter Vickers, Joan Drake, Lizzie Rowe, Nicola Craigs, Sarah Hyland, James Garvey, the members of the CHESS and K4U research groups at Durham University, my students at the University of Bayreuth, and Claudia (a chonky cat).

Finally, I would like to express my gratitude for the extensive financial support that I have received in the form of an AHRC Northern Bridge Doctoral Studentship (grant number: AH/L503927/1), a Durham Doctoral Studentship, and a Royal Institute of Philosophy Jacobsen Studentship.

[THIS PAGE INTENTIONALLY LEFT BLANK]

CHAPTER 1

Extrapolation in Evidence-Based Policy

1.1 Introduction

Evidence-Based Policy (EBP) is the movement according to which policy should be based on high-quality evidence for 'what works', i.e. evidence that persuasively speaks for the effectiveness of policy interventions in realizing[1] their intended goals. Seeking to mirror the success of Evidence-Based Medicine (e.g. Sackett et al. 1996), this movement has gained significant traction over recent decades, with numerous evidence-based initiatives and 'What Works Centres' being created in Australia, the United States, the UK, and other countries, to target a diverse range of policy issues, including education, economic growth, crime reduction, development, and others (see e.g. Cabinet Office 2013).

At the most general level, the motivation behind EBP seems uncontroversial. Consider the following statement by the UK Cabinet Office: "It is a fundamental principle of good public services that decisions are made on the basis of strong evidence and what we know works" (2013, i). So rather than grounding policy decisions in intuitions, hope, or speculation, the idea is to have high-quality empirical evidence directing our attempts to achieve given social and political ends.

When spelling out the details of this proposal, however, various highly contentious issues arise, including: what counts as evidence in the first place? What is 'high-quality' evidence? How do we best obtain it? How should this evidence be used in designing and implementing policies?

In recent years, EBP has been criticized on various fronts surrounding these issues. Several authors have raised methodological concerns about randomized controlled trials (RCTs), which EBP proponents consider the 'gold standard' for producing high-quality evidence (Heckman 1992; Worrall 2007; Cartwright 2007; 2010; Scriven 2008). Others worry about whether evidence can indeed play the role in political decision-making as envisioned by EBP proponents (Barnes and Parkhurst 2014; Parkhurst and Abeysinghe 2016; Cairney 2016; 2018) and whether evidence production and use might involve

[1] I will use Oxford spelling throughout this thesis.

unspoken value presuppositions (Khosrowi 2019; Reiss and Khosrowi 2019, ms.). Moreover, a particularly pressing criticism concerns the issue of *extrapolation*, i.e. using evidence concerning the effectiveness of a policy from a *study population* A to learn something about the effectiveness of that policy in a novel *target population* B. Here, Nancy Cartwright has levelled a sustained line of criticism arguing that evidence of policy effectiveness, by itself, is of limited use for decision-making, since it can only speak for the effectiveness of policies where they have been tested, i.e. in study populations (Cartwright 2009a; 2009b; 2011; 2012; Cartwright and Stegenga 2011; Cartwright and Hardie 2012; see also Angrist and Pischke 2010). Despite these criticisms, when it comes to making inferences about the effectiveness of policies in novel targets, EBP methodological guidelines have continued to leave unclear how such inferences can proceed successfully (see also Rodrik 2008, 20; Heckman and Vytlacil 2007, 4801).

This is a substantive shortcoming, as one of the central hopes in EBP is that it is useful to build 'libraries of evidence'[2] (also called 'intervention libraries', 'clearinghouses', 'warehouses', or 'toolkits'), where high-quality evidence pertaining to the effectiveness of specific (kinds of) policy interventions is collated in evidence-syntheses, and decision makers can go 'shopping' for policies that help address their needs (see also Duflo 2004). Building such libraries only seems useful, however, if we have a clear idea of how the evidence collated there can speak to our questions about novel target populations (Cartwright 2013a, 3).

Unfortunately, extrapolating from a given evidence-base is often difficult. It is rarely plausible, for instance, to use *naïve extrapolation*, which means to merely assume that whatever policy is effective in *A* will also be effective in *B* unless there are strong reasons to think otherwise (see Steel 2008 on *simple induction*; Reiss 2019; Fuller 2019 in the context of medicine). Individuals can differ in their psychological characteristics and economic circumstances; populations can differ with respect to social norms and institutions; the interventions themselves can differ depending on how they are implemented and who implements them, etc. More generally, the concern is that there are various ways in which any two populations can and will differ in causally relevant features that have bearing on the quality and magnitude of policy effects (see e.g. Vivalt

[2] Many of the UK 'What Works Centres' aim to offer such libraries, modelled after similar advances made by the National Institute for Health and Care Excellence (NICE).

2019). In light of likely differences between populations, naïve extrapolation would hence amount to little more than *hoping* that no such differences obtain. This seems undesirable at least insofar as it would undermine the emphasis on using high-quality evidence for establishing policy effects. High-quality evidence seems of little use if one does not know how to draw high-quality *inferences*, too.

We might wonder, then, how we can do better than naïve extrapolation. Here are two intuitions that seem helpful: First, it seems that some form of *similarity* between populations is often important, and sometimes perhaps even necessary, to allow an intervention to be similarly effective in a novel target as in a study population. So to *infer* whether an intervention will be similarly effective in a target as in a study population, we might need to determine whether such similarities obtain. Second, not all differences between populations present insurmountable obstacles to extrapolation; they only preclude *naïve* extrapolation. Even if populations differ importantly, we might still be able to anticipate *how* such differences bear on the effects to be extrapolated, and hence we might still be able to successfully predict policy effects in a target. Both of these ideas seem largely uncontroversial. Nevertheless, which similarities need to obtain and to what degrees, which differences can be accommodated and adjusted for, as well as how to acquire and use information pertaining to such similarities and differences, is far from obvious.

Given the challenging nature of the problem, it is perhaps no surprise that there are various proposals for *strategies* to help address it (Hotz et al. 2005; Crump et al. 2008; Steel 2008; 2010; Guala 2010; Cartwright 2012; 2013a; 2013b; Cartwright and Hardie 2012; Bareinboim and Pearl 2012; 2016; Muller 2013; 2014; 2015; van Eersel et al. 2019). These strategies try to achieve different things: some characterize sufficient conditions for when an effect will be the same in a target as in a study population (Cartwright 2013a; Cartwright and Hardie 2012); others propose methods for learning about similarities and differences between populations in order to decide whether effects will be qualitatively the same in the target as in a study (Steel 2008); and more ambitious strategies propose ways to adjust for relevant differences between populations, with the aim to permit prediction of quantitative effects in a target despite such differences (Hotz et al. 2005; Muller 2014; 2015; Bareinboim and Pearl 2012; 2016). Against the background of these proposals, some authors have suggested that the

problem of extrapolation has been "solved" (Marcellesi 2015; Bareinboim and Pearl 2016).

In this thesis, I will argue that this conclusion is too hasty and that none of the strategies on offer provides a compelling, general recipe for how we can extrapolate policy effects to novel settings. To this end, I will critically engage with these strategies, emphasising their advantages, challenging them on distinct disadvantages, and arguing that they exhibit a common shortcoming. In different ways, all existing strategies for extrapolation involve substantive assumptions about causally relevant similarities between populations. This is not a shortcoming per se, since some such assumptions are needed to permit any extrapolative inference more sophisticated than naïve extrapolation. However, supporting such assumptions, and thereby justifying the inferences enabled by them, often requires extensive causal knowledge of the target population. This is not only a concern about how demanding it can be to underwrite the assumptions required for extrapolation; it is a more serious problem, as it makes the strategies that involve these assumptions liable to fall prey to what I will call the *extrapolator's bind* (a generalized version of LaFollette and Shanks' 1996 and Steel's 2008 *extrapolator's circle*). In a nutshell, the extrapolator's bind requires that the supplementary information used to justify an extrapolation should not be so extensive that it allows us to learn the effect of interest on this basis alone, thereby rendering the experimental evidence to be extrapolated from redundant to our conclusion. This, I take it, is a substantial problem not only for strategies for extrapolation, but for EBP more generally, as it would undermine the promise that we can successfully learn about the effectiveness of policies in novel targets based on evidence collated in evidence libraries.[3]

Importantly, arguing that there are epistemic problems involved in underwriting the assumptions that strategies for extrapolation require is not intended to suggest that these strategies are somehow fundamentally inadequate. For the most part, I will grant that they successfully clarify the *abstract* conditions under which extrapolation can be successful in principle. However, they provide no *concrete* guidance concerning how one could overcome the substantial and non-trivial epistemic challenges of underwriting the assumptions that they involve, specifically with a view towards evading the

[3] See, however, Dekkers et al. (2010) and Banerjee and Duflo (2009) who take the extreme view that the only way to credibly establish that an effect will be the same in the target as in an experimental population is to repeat the study in the target.

extrapolator's bind. While getting the abstract conditions right is, of course, an important achievement, it is, by itself, insufficient for overcoming any concrete problems of extrapolation.

Of course, it has long been recognized that compelling inference schemas are only part of what helps us reach extrapolative conclusions. For instance, one of Cartwright's main aims, beyond making proposals for how to facilitate extrapolative inference at an abstract level, has been to raise awareness of just how difficult it can be to adequately support the assumptions required for such inference. More recently, Cartwright (forthcoming), and reinforcing proposals made in the realist evaluation literature (Pawson and Tilley 1997; 2001; Astbury and Leeuw 2010; Pawson 2013), has called for more efforts to develop and use theoretical resources that can be helpful for underwriting extrapolative inferences. Specifically, she emphasises the importance of so-called *middle-range theories,* which can be useful for clarifying whether the outcomes of interest in two populations are likely to be governed by similar causal mechanisms, as well as *programme theories,* which clarify *how* the policy interventions of interest are supposed to work and what contextual features are important for their effectiveness.

Yet, while such theoretical resources can be helpful for addressing some of the concerns to be developed here, it is also important to recognize that such resources are not always available, and that their usefulness is often (too) specific to a context.

Where abstract strategies remain too general, pointing to the importance of contextual resources relegates important epistemological issues to being settled by the concrete contextual details of specific cases. In doing so, appeals to theory neglect a host of finer-grained, but still relatively general, type-level features of problems of extrapolation and extrapolative inference that can bear importantly on the success of extrapolation. What is more, appeals to theory also do not say much about what role *empirical* strategies can play in underwriting extrapolation, even though these strategies can be highly useful in the absence of theory, in conjunction with it, and for building such theory in the first place.

What this suggests is that there is both a need as well as a place for a *framework for extrapolation* that operates at an intermediate level of analysis, i.e. between abstract strategies, which leave important epistemic challenges unaddressed, and concrete contextual resources, which, if and when available, can be helpful for facilitating

extrapolation, but, on their own, cannot provide general insights concerning how to tackle different kinds of extrapolation. As the arguments to be developed in this thesis will suggest, it is possible to build the basis for such a framework for extrapolation by saying more about important differences in the kinds of problems of extrapolation we might encounter; whether and how different strategies for extrapolation are able to address such problems, in principle and in practice; what kinds of assumptions they will (need to) make across different cases; how these assumptions might be underwritten by additional resources; and when doing so is unlikely to be successful. While, like existing strategies, the proposals to be developed here will not provide a general solution for problems of extrapolation, it is hoped that they can nevertheless help fill important blanks left open by existing proposals, and make significant contributions to connecting abstract strategies for extrapolation with its epistemic practice.

With these general aims in place, let me briefly outline the structure and contents of the subsequent chapters.

1.2 Outline

Chapter 2 provides a comprehensive discussion of problems of extrapolation and extrapolative inference. Here, I argue that extrapolation is a highly heterogeneous collection of problems and inferential activities. To this end, I distinguish between several dimensions along which problems of extrapolation can differ. This is followed by an overview of important differences in the epistemic aims pursued by extrapolative inference. Together, these analyses suggest that not only is extrapolation highly heterogeneous, but also that some problems of extrapolation are significantly more challenging than others, thus providing a useful background for evaluating existing extrapolation strategies with respect to what kinds of problems they can address, and what kinds of conclusions they can enable.

Chapter 3 takes a closer look at some of the basic assumptions that different strategies for extrapolation make. Here, I argue that they rest on a common core of assumptions that, taken together, ensure that it is in principle possible to predict the effects of an intervention in a novel target, despite important differences between populations. In addition, I provide a working analysis of what extrapolation is, at the most general level, and complement this analysis with important strictures on what counts as *successful* extrapolation. Here, building on Steel's (2008) *extrapolator's*

circle, I characterize a generalized version of this challenge, called the *extrapolator's bind*. I argue that evading the extrapolator's bind is an important part of what it means to achieve *successful* extrapolation. Together, these ingredients build the background for my subsequent efforts to evaluate existing strategies for extrapolation with respect to the assumptions they involve and whether supporting these assumptions can proceed in a way that enables successful extrapolation.

Chapter 4 is the first to take issue with concrete proposals for how to address problems of extrapolation. Here, I consider Cartwright's *argument-based* strategy for extrapolation. Cartwright (2013a) proposes the *Argument Theory of Evidence*, according to which extrapolative conclusions should be arrived at by means of valid and sound *arguments*, where effectiveness evidence comes together with further assumptions about how study and target populations are related in order to licence an extrapolative conclusion. To illustrate the capabilities of the *Argument Theory*, Cartwright offers an exemplary *effectiveness argument*, which involves assumptions that licence conclusions about whether an intervention can be efficacious for at least some individuals in a target and whether causal effects are the same in experimental and target populations. I argue that supporting these assumptions is epistemically over-demanding and raises important concerns about the extrapolator's bind. Following this, and looking beyond the *effectiveness argument*, I provide an updated assessment of how the contributions of Cartwright's more general *Argument Theory* could be understood.

Chapter 5 considers Steel's (2008) *mechanism-based* strategy for extrapolation. Steel's approach is the only existing strategy to explicitly acknowledge the extrapolator's bind (in the restricted form of the *extrapolator's circle*), and is specifically designed to evade it. Despite this, I argue that Steel's strategy encounters problems in evading the extrapolator's bind in a wide range of contexts. Specifically, I offer a distinction between two kinds of extrapolation, *attributive* and *predictive* (the latter being typical in EBP), and argue that Steel's strategy is unlikely to successfully overcome problems of predictive extrapolation and is hence unsuitable, by itself, to figure as a compelling strategy for extrapolation in EBP.

Chapter 6 focuses on *interactive covariate-based* strategies for extrapolation developed by econometricians (Hotz et al. 2005; Muller 2013, ms.; 2014; 2015). These strategies aim to accommodate and adjust for differences between populations in certain respects. However, in doing so, they involve substantive assumptions that populations

do not differ in other, more basic respects. I argue that supporting these assumptions raises important concerns about the extrapolator's bind. I then consider whether Steel's strategy could figure as a useful complement to evade this problem. I argue that it is unlikely to do so, however, at least in *predictive* extrapolation, and as long as one only uses quantitative observational evidence from a target to underwrite extrapolation. To remedy this, I suggest that other, previously neglected kinds of evidence may need to be considered by econometricians, including *qualitative* evidence.

Chapter 7 considers Bareinboim and Pearl's (2012; 2016) *graph-based* strategy. They offer a causal graph-based framework and an accompanying formal calculus to decide whether causal effects can be extrapolated at all and, if so, to help derive formulae to compute the effect of interest in the target. I argue that this strategy involves wide-ranging assumptions about the similarity of populations and that supporting these assumptions raises concerns about the extrapolator's bind. These problems are, again, aggravated by the distinction between *attributive* and *predictive* extrapolation, with the latter posing special problems that make graph-based strategies unsuitable for a wide range of extrapolations in EBP.

Integrating the insights developed in these chapters, *Chapter 8* works towards building the basis for a more general *framework for extrapolation*. Here, I discuss how background knowledge, theory, and empirical resources can productively interact to underwrite extrapolation. Building on this, I propose a list of substantive, general desiderata for helping future extrapolation strategies evade some of the challenges characterized in this thesis. Finally, I make some recommendations for how a general framework for extrapolation might be usefully complemented by future research on the role of *uncertainty* in extrapolation, and offer some suggestions for how EBP institutions might incorporate some of the insights provided in this thesis into their methodological recommendations.

Chapter 9 provides a concluding summary of the main contributions made in this thesis.

With this overview of the structure of the project in place, let me briefly give some general commentary concerning the scope of the arguments to be developed, outline some general assumptions I will make, and explain some issues that are related to those that I will consider here, but that I will not touch upon in more detail.

1.3 Scope, Assumptions, and Related Issues

First, as I will focus on issues of extrapolation primarily in the context of EBP, it is important to say more on what I take EBP to be. EBP can be understood in a narrower and wider sense. A narrow conception would say that EBP is concerned primarily with establishing policy effectiveness by means of RCTs (and meta-analyses thereof). This would follow traditional evidence-hierarchies developed in Evidence-Based Medicine and later adopted by EBP, where RCTs are the 'gold standard' method by which estimation of intervention effects should proceed (see e.g. Coe 2004; Goldacre 2013; Sanders and Halpern 2014). A wider conception, by contrast, would also allow other kinds of evidence, or really any evidence (i.e. understood in a thin sense as anything that can raise the probability of a hypothesis), to figure as a legitimate means for informing policy.

The partition of EBP that I will focus on sits somewhere in the middle between these conceptions. It is closely tied to cases where standard methods such as those recommended in EBP methodological guidelines and evidence-hierarchies are used, i.e. RCTs and meta-analyses. However, I will also depart from the narrow conception in that, there, EBP is often understood to focus mostly on issues of evidence-based policymaking in the developed world. My arguments will also extend, however, to the substantive empirical literature on (economic) development (see e.g. Duflo 2001; Miguel and Kremer 2004; Banerjee et al. 2007; Banerjee and Duflo 2011) and empirical microeconomics more generally (Angrist and Pischke 2010), sometimes called the *treatment effects literature* (Heckman 2005), where there is a similar emphasis on the importance of using experimental and quasi-experimental methods to investigate the effects of interventions (though not necessarily *policy* interventions; see Karlan et al. 2009; Banerjee 2007; Glennerster and Kremer 2011).

Here, following methodological emphasis on the importance of study design by Campbell and Stanley (1963), Cook and Campbell (1979), Leamer (1983), and others, it is claimed that microeconometrics has gone through a *credibility revolution* (Angrist and Pischke 2010), enabled by more rigorous identification strategies for causal effects and structural parameters that improve importantly on standard multivariate regression analyses. These strategies include not only RCTs but also instrumental variables approaches (Angrist 1990; Angrist and Krueger 1991), matching approaches (see Imbens and Wooldridge 2009), regression-discontinuity approaches (Angrist and Lavy

1999), and differences-in-differences approaches (Donohue and Wolfers 2005). Although these methods are frequently ranked lower than RCTs in evidence hierarchies, my subsequent arguments will also extend to cases where these methods are used to establish the effects of policies, or interventions more generally. Importantly, irrespective of the method used to estimate policy effects, I will simply assume that this has been successful in the sense that the identification assumptions of this method were satisfied and that the estimated effect is an unbiased estimate of the true causal effect in question (for various concerns about identification assumptions of RCTs see Heckman 1992; Worrall 2002; 2007; Cartwright 2010; Fuller 2018).

Concerning scope, it is important to note that the arguments developed here could also be taken to cover extrapolation in Evidence-Based Medicine, as well as in other fields, such as psychology, epidemiology, pharmacology, etc. In this sense, the treatment of EBP here could be understood as an extended case study for building arguments pertaining to extrapolation of causal effects more generally. While remaining open to this possibility, I will not make any commitments concerning the applicability of my arguments beyond EBP, recognizing that problems of extrapolation, aims, available empirical and theoretical resources, etc., might significantly differ in other areas, which could bear importantly on the cogency of the arguments to be developed here.

Second, I will not consider *all* strategies for extrapolation here, but only a selection that I take to capture the most important *types* of strategies. For instance, Robert Northcott has recently taken issue with problems of extrapolation. In the spirit of earlier work on the importance of predictive performance (e.g. Northcott 2017; 2019), he argues that prediction markets, in virtue of their incentive structures and ability to integrate information, could be helpful for 'outsourcing' extrapolative inference to market agents and thereby evading Steel's extrapolator's circle. Although providing a potentially elegant strategy for evading the extrapolator's circle, I will not consider this proposal in more detail here, as my arguments focus on how extrapolative inference proceeds and how particular inferences are justified, which is abstracted away from when considering how market arrangements can facilitate accurate prediction. Similarly, I will also not engage in detail with recent work by Beach and Pedersen (2019) who, in the context of qualitative comparative analysis, propose a 'snowballing-outward' strategy to help researchers assess the generalizability of causal claims in

political science. While I consider their proposals to be interesting, they seem more suitable for underwriting extrapolative inferences enabled by other strategies, rather than for characterizing how to make such inferences. I will also not engage with Guala's (2010) *analogical inference* reconstruction of extrapolation, which focuses primarily on highly controlled social science experiments, nor with van Eersel et al.'s (2019) proposals to use *latent-class regression* methods, which are similar to econometricians' proposals discussed in *Chapter 6*. Finally, I will not discuss Steel's (2010) more recent attempt to integrate causal graph-based methods and comparative process tracing, as the arguments in *Chapters 5* and *7* suggest that this approach will be vulnerable to similar problems as those discussed here.

Third, as RCTs will often play a central role in the kinds of extrapolation scenarios I am interested in, it seems useful to briefly reiterate the motivation behind using them, how they work, and what their (purported) virtues are (see Deaton and Cartwright 2017 for an excellent overview). Learning whether a policy 'works' often means learning what its *causal effects* are. This is difficult, particularly in policy, as individuals' outcomes, such as their welfare, educational achievements, health, wellbeing, etc., are often determined by a whole battery of influences beyond a policy of interest, often called *confounding factors*. So in learning 'what works', and in using that information for guiding subsequent policy action, it is important to isolate the policy effects we are interested in from the background noise induced by these confounding factors. This is a concern about *bias*. A *biased* measure of a policy effect is one that includes not only the effects of our policy, but also those induced by confounding factors. Obtaining *un*biased estimates of policy effects is believed to be crucial for informing policy action, as we might otherwise misallocate resources based on mistaken conclusions about policy effectiveness. To help address such concerns, the typical aim in EBP is to use methods that allow us to construct credible *counterfactuals*, i.e. measurements of states that tell us something about the differences in an outcome of interest in the presence and absence of the policy of interest respectively, other things being equal.

This way of motivating the use of RCTs is common in EBP (see e.g. Duflo et al. 2006; Zigler and Dominici 2014; Rossi et al. 2019, ch. 6) and largely follows the *potential outcomes framework* of Rubin (1974) and Holland (1986). Here, given an outcome of interest Y, the effect of an intervention for individual i, called the *individual treatment effect* (ITE), is the difference between her potential outcome $Y_i(1)$ given the

treatment and her potential outcome $Y_i(0)$ in the absence of treatment, other things being equal. While it would be desirable to measure ITEs for each individual in order to determine the effects of a policy, it is typically assumed that only one of the two values of Y_i can be observed for the same individual, so ITEs are considered unobservable magnitudes (Holland 1986). RCTs offer a remedy for this by permitting the estimation of *average treatment effects* (ATEs) instead of ITEs. This is achieved through 'balancing' the net effects of confounding factors by means of random assignment of subjects to experimental and control groups, and (if applicable) multiple blinding of trial participants, those administering treatment, and those recording and interpreting outcomes. Provided that randomization (and other precautions) are successful in that the net effects of confounders (including interactions among them, see e.g. Fuller 2018) are approximately balanced between treatment and control groups, and some further conditions pertaining to attrition are satisfied, an ideal RCT can help obtain, in expectation, an unbiased estimate of the ATE, defined as the expectation of the difference between the outcomes of treated and untreated units, indicated by $Y_t(1)$ and $Y_c(0)$ respectively:

$$ATE = E[Y_t(1) - Y_c(0)].$$

RCTs typically rank highest in evidence-hierarchies because their identification assumptions are believed to be easy to meet and they require few substantive assumptions to begin with, both of which are helpful in keeping concerns about bias at bay. I will not engage in a more extensive discussion of the role of randomization in addressing such concerns. These can be found elsewhere (Bloom 2006; Duflo et al. 2006; White 2013). Relatedly, I will also not engage in more detail with evidence-hierarchies put forward in EBP beyond highlighting in some places (e.g. *Chapter 8*) that my arguments suggest that such hierarchies might need to be revised when considering the suitability of the *supplementary* evidence used to underwrite an extrapolation, with a view towards how different kinds of evidence can sometimes productively interact in doing so (cf. Clarke et al. 2014; see also Borgerson 2009 in the context of evidence-based medicine).

Fourth, discussing issues surrounding extrapolation will invariably involve reference to concepts such as 'cause', 'causal effect', 'causal relationship', 'causal mechanism', 'intervention', and their cognates. There is a rich philosophical literature targeting metaphysical and epistemological issues pertaining to each of these concepts (see e.g.

Machamer et al. 2000; Glennan 2002; Woodward 2002; 2003; Bechtel and Abrahamsen 2005; Craver 2007 on mechanisms; Schaffer (2016) for an overview on the metaphysics of causation). I will not engage more deeply with attempts to explicate these concepts, nor will my arguments be strongly tied to any specific conception or account. That being said, since causal mechanisms will figure centrally in the arguments to be developed here, it seems useful to clarify that I will, for the most part, understand 'causal mechanisms' along the lines of Illari and Williamson's minimal characterization according to which "[a] mechanism for a phenomenon consists of entities and activities organized in such a way that they are responsible for the phenomenon" (2012, 120). Moreover, my elaborations of how causal mechanisms figure in producing the causal effects of policy interventions, and how learning about causal mechanisms can proceed, are broadly sympathetic to manipulability/interventionist accounts of causation such as Woodward's (2002; 2003). According to these accounts, and speaking broadly and non-reductively (see Craver 2007; Glennan 2009), *causes* are the kinds of things that yield effects (understood as changes or differences in some variable) when (potentially and discriminately) intervened on, and *mechanisms* are broadly understood as (potentially imperfect representations of) the causal arrangements (comprised by individual causal relationships between variables) that govern the production of these effects, or transmit them from an intervention variable onto an outcome. So no assumptions are made that the kinds of things I refer to by 'causes', 'causal relationships', and 'causal mechanisms' are in some metaphysically important sense fundamental or real. They might be, or they might not, but for the epistemic endeavour of extrapolating causal effects these issues do not seem to be of central importance. I will also not distinguish between causal mechanisms and causal *processes* (e.g. as per Salmon 1984; 1994; Dowe 1992), the latter of which could be understood as requiring some continuous activity among the entities that figure in them. I will merely think of these cases as subspecies of mechanisms, but gloss over any of the more intricate metaphysical and epistemological differences that one might be interested in if attempting to make a distinction. Finally, although I am sympathetic to arguments emphasising that reference to mechanisms can afford explanatory and predictive abilities in social science contexts, I will also not engage in more detail with the literature on the importance of mechanistic knowledge in social science (see e.g. Russo 2009; Hedström and Ylikoski 2010).

Finally, I will not take issue with the debate surrounding *internal* and *external validity*. These concepts, to my mind, have been largely confusing, and confused,

particularly in debates concerning apparent tensions between views asserting that internal validity is a prerequisite for external validity (see Hogarth 2005) and views asserting that there is a trade-off between the two (Campbell 1957; see also Guala 2003). For the present purposes, I will not make any specific commitments concerning the distinction between the two, their mutual relationship, or how they relate to extrapolation.

With these general caveats in mind, let me proceed to the substantive discussion, starting with a general outline of how problems of extrapolation are constituted, and what different types of extrapolation there are.

References

Angrist, J. D. (1990). "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records", *American Economic Review*, 80(3): 313–36.

Angrist, J. D., and A. B. Krueger. (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?", *Quarterly Journal of Economics*, 106(4): 976–1014.

Angrist, J. D., and V. Lavy. (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics*, 114(2): 533–75.

Angrist, J. D., and J.-S. Pischke. (2010), "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics", *Journal of Economic Perspectives*, 24(2): 3-30.

Astbury, B., and F. Leeuw. (2010). "Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation", *American Journal of Evaluation*, 31(3): 363-81.

Banerjee, A. (2007). *Making Aid Work*. Cambridge, MA: MIT Press.

Banerjee, A., and E. Duflo. (2009). "The Experimental Approach to Development Economics", *Annual Review of Economics*, Annual Reviews, vol. 1(1): 151-178.

— **(2011)**. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York (NY): Public Affairs.

Banerjee, A., E. Duflo, E., S. Cole, and L. Linden. (2007). "Remedying Education: Evidence from Two Randomized Experiments in India", *Quarterly Journal of Economics*, 122(3): 1235-64.

Bareinboim, E., and J. Pearl. (2012). "Transportability of causal effects: Completeness results", In: Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12), Menlo Park, CA.

— **(2016)**. "Causal inference and the data-fusion problem". *Proceedings of the National Academy of Sciences*, 113: 7345-52.

Barnes, A. and J. Parkhurst. (2014). "Can global health policy be depoliticised? A critique of global calls for evidence-based policy". In: *Handbook of Global Health Policy*, ed. G. Yamey and G. Brown, 157–173. Chichester: Wiley-Blackwell.

Beach, D, and R. Pedersen. (2019), *Process-Tracing Methods - Foundations and Guidelines*, 2nd edition. Ann Arbor: University of Michigan Press.

Bechtel, W., and A. Abrahamsen. (2005). "Explanation: A Mechanistic Alternative", *Studies in History and Philosophy of the Biological and Biomedical Sciences*, 36: 421–441.

Bloom, H. (2006). "The Core Analytics of Randomized Experiments for Social Research", MDRC Working Papers on Research Methodology. New York, NY: MDRC. Accessed 26 July, 2019 http://www.mdrc.org/sites/default/files/full_533.pdf

Borgerson, K. (2009). "Valuing evidence: bias and the evidence hierarchy of evidence-based medicine", *Perspectives in Biology and Medicine*, 52(2): 218-33.

Cabinet Office. (2013). *What Works: evidence centres for social policy*. London: Cabinet Office.

Cairney, P. (2016). *The Politics of Evidence-Based Policy Making*. London: Palgrave Macmillan.

— **(2018).** "The UK government's imaginative use of evidence to make policy", *British Politics*, 14(1): 1-22.

Campbell, D. (1957). "Factors relevant to the validity of experiments in social settings". *Psychological Bulletin*, 54(4): 297-312.

Campbell, D., and J. Stanley. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Cartwright, N. D. (2007). "Are RCTs the Gold Standard?", *BioSocieties*, 2(2): 11-20.

— **(2009).** "Evidence-Based Policy: What's to Be Done About Relevance", *Philosophical Studies*, 143(1): 127-36.

— **(2010)**. "What are randomised controlled trials good for?", *Philosophical Studies*, 147: 59-70.

— **(2011)**. "Predicting 'It will work for us': (Way) beyond statistics". In: F. R. Phyllis McKay Illari, and Jon Williamson (Ed.), *Causality in the Sciences*. Oxford: Oxford Scholarship Online.

— **(2012)**. "Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps", *Philosophy of Science*, 79: 973-89.

— **(2013a)**. "Evidence, Argument and Prediction". In: *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, The European Philosophy of Science Association Proceedings 2, Geneva: Springer.

— **(2013b)**. "Knowing what we are talking about: why evidence doesn't always travel", *Evidence & Policy*, 9(1): 97-112.

— **(Forthcoming)**. "Lullius Lectures 2018: Mid-level theory: Without it what could anyone do?" In: C. Martínez Vidal and C. Saborido (ed.), *Nancy Cartwright's Philosophy of Science*, Special Issue of *Theoria*.

Cartwright, N. D., A. Goldfinch, and J. Howick. (2009). "Evidence-Based Policy: Where Is Our Theory of Evidence?", *Journal of Children's Services*, 4(4): 6-14.

Cartwright, N. D., and J. Hardie. (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: Oxford University Press.

Cartwright, N. D., and J. Stegenga. (2011). "A Theory of Evidence for Evidence-Based Policy". In P. Dawid, W. Twining., and M. Vasilaki (Eds.), *Evidence, Inference and Enquiry* (Proceedings of the British Academy). New York: Oxford University Press.

Clarke, B., D. Gillies, P. Illari, F. Russo, and J. Williamson. (2014). "Mechanisms and the Evidence Hierarchy", *Topoi*, 33: 339–360.

Coe, R. (2004). "What Kind of Evidence Does Government Need?", Evaluation & Research in Education, 18(1): 1–11.

Cook, T., and D. Campbell. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton-Mifflin.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2008). "Nonparametric tests for treatment effect heterogeneity", *The Review of Economics and Statistics*, 90 (3): 389–405.

Dekkers, O. M., E. von Elm, A. Algra, J. A. Romijn, and J. P. Vandenbroucke. (2010). "How to assess the external validity of therapeutic trials: a conceptual approach", *International Journal of Epidemiology*, 39: 89–94.

Deaton, A., and N. D. Cartwright. (2017). "Understanding and Misunderstanding Randomized Controlled Trials", Working Paper. Cambridge (MA): NBER.

Donohue, J., and J. Wolfers. (2005). "Uses and Abuses of Empirical Evidence in the Death Penalty Debate". *Stanford Law Review*, 58: 791–845.

Dowe, P. (1992). "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory", *Philosophy of Science*, 59: 195–216.

Duflo, E. (2001). "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment", *American Economic Review*, 91: 795-813.

— (2004). "Scaling Up and Evaluation." In: F. Bourguignon and B. Pleskovic (eds.), *Annual World Bank Conference on Development Economics, 2004: Accelerating Development*, pp. 341–69. Washington, D.C.: World Bank; Oxford and New York: Oxford University Press.

Duflo, E., R. Glennerster, and M. Kremer. (2006). "Using Randomization in Development Economics Research: A Toolkit." Cambridge, MA: Department of Economics, Massachusetts Institute of Technology and Abdul Latif Jameel Poverty Action Lab. http://www.povertyactionlab.org/sites/default/files/documents/Using%20Randomization%20in%20Development%20Economics.pdf (accessed July 26, 2019).

Fuller, J. (2018). "The Confounding Question of Confounding Causes in Randomized Trials", *The British Journal for the Philosophy of Science*, axx015.

— (2019). "The myth and fallacy of simple extrapolation in medicine", Synthese, online first, https://doi.org/10.1007/s11229-019-02255-0.

Glennan, S. (2002). "Rethinking Mechanistic Explanation", *Philosophy of Science*, 69: 342–53.

— (2009). "Productivity, Relevance and Natural Selection", *Biology and Philosophy*, 24:325–339.

Glennerster, R., and M. Kremer. (2011). *Small Changes, Big Results: Behavioral Economics at Work in Poor Countries*. Boston Review: A Political and Literary Forum.

Goldacre, B. (2013). "Building Evidence into Education." London: Department for Education, https://media.education.gov.uk/assets/files/pdf/b/ben%20goldacre%20paper.pdf (accessed 10 August 2019).

Guala, F. (2003). "Experimental localism and external validity", *Philosophy of science*, 70(5):1195-1205.

— (2005). "External Validity". In: Guala, F. *The Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

— (2010). "Extrapolation, Analogy, and Comparative Process Tracing". *Philosophy of Science*. 77: 1070–82.

Heckman, J. (1992). "Randomization and Social Policy Evaluation". In: *Evaluating Welfare and Training Programs*, ed. C. F. Manski and I. Garfinkel, 201-230. Boston (MA): Harvard University Press.

— (2005). "The scientific model of causality", *Sociological methodology*, 35: 1–97.

Heckman, J., and E. Vytlacil. (2007). "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation", In: *Handbook of Econometrics,* Vol. 6, Part B, ed. J. Heckman and E. Leamer, pp. 4779-4874. North Holland: Amsterdam.

Hedström, P., and P. Ylikoski. (2010). "Causal Mechanisms in the Social Sciences", *Annual Review of Sociology*, 36: 49–67.

Hogarth, R. M. (2005). "The challenge of representative design in psychology and economics". *Journal of Economic Methodology*, 12(2): 253-263.

Holland, P. (1986). "Statistics and Causal Inference", *Journal of the American Statistical Association*. 81(396): 945-60.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer. (2005). "Predicting the efficacy of future training programs using past experiences at other locations", *Journal of Econometrics*, 125: 241–270.

Illari, P. M., and J. Williamson. (2012). "What is a Mechanism?: Thinking about Mechanisms Across the Sciences", *European Journal for Philosophy of Science*, 2: 119–135.

Imbens, G. W., and J. M. Wooldridge. (2009). "Recent Developments in the Econometrics of Program Development", *Journal of Economic Literature*, 47(1): 5–86.

Karlan, D., N. Goldberg, and J. Copestake. (2009). "Randomized controlled trials are the best way to measure impact of microfinance programs and improve microfinance product designs", *Enterprise Development and Microfinance*, 20(3): 167–76.

Khosrowi, D. (2019). "Trade-offs Between Epistemic and Moral Values in Evidence-Based Policy", *Economics & Philosophy*, 35(1), 49-78.

LaFollette, H., and N. Shanks. (1996). *Brute Science: Dilemmas of Animal Experimentation*, New York: Routledge.

Leamer, E. (1983). "Let's Take the Con Out of Econometrics", *American Economic Review*, 73(1): 31–43.

Machamer, P.K., L. Darden, and C.F. Craver. (2000). "Thinking about Mechanisms", *Philosophy of Science*, 67: 1–25.

Marcellesi, A. (2015). "External Validity: Is There Still a Problem?", *Philosophy of Science*, 82(5): 1308-17.

Miguel, E., and M. Kremer. (2004). "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities", *Econometrica*, 72(1): 159-217.

Muller, S. M. (2013). „External validity, causal interaction and randomised trials: the case of economics", Unpublished manuscript.

— **(2014)**. "Randomised trials for policy: a review of the external validity of treatment effects", Southern Africa Labour and Development Research Unit Working Paper 127, University of Cape Town.

— **(2015)**. "Interaction and external validity: obstacles to the policy relevance of randomized evaluations", *World Bank Economic Review*, 29(1): 217-25.

Northcott, R. (2017). "When are Purely Predictive Models Best?", *Disputatio*, 9(47): 632-56.

— **(2019).** "Prediction versus accommodation in economics", *Journal of Economic Methodology*, 26(1): 59-69.

Parkhurst, J., and S. Abeysinghe. (2016). "What constitutes "good" evidence for public health and social policy-making? From hierarchies to appropriateness", *Social Epistemology,* 30: 665-679.

Pawson, R., and Tilley, N. (1997). *Realistic Evaluation*, London: SAGE Publications.

— **(2001)**. "Realistic Evaluation Bloodlines", *American Journal of Evaluation*, 22: 317-324.

Pawson, R. (2013). *The science of evaluation: a realist manifesto*, London: SAGE Publications.

Reiss, J. (2019). "Against external validity", *Synthese*, 196(8): 3103-21.

Reiss, J., and Khosrowi, D. (2019, ms.). "Evidence-Based Policy and Its Hidden Costs of Justice", Unpublished manuscript, Durham University.

Rodrik, D. (2008). "The New Development Economics: We Shall Experiment, but How Shall We Learn?". HKS Working Paper No. RWP08-055.

Rossi, P. H., H. E. Freeman, and M. W. Lipsey. (2019). *Evaluation - A Systematic Approach*. Thousand Oaks: SAGE.

Rubin, D. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies", *Journal of Education Psychology*, 66: 688–701.

Russo, F. (2009). *Causality and Causal Modeling in the Social Sciences: Measuring Variation*, Dordrecht: Springer.

Russo, F., and J. Williamson. (2007). "Interpreting Causality in the Health Sciences", *International Studies in the Philosophy of Science*, 21: 157–170.

Sackett D., W. Rosenberg, M. Gray, B. Haynes, and S. Richardson. (1996). "Evidence based medicine: what it is and what it isn't", *BMJ*, 312 :71.

Salmon, W.C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

— (1994). “Causality Without Counterfactuals”, *Philosophy of Science*, 61: 297–312.

Sanders, M., and D. Halpern. (2014). “Nudge Unit: Our Quiet Revolution Is Putting Evidence at Heart of Government”, *The Guardian*, 3 February 2014.

Schaffer, J. (2016). "The Metaphysics of Causation", The Stanford Encyclopedia of Philosophy (Fall 2016 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics.

Scriven, M. (2008). "A Summative Evaluation of RCT Methodology & An Alternative Approach to Causal Research", *Journal of Multi-Disciplinary Evaluation,* 5(9): 11-24.

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*, Oxford University Press.

— (2010). "A New Approach to Argument by Analogy: Extrapolation and Chain Graphs", *Philosophy of Science*, 77(5): 1058-69.

van Eersel, G. G., G. V. Koppenol-Gonzalez, and J. Reiss. (2019). "Extrapolation of Experimental Results through Analogical Reasoning from Latent Classes", *Philosophy of Science*. 86(2): 219-235.

Vivalt, E. (2019). “How Much Can We Generalize from Impact Evaluations?”. Unpublished manuscript, ANU, Canberra. Retrieved from: http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf (accessed February 28, 2019)

White, H. (2013). “An introduction to the use of randomised control trials to evaluate development interventions”, *Journal of Development Effectiveness*, 5(1): 30-49.

Woodward, J. (2002). “What Is a Mechanism?: A Counterfactual Account”, *Philosophy of Science*, 69: 366–77.

— (2003). *Making Things Happen: A Theory of Causal Explanation*, New York: Oxford University Press.

Worrall, J. (2002). “What evidence in evidence-based medicine”, *Philosophy of Science*, 69: 316-30.

— (2007). “Why there's no cause to randomize”, *British Journal for the Philosophy of Science*, 58: 451-88.

Zigler, C. M., and F. Dominici. (2014). "Point: Clarifying Policy Evidence With Potential-Outcomes Thinking—Beyond Exposure-Response Estimation in Air Pollution Epidemiology", *American Journal of Epidemiology*, 180(12): 1133-40.

CHAPTER 2

What's Extrapolation?

2.1 Introduction

To build a basis for my subsequent contributions, it is important to develop a comprehensive characterization of what extrapolation is, what makes extrapolation challenging, and what kinds of extrapolation there are. To develop this background I will proceed as follows:

In *Section 2* I begin with a sketch of what I take extrapolation to be, at the most general level. This will be the starting point for developing a more comprehensive and precise view of extrapolation as a highly heterogeneous collection of inferential activities that, while amenable to a single, systematic analysis, can exhibit important differences.[1] Building such a view is useful in three ways: 1) it allows us to distinguish existing strategies for extrapolation with respect to what kinds of problems of extrapolation they can, in principle, address (and criticize them accordingly for their limitations); 2) it helps articulate why some kinds of extrapolation are more difficult than others, and how there may be cases that are unlikely to be overcome by any strategy; 3) it helps pinpoint what exactly the criticisms to be developed in subsequent chapters latch onto.

To build this view, *Section 3* elaborates why extrapolation is challenging. I approach this issue from two angles. One concerns the underlying reasons for why extrapolation is challenging, i.e. different kinds of causally relevant differences between populations. The other concerns the symptoms of such differences, i.e. *heterogeneous causal effects* that differ systematically between individuals and between populations. I will begin with the symptoms because it is the symptoms of heterogeneous causal effects that we face at the level of observable phenomena, i.e. the level at which effect estimation, prediction, and extrapolation proceed. After offering a brief overview of different ways in which these symptoms can manifest, I turn to elaborate on the various underlying reasons for why these symptoms can occur. In doing so, I begin from standard

[1] cf. Steel (2010; 2013), who recognizes that different kinds of extrapolation should be distinguished concerning the type of causal claim to be extrapolated and that different kinds of extrapolation might require different conditions to proceed successfully.

approaches to explaining heterogeneous causal effects, which proceed in terms of so-called *moderating variables*, i.e. variables that causally interact with a treatment to induce differences in causal effects between individuals and between populations. I then expand on several additional, but heretofore unrecognized, types of heterogeneity, which are induced by *mediating variables,* and some kinds of intervention and outcome variables, which I call *constrained* intervention and outcome variables. I argue that these variables, too, can induce heterogeneous effects, and that differences in the distribution of such variables can hence pose important obstacles to successful extrapolation.

Building on this, *Section 4* offers a more comprehensive analysis of causally relevant differences between populations. Specifically, in addition to differences in causal effects induced by differences in *variables*, i.e. moderating, mediating, and constrained intervention and outcome variables, I expand on two additional levels at which causally relevant differences can obtain: 1) differences in functional form and parameters (causal and statistical) and 2) differences in the basic structure of causal mechanisms.

With this overview of causally relevant differences in place, *Section 5* synthesizes a general analysis of problems of extrapolation as a highly heterogeneous set of inferential challenges. Extrapolation here is distinguished along several dimensions, including the types of underlying causally relevant differences that pose obstacles to extrapolation, the kinds of causal queries at issue, the envisioned kind and fidelity of the inferences to be drawn, etc. The main aim is to make clear that extrapolation comes in various kinds, some of which are significantly more challenging than others. This not only improves on standard unanalysed notions of extrapolation used in the literature, but also helps build a framework in which my subsequent critical assessment of extant strategies for extrapolation can proceed, e.g. clarifying which of these strategies are useful for addressing which kinds of problems, and which types of problems pose distinct, and, at times, insurmountable, challenges for these strategies respectively.

2.2 What's an Extrapolation? The Short Story

To get us started, let me offer a basic and preliminary sketch of what I take extrapolation to be at the most general level. Surprisingly, there are few explicit attempts in the literature to characterize extrapolation. One exception is Daniel Steel (2008), who begins by offering some examples of what he considers to be instances of

problems of extrapolation: e.g. when we know that a particular substance is a carcinogen in rats and would like to know whether it is also in humans, or when we have evidence from an RCT that a particular social policy intervention is effective and want to know whether it will be similarly effective in other locations, or when scaled up. Against the background of these examples, Steel proposes the following broad characterization:

> "In each of these cases, one begins with some knowledge of a causal relationship in one population, and endeavours to *reliably* draw a conclusion concerning the relationship in a *distinct* population." (2008, 3)

We have a reasonably clear understanding of what it means to possess *knowledge* of a causal relationship or causal effect in an experimental population. This is the case whenever we have identified a causal effect in an experimental population, the method used for identifying this effect is principally adequate to this purpose, and its conditions of applicability are satisfied beyond some reasonable degree of confidence, i.e. its main identification assumptions are justifiably believed to be met.

What is currently open is what it means to draw *reliable* conclusions about a causal relationship or effect in some target population, as well as what it means for the target population to be *distinct* from the experimental population. As I will argue more fully throughout this chapter, the precise nature of the relationship between experimental and target populations is an important issue when it comes to distinguishing different kinds of problems of extrapolation according to how challenging they are.

For now, it suffices to give a rough first-pass analysis of extrapolation. Extrapolation minimally involves 1) a causal effect, relationship, or claim to be extrapolated (from), 2) a causal effect, relationship, or claim pertaining to a distinct population to be inferred as a result of the extrapolation, and 3) a *basis for extrapolation* which justifies an inference from 1) to 2); this is typically a conjunction of background theory (including a theory of causality and causal inference), empirical assumptions about the populations of interest, and supplementary knowledge and evidence that underwrite these assumptions.

I will not make more specific commitments as to whether I take extrapolation to be an inductive or deductive (or other) form of inference. While widely taken to be inductive, Cartwright's (2012; 2013) proposals for how to address problems of extrapolation involve deductive arguments. For the moment, it is enough to note that the

mode of inference joining 1) and 3) to infer 2) is such that it aims to be either logically valid, or invalid but strongly compelling in yielding the extrapolative conclusion of interest. With this in mind, let me proceed to expand on the reasons for why extrapolation is challenging.

2.3 Heterogeneous Treatment Effects

The causal effects of one and the same intervention can, and will typically, differ between different individuals, and between different populations. This is why we need to worry about extrapolation in the first place. If causal effects were the same everywhere and for everyone, there would be no need to be concerned: what is effective in one place, time, or individual will be effective in any other, and to the same extent.

In virtually all areas targeted by EBP initiatives, however, causal effects are usually not the same across individuals, places, or times. For instance, in development economics, the effects of microfinance interventions may differ significantly between individuals as a function of whether they will pursue profitable or unprofitable business plans with the microfinance loans. This, in turn, may differ as a function of prior business ownership experience or education. Similarly, in educational policy, the effects of interventions such as supplementary teaching or class size reductions may differ significantly between settings. Students might be stigmatized for receiving supplementary teaching in one population, but not in another. Labour supply of qualified teachers may differ between populations, so increased demand for teachers will have different effects on average teacher quality in different populations, which in turn may bear importantly on student achievement outcomes. In economic policy, the effects of interventions such as universal basic income or minimum wage policies may differ significantly between populations as a function of differences in population-specific features of the labour market, and social and institutional arrangements that bear on agents' labour supply decisions in response to changes in incentives. More generally, the empirical literature gives us good reasons to believe that there will frequently be substantial differences in causal effects (both in magnitude and sign) of the same or similar interventions between populations (Vivalt 2019).

In the econometrics literature, causal effects that differ between individuals, times, or places are called *heterogeneous treatment effects* (HTEs). Formally, in accordance with the standard potential outcomes framework by Rubin (1974) and Holland (1986), HTEs

obtain among individuals i and j whenever individual treatment effects (ITEs), i.e. differences between $i's$ and $j's$ respective potential outcomes in the presence and absence of treatment, differ. Let Y be the outcome of interest, X the intervention variable, u an idiosyncratic error that captures the effect of other variables on Y, and $g(\cdot)$ be a function that captures the causal relationship that obtains between X and Y. Let the outcomes equations for i and j be:

$$Y_i = g_i(X_i, u_i)$$

$$Y_j = g_j(X_j, u_j)$$

Let X be a binary variable $X\epsilon[0,1]$, with $X = 0$ denoting untreated and $X = 1$ denoting treated status respectively. Then, the ITEs of an intervention ΔX (Δ indicating that the value of X is changed) for i and j respectively are just:

$$\tau_i = Y_i(1) - Y_i(0)$$

$$\tau_j = Y_j(1) - Y_j(0)$$

Then, the causal effects of a given intervention ΔX are heterogeneous if and only if:

$$\tau_i \neq \tau_j$$

HTEs are often considered to be systematic in nature, that is, they are not considered to be random, inexplicable variations that cannot be accounted for by reference to some substantive causally relevant difference between i and j. Rather, the standard view in the literature is that differences in ITEs between individuals, and more generally HTE between populations and settings, prevalently obtain as a result of underlying causally relevant differences. That is, if there are no genuine causally relevant differences between individuals who experience different effects in response to one and the same intervention, this residual variation would be due to the indeterministic nature of the mechanisms underlying the events of interest as well as measurement error – but there will not be any significant, but ultimately and in-principle inexplicable differences in causal effects experienced by homogeneous individuals.[2] I will follow this convention for the remainder of the discussion, although I will add some important qualifications to it in *Chapter 3*.

[2] In contrast to the theoretical literature, methodological guidelines in EBP often take a different stance on HTE by treating it as a symptom of measurement error (see Khosrowi 2019 for an overview).

2.3.1 Levels of Heterogeneity

HTEs can obtain at various levels. First, HTEs can obtain at the level of individuals. This kind of heterogeneity is difficult to observe since ITEs can often not be directly observed in principle, and at best approximated (Holland 1986). However, there are cases where, even without accurate estimation of ITEs, we have good reasons to believe that effects are heterogeneous. Take for instance a job-training programme, which aims to increase participants' job market prospects by offering workshops that help improve their CV quality. Say we have two individuals, i and j, where i's pre-intervention CV quality is bad, and j's CV quality is excellent. It seems plausible to think that the causal effect of the training programme will be smaller for j than for i, since j's CV is already close to being perfect beyond improvement.

A second kind of HTE obtains at the site level. Here the idea is that characteristics of the particular place (or time, or both) where an intervention is implemented can induce differences in causal effects between sites. Suppose we are interested in the effects of an intervention that aims to increase student performance by offering free supplementary teaching to students to help them review material discussed in class. Suppose we have two school districts, i and j, where students in i are stigmatized by their peers for being in need of supplementary teaching, making them less confident in their abilities and decreasing their performance on tests, but students in j are not stigmatized in this way. Here, the causal effects of the intervention may systematically differ at the site level: all students in i experience stigmatization effects and no students in j do, so these features only differ *between* i and j but not *within* either population. More generally, site-level heterogeneity obtains in cases where environments have characteristics relevant to the effect of interest and these characteristics carry over to individuals in virtue of their being in a particular place, time, or setting. This can include social norms, institutional arrangements, sociodemographic features (when sites are homogeneous in this respect), and other kinds of 'blanket features' that generally, or predominantly, affect individuals in virtue of belonging to a particular population.

Finally, there is also implementer-level heterogeneity (see e.g. Muller 2015). Here, the idea is that the agents who implement the interventions of interest, e.g. local government officials, NGO workers, etc., can have specific causally relevant features that bear on the effects experienced by the agents to whom they administer the treatment. In the simplest case, populations A and B are identical in features relevant to

the causal effect of interest, and treatment in A and B is administered by agents i and j respectively. If i and j possess causally relevant features that differ, such as different levels of experience in implementing the intervention of interest, this can induce differences in the effects experienced by individuals who receive treatment from i and j respectively. Implementer-level heterogeneity is important because even if populations are believed to be relevantly similar in individual and site-level characteristics, inadequately or differently-trained implementers can still induce significant differences between the effects expected in the target and those that will eventually obtain.[3]

In addition to these three levels, there may be yet other levels of description at which HTEs can obtain. These will be bracketed for the remainder of the discussion as the above levels are arguably the most important ones.

Let me proceed to elaborate in more detail on the underlying reasons for why HTEs obtain: causally relevant differences between individuals, settings, implementers, and so forth. Such differences can be realized in various ways. I begin with the key concept that is used in the extant literature to explain HTEs: *moderating variables*. I then proceed to identify three further, heretofore unrecognized sources of HTEs: mediating variables and constrained intervention and outcome variables.

2.3.2 Moderating Variables

Following Baron and Kenny (1986, 1174) moderating variables are qualitative or quantitative variables MO that can change the magnitude and/or sign of the marginal causal effect between a pair of variables X and Y. So different realizations of MO will induce, or at least coincide with (more on this shortly), different causal effects of a given intervention.

To give a simple example, let Y be individuals' self-reported headache intensity, let $X\epsilon[0,1]$ represent whether an individual takes Aspirin, and let MO be age. Let us suppose that, for some underlying physiological reason, the marginal causal effect of setting X from $X = 0$ to $X = 1$ for individuals i and j differs as a result of differences in their age MO.

[3] Note that, in principle, site- and implementer-level HTEs can be explained at the individual level, e.g. when we describe features of sites or implementers as features that are 'imposed on' individuals.

Graphically, this situation is often represented by an arrow-on-arrow arrangement as in *Figure 1* below (see Weinberg 2007; see Elwert 2013, 255 for arguments against this representation). Here, there is a causal arrow from MO that points *onto* the causal arrow from X to Y, to indicate that the causal relationship between X and Y is co-determined by, and changed under intervention on, MO.

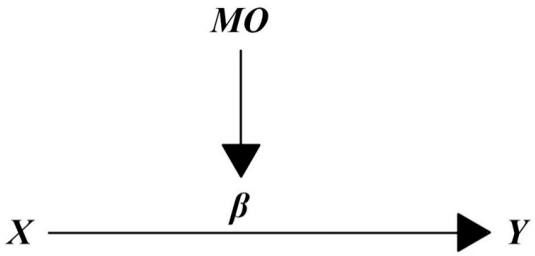

Figure 1: Moderation

There are various types of moderating variables that can be distinguished. For instance, moderators can be continuous, dichotomous, or categorical. Moreover, the way in which causal effects of X on Y are co-determined by moderating variables can differ importantly as well.

First, an effect can be *fully moderated*, in which case all contributions of X to Y are produced in interaction with the moderator and there is no separate, direct effect of X on Y that would remain unaffected by changes in MO. This can be represented by:

$$Y = g(X, MO, \beta) + u$$

Here, u captures the effect of all variables on Y that do not interact with X, and $g(\cdot)$ is an interactive function of X and MO, such as $g(X, MO, \beta) = \beta * X * MO$. Here, $g(\cdot)$ cannot be additively separated into two functions $f(\cdot)$ and $h(\cdot)$ where either of these functions does not depend on MO. In this case, MO and X are equally privileged in bringing about or curtailing changes in Y, e.g. when $MO = 0$, there is no marginal effect of X on Y. Likewise, if $X = 0$ then changes in MO cannot effect changes in Y. In these cases, it would be possible to say that what counts as the intervention variable and what counts as the moderating variable is a matter of our epistemic or pragmatic interests and not a substantive distinction between different types of causally relevant variables.

This is different in the case of *partially moderated* effects. Here, the moderator interacts with the treatment to produce an idiosyncratic contribution that depends on the

value of the moderator, but there is also a separate, unmoderated effect of the intervention variable on the outcome; i.e. a baseline effect of X on Y that we get no matter the value of MO. This can be represented by:

$$Y = g(X, MO, \beta, \gamma) + u$$

In contrast to fully moderated effects, here we allow that $g(\cdot)$ is additively separable into $f(\cdot)$ and $h(\cdot)$, and either $f(\cdot)$ or $h(\cdot)$ do not depend on MO, such as when:

$$Y = f(X, \beta) + h(X, MO, \gamma) + u$$

Here, $f(\cdot)$ does not depend on MO, but $h(\cdot)$ does. $f(\cdot)$ hence captures the unmoderated, baseline effect of X on Y, and $h(\cdot)$ captures the moderated effect that is produced in interaction of X and MO.

In this case, the intervention variable is causally privileged over the moderating variable. While, in principle, a wide range of changes in Y may still be effected by an intervention on either X, or MO, or both, not all changes in MO will have effects on Y. Unlike in the fully moderated case, this implies that we cannot represent the same situation with two different graphs by exchanging MO and X:

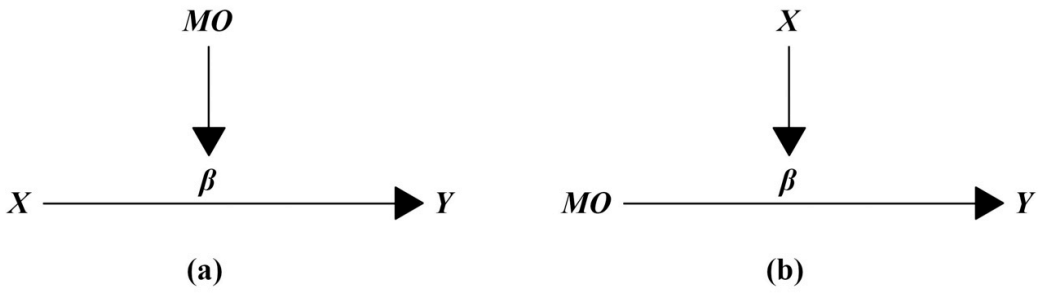

Figure 2: MO and X are not exchangeable between a) and b)

For partially moderated effects, these diagrams would be inequivalent, since *in Figure 2a)* interventions on X can have an effect on Y even if $MO = 0$, but not the other way around, and in Figure 2b) the situation is reversed. For instance, if in *Figure 2a)* $h(X, MO, \beta) = \beta * X * MO$, then if $X = 0$, changes in MO do not effect changes in Y. This is in contrast, for instance, to Cartwright's conception of *causal support factors* (Cartwright 2013), which seems closer to the fully interactive interpretation of moderating variables. I will discuss the relation between moderating variables and causal support factors in more detail in the appendix to *Chapter 4*. There, I will also discuss cases of *non-linear moderation, i.e.* where the marginal causal effect of X on Y varies non-linearly over MO, including cases where $g(\cdot)$is a stepwise function of MO,

such as when MO has a threshold value λ below which $\Delta Y/\Delta X = 0$, and above which $\Delta Y/\Delta X = \gamma$.

There are several additional variations on the above cases that have been discussed in the literature (see e.g. Kraemer et al. 2001; 2002; Marsh et al. 2013). These will not be discussed here, as the qualitative distinctions above are sufficient to make clear how the most important types of moderating variables can induce differences in causal effects between individuals, and between populations.

With these general points in place, let me turn to an important clarification. The concept of moderating variables seems innocuous at first, but there is an important distinction to be made between a *causal* and *statistical* interpretation of moderating variables. More specifically, the definition of moderating variables provided above, i.e. variables MO that induce differences in a causal effect between X and Y, is a *causal* notion of moderating variables: it makes clear that differences in MO *causally induce* differences in the effect of X on Y. This also implies that there can be interventions on MO that will induce differences in the potential outcomes of individuals, and hence in the causal effects experienced by individuals with respect to one and the same intervention on X.

Yet, this definition alone does not yet help us identify moderating variables from data. For this, an alternative operationalization is needed. One candidate is to say that moderating variables are variables MO, where the causal effect of X on Y differs over different levels of MO. This can be tested by modelling a statistical interaction between X and MO, e.g. as $Y = \alpha * X + \beta * MO + \gamma * MO * X + u$. Here, MO is a moderator if the interaction $MO * X$ is significantly correlated with the outcome (and potentially also uncorrelated with X, see Baron and Kenny 1986). This would be a *statistical notion* of moderating variables. However, a variable MO satisfying this operationalization is neither necessary nor sufficient for it being a moderating variable in the causal sense. *Figure 3* illustrates:

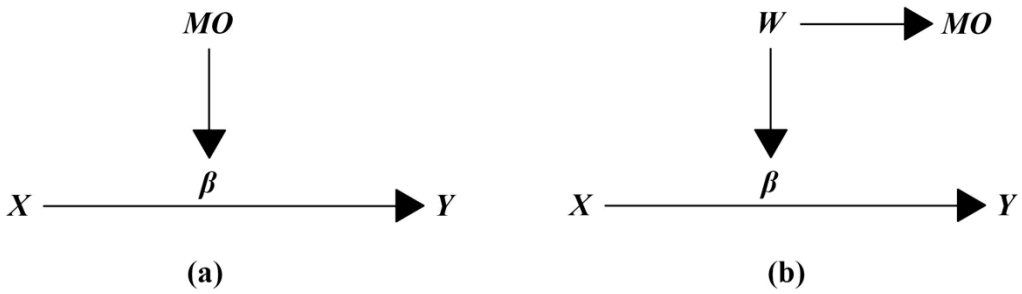

Figure 3: MO is a causal moderator in a) but not in b)

In *Figure 3a)*, MO is a genuine, causal moderating variable, represented by the arrow pointing from MO onto the path connecting X and Y. In *Figure 3b),* W is a causal moderating variable, and MO is a child of W. Because MO and W are correlated, and W is correlated with the magnitude of the causal effect of X on Y, MO will satisfy the statistical definition of moderating variables, but not the causal one.[4]

This distinction is important. While it is not essential to distinguish between causal and statistical moderators for investigating whether there is heterogeneity in effects across MO, or for some cases of predicting differences in causal effects of X on Y across MO in a novel target[5], the causal/statistical distinction is crucially important for *intervention*. If one suspects that MO is a causal moderating variable of an X-Y-effect, e.g. on the basis that higher levels of MO coincide with larger effects of X on Y, then this may give us reasons to believe that intervening on MO in addition to X can help us achieve larger causal effects by creating favourable realizations of MO that are believed to further increase the effects of interest. Such attempts will generally remain unsuccessful, however, if the variable to be intervened on is not a genuine causal moderator, but merely a close correlate of some other variable W that is a causal moderator, such as in 3 b).

These cases can, in principle, be disambiguated from observational data but only if W is known and can be conditioned on, which is not always the case and may be difficult if W is a latent variable (such as a psychological characteristic) that cannot be

[4] Similarly, there can be cases where a variable MO downstream of Y will satisfy the statistical notion of moderating variables, as many variables that are (co-)determined by Y will correlate not only with Y but also with changes in Y induced by X, i.e. causal effects. These cases, of course, would not satisfy the causal notion of moderating variables, and for identification of moderating variables from statistical data, our statistical notion will consequently need to be refined to exclude these cases.

[5] Specifically, cases where MO and W (or some relevant analogue) play the same causal role in both populations, and one has strong reasons to believe that this is so.

easily measured in the first place. Alternatively, an obvious test to help tell whether a variable MO is truly a moderator in the causal sense would be to manipulate MO, or indeed both X and MO, in an experiment (see Imai et al. 2013).

This completes my overview of moderating variables. Let me expand on other types of variables that can induce HTEs: mediating and constrained outcome/intervention variables.

2.3.3 Mediating Variables

A second important class of variable that has the capacity to induce differences in causal effects between individuals, and between populations, is that of *mediating variables*. In contrast to moderating variables, which induce differences in causal effects by meddling with the relationship between X and Y, mediating variables are variables that sit, as it were, on the causal pathway(s) between X and Y. *Figure 4* represents the simplest case, where there is one causal pathway connecting X and Y, and this pathway is mediated by Z, so all variation in Y induced by changes in X is transmitted through Z, and manifests itself as variation in Z.

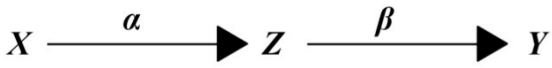

Figure 4: Mediation

At this stage, readers familiar with the literature on causal inference, extrapolation of causal effects, and related topics might wonder why mediating variables are discussed as candidates for inducing HTEs. Indeed, on standard treatments of mediating variables, this would be surprising. Mediating variables *transmit* causal effects, but they do not *meddle* with them. Consider the probabilistic dependence and independence characteristics of Z in *Figure 4*. Z is correlated with X and Y, and uncorrelated with Y conditional on X. And while Z is correlated with Y unconditionally, it is not correlated with differences in potential outcomes $Y(1) - Y(0)$, i.e. causal effects. Specifically, higher pre-intervention values of Z will coincide with higher post-intervention values of Y (assuming that α and β are positive). So Z is correlated with the *levels* of potential outcomes, and hence the *levels* at which differences in potential outcomes, qua causal

effects of interventions on X, are realized. But Z is uncorrelated with *differences* in potential outcomes, i.e. the *magnitude* of causal effects. This can be easily understood with basic differential calculus. Let

$$Y = \alpha * Z$$

$$Z = \beta * X$$

then

$$Y = \alpha * \beta * X$$

and $\Delta Y/\Delta X = \alpha * \beta$, so causal effects, do not hinge on initial values of Z (or indeed X).

To give an intuitive example, if a job training programme is supposed to improve an individual's job prospects by means of increasing the quality of her CV, then her pre-intervention CV quality will determine whether her CV will improve from, say, moderate to good, or rather from good to very good, i.e. it will affect the *levels* at which treatment effects obtain. But the pre-intervention CV quality will not affect the *magnitude* of the effect on CV quality induced by the training programme, if we assume for the moment that the difference from moderate to good is evaluated the same as the difference from good to very good (I will say more on this shortly).

However, there are three important and related types of cases where, contra these intuitions, differences in mediating variables can induce differences in causal effects: 1) non-linearly associated mediators, 2) bounded mediators, and 3) dichotomous mediators.

First, non-linearly associated mediators are mediators that transmit variation from X to Y in a non-linear fashion. Specifically, let us assume the simplistic causal model from *Figure 4*. Let the equations describing this model be:

$$Y = g(\alpha, Z)$$

$$Z = f(Z_0, \beta, X)$$

where $g(\cdot)$ is the function capturing how Z bears on Y and $f(Z_0, \beta, X)$ hinges on the pre-intervention value of $Z = Z_0$, such as when $f(Z_0, \beta, X) = Z_0 + \beta * X$. Now, if $g(\cdot)$ is non-linear in Z, such as when $g(\alpha, Z) = \alpha * Z^2$ then the pre-intervention value of Z, Z_0, will bear on the magnitude of the causal effect of interest. This is easy to see if we take the first derivative of Y with respect to Z: $g'(Z) = 2\alpha Z$. The marginal effect

induced by a given intervention on X will depend on the initial value of Z, which in turn depends on the initial value of X.

As an intuitive example, consider the case of an educational intervention that aims to improve students' ability to translate Latin texts to English by means of lessons that help increase the scope of their vocabulary. We can imagine that students with more extensive pre-intervention vocabulary may benefit more from the lessons, as the newly learned words complement their pre-existing vocabulary in ways that allow them to translate whole text passages without substantial difficulties. For students with low pre-intervention vocabulary scope, the same increment of newly learned words may have smaller effects as complementarity with pre-existing vocabulary may not materialize for lack of such pre-existing vocabulary.

This can also work the other way around. The marginal effect of grammar tutorials to increase students' understanding of Latin grammar in order to improve their translation performance may importantly depend on pre-intervention understanding of grammar. If the lessons are pitched at a relatively low level, then students with excellent pre-intervention understanding of Latin grammar may not be able to reap substantial benefits given much of the lesson covers material that is redundant to their well-developed pre-intervention understanding of Latin grammar.

More generally, non-linear functional form associations can be an important source of HTEs between individuals and between populations. In some cases this is obvious. For many educational interventions it is not surprising that interventions that are targeted to improve low- or high-ability students' performance respectively will not be as effective if delivered to the other group. There are clearly other cases, however, where this is not as obvious. For instance, dose-response relationships are a central concern in epidemiological studies, and while there is a generally well-motivated suspicion that dose-response relationships might be non-linear, there are usually no obvious ways to determine even rough features of the functional form before extensive empirical investigation. Similarly, in many EBP settings it will not only often be difficult to estimate functional form associations from observational or experimental data, but there is also, perhaps unfortunately, often little concern for investigating such issues to begin with. So while non-linear functional form associations of mediators are clearly important for HTEs, and consequently for extrapolation, they remain understudied.

A second, related case where mediating variables can induce HTEs concerns *bounded mediating variables*, i.e. variables that have a natural or otherwise induced upper or lower bound. For instance, consider again the job-training programme that is supposed to increase jobseekers' CV quality. Here, it seems plausible to think that CV quality has both an upper and lower bound, beyond which changes to a CV will not matter for the assessment of its quality. Let us assume that two individuals i and j have good and excellent CVs respectively. Specifically, j's CV is excellent beyond improvement; there is no feasible change to her CV that would induce changes in the assessment of its quality, or, indeed, j's subsequent job prospects. If this is the case, then one and the same intervention will have different marginal effects on the outcome for i and j respectively, depending on their pre-intervention CV quality. This is essentially a special case of non-linear functional form association discussed before but merits separate discussion because the reasons for why certain variables are bounded can sometimes be more easily appreciated (and measured) than whether or not they are non-linearly associated with an outcome of interest.

Finally, the third case where mediating variables can induce HTEs concerns *binary mediating variables*. Let the intervention of interest be a GMAT[6] training class and the outcome be long-term average earnings. Let us assume that earnings are importantly determined by whether agents attend ivy-league universities and that there is a sharp threshold GMAT score λ that agents need to surpass in order to attend ivy-league universities. The mediator of interest here is ivy-league university attendance, which is a range-property supervening on GMAT score. For simplicity, we assume that scoring above λ on their GMAT will make agents attend ivy league college (without defiers, i.e. so all agents above λ are guaranteed to attend ivy league college), and scoring below λ will preclude them from doing so (without defiers, i.e. no agent below λ will attend ivy league college). Let us assume that GMAT performance is determined by a latent variable called *GMAT Skill*. If pre-intervention GMAT Skill is beyond a level where individuals surpass λ with near certainty, then this means that a GMAT training class will not be effective in changing their propensity to get into ivy-league universities, and hence their subsequent earnings. Individuals are *saturated* with respect to GMAT Skill. Even though skill can still be improved in a way that would increase individuals' test scores even further beyond λ, such improvements do not bear on the subsequent

[6] *Graduate Management Admission Test*, an admission test used for governing entry into various study programmes.

outcomes of interest. Hence, similarly to the above cases, the case of binary mediators is one where individuals' potential outcomes are a non-linear (in this case stepwise) function of the pre-intervention value of a mediating variable. So, under some conditions, pre-intervention values of binary mediators can induce important heterogeneity in causal effects between individuals, and consequently between populations.

Let me expand on two additional sources of HTEs at the level of variables, which are related to the cases discussed above.

2.3.4 Constrained Intervention and Outcome Variables

So far I have discussed the role of variables that sit on, as it were, or otherwise directly affect the causal pathways connecting treatment and outcome variables, i.e. moderating and mediating variables. Similar to these variables, both the outcome variables as well as the intervention variables themselves can also induce HTEs (see e.g. Keane 2010).

First, pre-intervention values of intervention variables can induce HTEs in the following way: Let X be a binary variable with values $X\epsilon[x', x'']$ and $x'' > x'$. Let X_0 be the pre-intervention value of the intervention variable X. Let the intervention of interest be such that it sets $X = x''$ for all individuals. Then, for two individuals i and j, with $X_{i,0} = x'$ and $X_{j,0} = x''$, the intervention will change X only for i but not for j. Hence, if $\Delta Y/\Delta X \neq 0$ for i, causal effects experienced by i and j will be different.

Second, essentially the same applies when X has a lower or upper bound and either i or j exhibit X_0 at the lower (upper) bound X_{down} (X_{up}), and the intervention sets $X = X_{down}$ (X_{up}), or near $X = X_{down}$ (X_{up}) for all individuals.

A third way in which pre-intervention values of X matter for effect magnitudes is when an intervention adds some constant amount ΔX to X_0, individuals differ in their pre-intervention value X_0, and when the outcome is non-linear in X (or non-linear in any mediator Z on the pathway between X and Y).

Similar concerns apply to outcome variables. When outcome variables are bounded, then causal effects experienced by individuals can differ with respect to their pre-intervention values of the outcome Y_0. If individuals are already within a region ε smaller than $\Delta Y/\Delta X$ from the upper or lower bounds of Y, then, other things being

equal, an intervention on X that will move individuals closer towards either bound will have smaller effects for individuals within ε of Y_0 than for individuals at or outside of ε.[7]

For now, this overview is sufficient to make clear that, in different ways, causal effects can vary between individuals, and hence between populations, depending on the values that different kinds of variables associated with individuals assume, specifically moderating, mediating, intervention, and outcome variables. Even without further consideration of the likelihood of substantial differences between individuals in these respects, the sheer number of ways in which individuals may differ in causally relevant characteristics suggests that HTEs are likely to obtain.

2.4 Species of Causally Relevant Similarities and Differences

So far the discussion of causally relevant differences that can induce HTEs, and hence present obstacles to extrapolation, has focused on the role played by differences in the values of variables. However, this is not the only level at which causally relevant differences can obtain between individuals and between populations. Specifically, causally relevant differences at the level of variables only mark one of at least three different levels at which such differences can obtain. In addition to differing in the realizations of variables, individuals and populations may also differ at the level of the *functional form* of causal relationships and the *parameters* governing the causal effects between variables, as well as at the level of the *basic structure of the causal mechanisms* governing the outcomes of interest. Let me provide a brief overview of these three levels to establish a more comprehensive picture of the various ways that causally relevant differences can obtain between individuals and between populations.

2.4.1 Variables

The first level at which individuals and populations may exhibit causally relevant differences concerns the *realizations/distributions of variables*. The kinds of variables that can induce such differences at this level have been extensively discussed above.

[7] See e.g. Sundell et al. (2008) for a case from social psychology where the status quo before intervention is so good that treatment cannot improve much on it.

So far I have mostly focused on differences in variables at the individual level. In extrapolation we typically care about differences in ATEs between populations, however. So a more general statement of the concern about differences in variables is that ATEs between populations can differ if the *distribution* of variables that can induce differences in causal effects differs between populations.

For instance, if the mean of a moderating variable MO differs between populations A and B, then ATEs in A and B will often differ. Similarly, while the mean of MO might be the same in A and B, its variance may differ, so the variance of ITEs in A and B may differ as well. Analogous differences can apply to other moments of distributions, too, such as kurtosis or skewness. Here, one and the same intervention with one and the same ATE in both populations may put more or less distributional mass on extreme values in one population than in another. In general, differences in distributions of variables that bear on the magnitude or sign of the effects of interest may not necessarily manifest themselves in differences in ATEs, but in ways that are less immediately obvious. This is important whenever extrapolation involves not only concerns for whether ATEs will be similar in the target as in an experimental population but involves more ambitious aims regarding welfare analysis, e.g. whether an intervention that satisfies certain distributive desiderata in an experimental population will also satisfy these desiderata in a distinct target.[8]

In addition to these basic concerns, there are also some more intricate concerns that can add significant complexity to the challenges posed by differences in variable distributions. Specifically, individuals, and populations, may frequently differ in more than one variable relevant for the causal effects of interest. This is important in multiple ways. For one, even minor differences in variables, if there are many such differences, can compound to significant differences in ATEs between populations if these differences have the same qualitative bearing on the effects of interest. Conversely, differences in causal effects induced by differences in variables can also attenuate each other or cancel each other out, so observing differences in variables known to induce HTEs does not licence the conclusion that ATEs will be different unless one is

[8] This is not to suggest that learning the distribution of ITEs in an experimental population is ever straightforward. Sophisticated subgroup-analyses can be helpful for this purpose, but they are subject to important methodological concerns (see Varadhan and Seeger 2013; Khosrowi 2019 for overviews). While still rare, it is hoped that such analyses will be more common in the future. To the extent that they will be, it is important to recognize that extrapolating conclusions about distributional features to novel targets poses additional challenges to extrapolation; these will not be discussed here.

somewhat confident that there are no other differences that might attenuate or cancel out those of primary interest. Finally, it is important to note that moderating variables may also interact not just with the treatment itself but also with other moderating variables (see e.g. Fuller 2018). This adds further complexity to the assessment of how differences at the level of variables will bear on differences in causal effects between an experimental and target population. I will return to discuss some of these complexities in more detail in the appendix to *Chapter 4* when exploring Cartwright's conception of *causal support factors*. For now, let me expand on two additional levels at which causally relevant differences can obtain.

2.4.2 Parameters and Functional Form

The second level at which populations can differ in causally relevant respects concerns the *structural parameters* associated with the variables that figure in the causal mechanisms governing the effects of interest and the *functional form association* of these variables. X can be causally relevant for Y in an experimental and target population, yet the particular *way in which* X is relevant for Y, i.e. the structural parameter capturing its effect on Y, or the functional form of the structural equations best representing the causal relationships between X and Y, can differ between populations.[9]

In the simplest case the mechanism connecting X and Y is one unmediated path from X to Y, and the parameter β governing the marginal effect of X on Y, or the functional form association between X and Y, differs between populations A and B, so the causal effects of a given intervention on X will differ between A and B.

There are several variations on this simplistic setting. For instance, the way in which moderating variables affect causal relationships between X and Y can differ between populations, too. For instance, for one and the same intervention, higher values of MO might induce larger effects in population A, but smaller effects in population B. Here, the sign of the parameter associated with MO differs between populations, but differences in magnitude can be similarly important in bringing about significant differences in causal effects between populations.

[9] I will not discuss issues of functional form differences in detail at this stage – these will be discussed in later chapters in the context of concrete extrapolation strategies that make assumptions about functional form *similarities*.

Another important class of cases involving differences in parameters concerns the parameters associated with the pathways determining and originating from mediating variables between X and Y. This is important for instance in cases of partially mediated effects, e.g. when there is a causal pathway from X to Y that is mediated by Z, but there is also a second, unmediated causal pathway from X to Y. *Figure 5* visualizes:

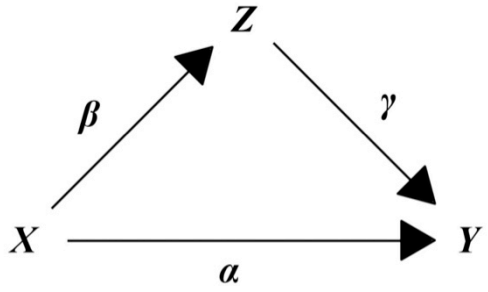

Figure 5: Partial mediation

Here, the relation between the path parameters α, β, and γ determines what proportion of the X-Y-effect is mediated by Z and what proportion is unmediated. Differences in α, β, and γ can importantly change this proportion, and hence have significant bearing on differences in causal effects between populations.

This applies to somewhat more complex cases of *moderated mediation* and *mediated moderation* as well (see Muller et al. 2005):

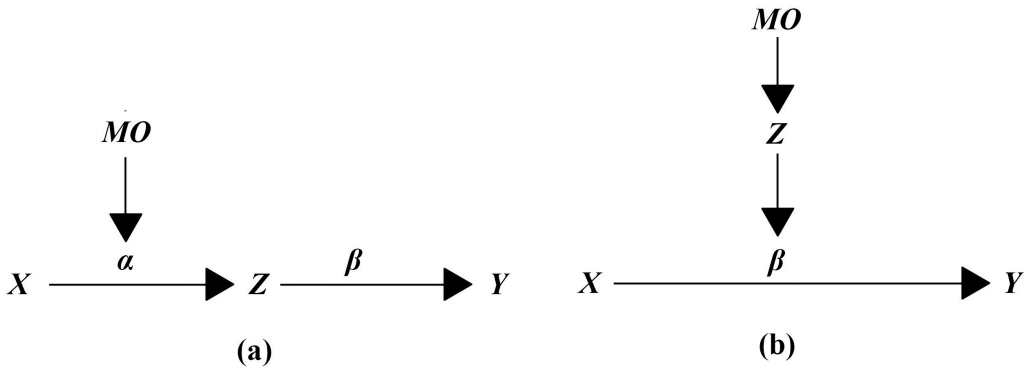

Figure 6 a) moderated mediation*; b)* mediated moderation

In moderated mediation (*Figure 6a*), either the path from X to a mediator Z, or the path from Z to the outcome Y are moderated (or both). In mediated moderation (*Figure 6b*), the effect of a moderator on another causal pathway is mediated by some variable

Z. Both cases can potentially further aggravate differences in causal effects brought about by differences in parameters when MO differs between populations.

Another type of parameter worth mentioning is the *meta-parameter*, i.e. a parameter that captures the rate at which another variable or parameter can change. Such parameters may often be constant for individuals but may differ importantly between individuals. For example, behavioural nudge interventions often aim to intervene with agents' behavioural response profiles with respect to a given institutional/incentive structure. In these cases, it often seems that the primary targets of interventions are not variables, such as household endowment, but instead structural parameters. For instance, a well-known nudge intervention called the Save More Tomorrow programme (Thaler and Benartzi 2004) aims to increase employees' retirement savings contributions. This intervention aims at meddling with the ways that agents perceive the subjective value of present and future payments. Here, the idea is that because agents tend to discount the value of future rewards over present rewards, the scheme can increase agents' savings for retirement by offering them a choice to commit to saving a share of future salary raises towards retirement rather than consumption in the period when the salary raises are received. This could be reconstructed as intervening on a structural parameter, i.e. the parameter that converts income in period t into savings and consumption respectively, by having agents commit to choices that derive from the parameter value that applies to future rewards, rather than having them behave in accordance with their parameter for present rewards, and succumb to the temptation of consuming more at the time when the salary raises commence. So, the programme is supposed to work in a setting where the values of variables are fixed, i.e. income schedules, relative prices etc. are given, and intervenes with the parameters that figure in agents' cognition when making decisions about diverting income into present and future consumption respectively.

This case suggests that some interventions are more appropriately conceptualized as interventions on parameters rather than variables. What is important about this is that the effectiveness of such interventions seems to hinge on the agent-specific malleability of these parameters, e.g. the extent to which, for a certain framing of a choice setting, agents are willing to commit to changes of parameters that will bear on their future savings contributions, i.e. the outcome variable that the intervention is supposed to help increase. This malleability, i.e. the *meta-parameter* governing the sensitivity of a

parameter to interventions, may importantly differ between agents, and the effectiveness of an intervention aiming to change such parameters may consequently differ importantly between individuals, and between populations.

In summary, as the cases outlined above make clear, it is not just differences in variables between individuals and populations that can induce differences in causal effects. In various ways, differences in parameters associated with paths that connect the treatment and outcome variables of interest, as well as the functional form of the relationships between variables, can induce such differences as well.

As is the case with differences in variables, it is important to keep in mind that learning about differences between populations in parameters/functional form does not imply that causal effects will be different, as these differences can be partly or fully mitigated by yet other differences. Similarly, learning about some similarities in parameters/functional form does not imply that causal effects will be the same or similar between populations, as there may be other, unknown differences that can still induce differences in effects between populations.

2.4.3 Basic Structure of Mechanisms

Finally, the third level at which causally relevant differences between populations can obtain concerns the basic structure of the causal mechanisms that govern the effects of interest. For instance, X can be causally relevant for Y in A but not in B, e.g. because there is no causal pathway connecting X and Y in B, or the mechanism is disrupted in some subpopulation of individuals in B.

More generally, differences at this level concern qualitative features of causal mechanisms. Several kinds of differences are important here, including whether there is some causal relationship between a pair of variables X and Y and what the direction of the causal arrow connecting these variables is. Differences in these dimensions can obtain between individuals, and between populations. Although some differences seem somewhat unlikely (e.g. reversals of arrows between two individuals), others seem to pose more realistic threats to successful extrapolation, such as when a social norm in A induces economic decision-makers to respond to the presence of a certain salient social feature Z, but the same is not true in B. Here Z is involved in the decision-making mechanisms in A but uninvolved in B.

Similar to the considerations offered above, differences at the level of the basic structure of mechanisms do not imply differences in causal effects, and similarities do not imply similarities in causal effects between populations. If there is a mediated causal pathway $X \rightarrow Z \rightarrow Y$ in A but one of the two relationships is severed in B, X can still be causally relevant for Y if there is another causal pathway from X to Y. Similarly, if $X \rightarrow Z \rightarrow Y$ in both populations, then the efficacy of X for Y can be curtailed if there is a counteracting pathway that has opposite causal relevance and can suppress changes in Y brought about by changes in X.

2.4.4 Causally Relevant Differences Summarized

The previous discussion offers a general framework for thinking about how problems of extrapolation are constituted, i.e. by causally relevant differences between individuals and populations at different levels of causal analysis: variables, parameters/functional form, and the basic structure of causal mechanisms. Differences can obtain at any or all of these levels, and to varying degrees, thus posing problems of different difficulty, for instance, because some differences at some levels (such as the basic structure of mechanisms, difficult to measure parameters, or latent variables) are epistemically less readily accessible than others.

What this analysis also makes clear is that some differences are more fundamental than others. Specifically, as I will discuss in more detail in subsequent chapters, some extant strategies for extrapolation explicitly attempt to account for causally relevant differences between populations. Yet, while these attempts can sometimes be successful, such strategies often consider only differences at the level of the distributions of variables, but not differences in parameters, functional form, and basic causal structure. It seems, however, that differences at these levels are more fundamental in the sense that unless one is confident that populations are similar at these levels, there is no point in trying to accommodate, and adjust for, differences at higher levels. For instance, if populations differ at the level of the basic structure of causal mechanisms, such as when there is some causal pathway from X to Y in A but not in B, then adjustment for differences at higher levels does not change or add to the conclusion that interventions on X will have no effects on Y in B. The converse is not true: if there are differences in the distribution of a moderating variable MO between A and B, then our conclusions about causal effects in B will still be responsive

to further information about similarities and differences in parameters and basic structure.

The framework outlined above is useful in several general ways: it allows us to distinguish different problems of extrapolation according to severity; it helps distinguish the epistemic challenges involved in learning about causally relevant differences and similarities between populations; and it is useful for assessing the scope of extant strategies for extrapolation in addressing different types of problems, including what assumptions they involve about similarities at each of the above levels.

Before this framework can be successfully put to use, however, some clarifications are in order: while variables and basic structural features (such as whether or not there is a causal relationship between two variables) seem like uncontroversial units of causal analysis, a critical reader may wonder about the nature of the parameters discussed here. More specifically, while it is reasonably clear what a variable is, and what it means for a causal relationship between X and Y to be present, absent, have different directions, or to have a specific functional form, my analysis may raise the question of what, after all, parameters are and whether parameters constitute a genuine level of causal analysis.

These questions are important because it may seem that parameters are merely a convenient summary, obtained by measurement, that accounts for what happens to an outcome Y if we induce a marginal (e.g. one unit) change in a causal parent X. As such, a parameter would be nothing more than a measure of a marginal causal effect. Specifically, saying that the parameter on the path from X to Y is such-and-such would not assert anything about whether there is a more fine-grained causal structure underlying these marginal causal effects, what this structure looks like, or whether the path between X and Y, as well as the parameter itself, is indeed a primitive unit of causal analysis.

This could be a challenge for the persuasiveness of my analysis, because it may seem that, in many cases, causally relevant differences that purportedly obtain at the level of parameters can in fact be attributed to, and explained in terms of, other differences at the level of variables or the basic structure of mechanisms.

For instance, let X be a job-interview training programme, let Z be a latent mediating variable capturing applicants' interview skills, and let Y be success in job interviews. At face value, the path from X to Z may have different parameters in different individuals,

capturing how the change in skills from exposure to the training programme differs between them. However, it may seem that these differences are not brute differences, and may instead have a deeper explanation, in the sense that they are attributable to differences that obtain at the level of a more fine-grained causal structure underpinning the relation between X and Z; most likely the cognitive mechanisms governing how individuals convert various inputs obtained in training classes into domain-general interview skills.

To further clarify this concern, it might be useful to distinguish between two notions of parameters: *causal* and *statistical.* Causal parameters would be those that are genuinely causal in nature and could (at least in principle) be intervened on directly, whereas statistical parameters would be mere representations of factual and counterfactual features that an underlying causal structure regularly gives rise to. As an example of a causal parameter, we might consider the size of a water valve's aperture, i.e. a parameter that corresponds closely to, or is identified with a physical quantity that is nomologically related to an outcome of interest, such as the flow of water, or the time it takes to fill a container placed under the valve's opening. Such parameters, in many cases, can be straightforward targets of actual or hypothetical interventions that would change the factual and counterfactual quantities that constitute the effects of interest.

An example for a statistical parameter could be students' learning speed, e.g. the marginal change in propositional knowledge held by a student with respect to a given exposure to training classes. This parameter seems better understood as convenient shorthand for lower-level features of the mechanism governing the production of individuals' propositional knowledge. Intervention here seems more difficult, as learning speed is not easily directly manipulated by intervention on a single variable or entity with variable properties (such as the water valve). This does not mean that learning speed is somehow immune to interventions, just that it is more difficult to intervene on directly and that successful interventions on learning speed may need to target variables and causal structure at those lower levels that the parameter supposedly represents, e.g. we could call for students' attention, allow them to exercise between tasks, set up disincentives for being inattentive, give them drugs to focus better, etc. While the effect of such interventions on the parameter in question will plausibly be transmitted or realized by lower-level causal arrangements that the parameter is supposed to represent, this does not mean that the parameter itself cannot be changed.

Rather, it just cannot be *directly* changed, since there is no single variable or entity with malleable properties to be directly intervened on (as in the water valve case).

While I think that there is an interesting difference between causal and statistical parameters, I do not think that the difference between the two matters importantly for most of the arguments developed in this thesis.

First, while it might be possible to re-describe differences at the level of statistical parameters in terms of differences at the level of variables and basic structure of the mechanisms governing the relations that the parameter is supposed to represent, this does not imply that thinking about differences in statistical parameters does not constitute a useful level of analysis. Thinking about causal mechanisms in social sciences will often proceed at levels of abstraction where putatively primitive causal relationships between variables can, in principle, be unpacked in terms of sub-mechanisms that constitute, or otherwise give rise to, these relations (see Craver 2007, 7 for similar points on neuroscience; see also Craver and Bechtel 2007). However, for many epistemic activities relevant to social scientists, higher levels of abstraction will often be adequate, or indeed more adequate than lower-level characterizations of mechanisms (e.g. in terms of neurophysiological mechanisms), to the epistemic aims involved in these activities (see also Little 1993).

Second, for related reasons, both the estimation of causal effects as well as their extrapolation usually proceed, often successfully, at higher levels of abstraction. Hence, it can still be useful to think of causally relevant differences at the level of parameters, even if, in some cases, parameters are not a genuine level of causal analysis and causally relevant differences at this level are ultimately attributable to lower-level differences.

Third, the concern about whether the levels of causal analysis outlined above are genuine also applies the other way around. For instance, one might argue that we can represent the presence or absence of causal relationships between X and Y already at the second level, e.g. by setting the parameter for such relationships to zero. Again, while this is possible, I prefer to resist such moves to accommodate apparent differences at some level by expressing them in terms of differences at other levels.

In general, my aim is to keep two issues distinct. First, there are quantitative issues concerning the magnitudes (including signs) of marginal causal effects; these seem best represented at the level of parameters. Second, there are qualitative issues concerning

the presence, absence, direction, and functional form of causal relationships; these seem best represented at the level of the basic structure of the mechanisms.[10]

So while I acknowledge that there might be interesting discussions about the nature of the levels invoked in my analysis, I will not go deeper into this issue, in part because doing so would likely raise all kinds of trouble, including issues of constitutive causal relevance, reduction, downward causation, mental causation and other controversial subjects (see e.g. Craver 2007; Craver and Bechtel 2007). For the remainder of the discussion, I will hence assume, mostly for pragmatic reasons, that the levels of analysis outlined here constitute a fruitful basis that helps elucidate problems of extrapolation and ways to address them.

2.5 Varieties of Extrapolation

As the preceding analysis makes clear, causally relevant differences can obtain in several different ways, and to different degrees.[11] This, by itself, already suggests that problems of extrapolation can vary significantly in severity. In what follows, I expand on this, as well as on several other important dimensions along which problems of extrapolation can differ. At the most general level, I will distinguish between *ontic* and *epistemic* dimensions of extrapolation. The *ontic* dimension concerns mind-independent or mind-invariant facts[12] about the nature of the causal makeup of the populations of interest and how the populations are related. In contrast, the *epistemic* dimension concerns issues regarding our knowledge of these facts, our ability to learn such facts, and the epistemic aims of the inferences we wish to support by reference to such facts. Let me begin with the ontic dimension.

2.5.1 The Kinds and Degrees of Causally Relevant Differences

As elaborated above, experimental and target populations can differ in various ways, and at different levels of causal analysis. They may differ at only one of the levels or at all of them simultaneously, and these differences can obtain in stronger or milder ways.

[10] Departing from this organization, in *Chapter 6* and *7* I will discuss issues of functional form and parameters together and target issues concerning the presence, absence, and direction of causal relationships separately. This is owed to how some strategies for extrapolation organize these issues.
[11] For a related discussion in computer science, see Subbaswamy et al. (2019).
[12] Of course, such facts (psychological ones) can be *about* minds, but they can still be intersubjectively accepted and hence *mind-invariant*.

To give an intuitive assessment of the scope of variation, consider what is possibly the mildest case where two populations A and B of equal size only differ in the distribution of one variable, say a moderating variable MO, and the difference in the distributions of MO is minimal in that all individuals in A and B have the same values of $MO = mo$, except for one individual i in either population, who has a value of $MO_i = mo' \neq mo$, but where that value is reasonably close to $MO = mo$. This would be a case of causally relevant differences that pose no important obstacle to any strategy for extrapolation. Indeed, if it were known that populations are related in this way, then even naïve extrapolation would presumably yield a reasonably accurate prediction of the causal effect of interest in the target.

At the other extreme, we can imagine cases where individuals differ at the lowest level of the basic structure of causal mechanisms in such a way that there is a causal pathway from X to Y for all individuals in A, but there is no such pathway for some individuals in B, a different pathway, involving different moderating and mediating variables for yet other individuals in B, and so forth. So X is causally relevant in A, not relevant for some in B, and relevant, but in potentially dramatically different ways, for yet other agents in B. A and B might additionally differ wildly at the levels of parameters, functional form, and distributions of variables.

As these considerations make clear, the kinds of causally relevant differences obtaining between populations can vary significantly in severity where, even before considering particular strategies for extrapolation, it seems clear that some of these differences are substantially more challenging than others.

Causally relevant differences are not the only facts about how populations are related that will bear on our ability to extrapolate successfully.

2.5.2 Distinctness

Another important dimension concerns the entities that populations are made up of and how these entities are spatiotemporally related. Recall the pre-analysis sketch of extrapolation offered at the beginning of this chapter: extrapolation is an inference starting from knowledge of a causal relationship in some population A to infer a causal conclusion about some distinct population B. As suggested, an important question

raised by this sketch is what exactly is meant by populations being *distinct*.[13] In addition to exhibiting relationships of similarity and difference at various levels, there are also important relationships between A and B concerning who or what they are composed of.

Generally, although I have treated A and B as stand-ins for populations, A and B can denote various units of analysis, including whole populations, individuals, places, or all of these at different times. This gives rise to various sorts of relationships between the systems to be extrapolated from and to.

For instance, A and B can be two individuals, populations, or settings, or be of different kinds respectively, e.g. where A is a population and B is an individual (such as in evidence-based medical diagnosis). A and B can also be temporally distinct, where, for instance, A and B can refer to the same individuals at different points in time. A and B, at the level of populations, may also overlap (even significantly). For instance, A can be a proper and small subset of B, such as when pilot studies are used before an intervention is rolled out in a whole population, where the pilot study recruits (potentially representative) samples from a superpopulation B.[14]

To offer a sense of the spectrum that is being covered here, consider the mildest case, where A and B are populations that differ only in one individual i, where $A \cup i = B$. At the other end of the spectrum there are populations that are completely distinct in any or all of the senses above, so $A \cap B = \{\emptyset\}$, where A and B might be temporally far apart, might be of different kinds, e.g. A is a population, B an individual (such as in individual-specific effectiveness prediction), etc. All of these features underpinning how experimental and target systems may be distinct from one another can sometimes covary strongly with different flavours and severities of causally relevant differences

[13] One might also wonder whether it is important to explicate what I mean by 'population'. I understand populations merely as collections of individuals situated in a specific context with causally relevant contextual features (e.g. of an environment, an institutional structure etc.) being 'attached to' individuals in virtue of their being situated in that context. I am open to more involved conceptions, e.g. where populations are, in part, defined by causal characteristics. However, for the purpose of targeting extrapolation in EBP it seems best to take a 'blank slate' view, where we start by identifying populations as just individuals being in a certain place/time, but without being able (before additional investigations, anyway) to say more on causally relevant characteristics at the group-level.

[14] Contra authors such as Shadish et al. (2002) I do not find such cases to be particularly interesting as they are arguably not typical in EBP. Here, it is more common to use evidence from populations A that are not usefully understood to have been sampled/selected from the target B. Moreover, unlike some (e.g. Dias et al. 2012; Stuart et al. 2018), I also do not find it helpful to think about A and B being sampled/selected from some general superpopulation C, such as 'school-aged children in Sub-Sahara Africa'. Doing so is often unhelpful because 1) C's characteristics are rarely known, 2) it is usually unclear what sampling/selection process gave rise to the distributions of causally relevant characteristics in A and B, and 3) thinking about the differences between A and B as a result of sampling/selection does not add much beyond recognizing that individuals in A and B are different in important ways.

outlined above, so they can sometimes figure as an intuitive proxy for how likely it is that populations are relevantly causally similar or different (cf. Campbell 1986). It is still important, however, to keep these issues distinct, as A and B being less or more distinct does not always, nor perhaps typically, speak reliably for whether they are relevantly similar or different.

2.5.3 The Nature of the Intervention

Another important way in which problems of extrapolation can differ is with respect to the nature of the intervention at issue, and with respect to how the actual and envisioned interventions in A and B are related.

A first important layer of analysis concerns general differences in interventions, not between A and B, but rather between interventions in general, even if they are the same in A and B. Interventions can target single variables, multiple variables, parameters, or causal structure. They can be point interventions that set the value of a variable $X = x$ at a certain point in time or interventions that aim to sustain $X = x$ for prolonged periods, shielding it from changes induced by (variation in) its causal parents.

Interventions can also differ between A and B, such as when the implementation quality of one and the same intervention (at some level of abstraction) relevantly differs between A and B, e.g. implementers in A are more skilled in setting X within some region ε around $X = x$, but implementers in B (for instance due to lack of training) might err more broadly around $X = x$, or even systematically on either side of $X = x$, thus inducing further causally relevant differences in the causal effects of one and the same *envisioned* intervention in A and B respectively.

Interventions can also differ more radically, such as in cases where an intervention in A targets a single variable X, but targets (knowingly or unknowingly) multiple variables when implemented in B.

Another important source of differences in interventions are so-called structure-altering interventions (see e.g. Steel 2008, ch.8), where one and the same initial intervention on X in A and B yields different downstream changes to the causal structure and parameters governing the effects in A and B respectively. This is essentially an extended version of the *Lucas critique* (Lucas 1976). The idea here is that interventions can sometimes meddle with features of the mechanisms connecting an

intervention and outcome variable (usually parameters). An extended version of this concern relevant for extrapolation is that such (unanticipated or unintended) changes to mechanisms effected by an intervention can also differ between populations, e.g. when an intervention I is structure-altering in A but not in B, or differently structure-altering in A and B.

This completes the overview of important differences between problems of extrapolation at the ontic level, i.e. the level pertaining to mind-independent and invariant facts about the experimental and target populations and how they relate, as well as the interventions to be implemented in these populations. These facts have important bearing on our ability to successfully extrapolate causal effects in a principled way. For instance, if populations are radically different at the level of the basic structure of causal mechanisms, or if interventions differ radically between populations, then, short of unrealistic cases where we have full knowledge of the causal makeup of both populations and how they relate, successful extrapolation will be precluded.

2.5.4 Epistemic Differences

As already suggested, there is also a second, *epistemic* layer involved in extrapolation, which concerns our knowledge of facts of similarity and differences between populations, our ability to acquire such knowledge, the aims we use such knowledge for, etc. In what follows I expand on several epistemic dimensions in which problems of extrapolation can differ importantly with respect to the nature of the extrapolative inference to be drawn, i.e. the mode of inference, the type of conclusion envisioned, the knowledge and evidence used to obtain this conclusion, etc. Here, extrapolation can differ in the following ways.

First, our background causal knowledge can differ importantly between cases. For instance, the intervention of interest might be well understood, including the particular causal pathways through which it is supposed to induce the envisioned changes in the outcome of interest, or we might possess pre-existing causal knowledge about relevant (and likely) similarities in the mechanisms governing the effects of interest. We might also possess more abstract knowledge that increases our confidence in such similarities. For instance, there can be cases where we have good reasons to believe that individuals and populations are relevantly similar in virtue of being members of some general type, as for instance in biomedical research where mechanisms can sometimes be justifiably

assumed to be similar between individuals and populations in virtue of common ancestry and heritability of physiological features that jointly give rise to causal homogeneity among individuals at many important levels of causal analysis (see e.g. Ankeny and Leonelli 2011). On the other hand, there will also be cases where such background causal knowledge is not available, such as for novel interventions that are not well understood, with no explicit and suitably supported causal theory of the mechanisms governing the effects of interest being available, and only local evidence of causal efficacy, i.e. evidence that some variable X is causally relevant for Y, but no evidence that elucidates features of the underlying mechanism governing this relation.

A second, related epistemic dimension concerns the accessibility of knowledge regarding causally relevant differences and similarities. This is important particularly in cases where our background causal knowledge is insufficient to support the assumption that populations are relevantly causally similar. In such cases, it needs to be explicitly supported that populations are sufficiently similar in ways that matter importantly for the effects of interest, e.g. at the different levels of causal analysis outlined in *Section 4*. In *Chapter 8*, I will expand in more detail on empirical strategies to generate such knowledge, and important difficulties encountered in doing so. For now, it is enough to note that while learning the values and distributions of variables is comparably easy, it is substantially more difficult to identify which variables are moderating and mediating the effects of interest, and even more challenging to estimate parameters and identifying functional form associations of such variables as well as identify other, qualitative features of the basic structure of mechanisms.

Third, extrapolation can also differ significantly with respect to the type of causal query to be answered by an extrapolation. Here we can distinguish between various different types of queries, including:

1) Will intervening on X have *some* effect in B if it does so in A?
2) Will intervening on X have a *similar* effect in B as in A?
3) Will the effect of an intervention on X in B be at least as large as in A?
4) What is the magnitude of the causal effect of an intervention in B if it is such-and-such in A?
5) What would an intervention I' in B need to look like to achieve the same effect in B as I in A?

6) What would an intervention I' in B need to look like to achieve some specific effect in B?

7) What is the effect of a never-before-experienced intervention I' in B, given the effect of some other, related intervention I in A?

8) What are the characteristics of a population B where I would have the same or similar effect as in A?

More generally, we can distinguish between queries concerning qualitative and quantitative effects, between queries concerning the achievement of a specific level of an outcome, or rather a specific change in an outcome (both in absolute and relative terms), whether the effect should obtain at a particular point in time or be sustained over longer periods, etc. Queries may also differ importantly concerning the envisioned fidelity of the inference to be drawn. They might ask for anything from a broad assessment of the qualitative effects in the target, the quantitative magnitude of the effect, to details of the distribution of causal effects in the target, and all of these with varying degrees of envisioned accuracy and precision.

Even without going into more detail at this stage, it seems clear that some of these queries will be considerably more difficult to answer than others, primarily because the assumptions required to answer them, and the evidence required to support such assumptions, will be more extensive as we move from the top to the bottom of the list.

A fourth dimension concerns differences in the mode of inference used to extrapolate. For instance, extrapolation can be inductive, where the level of support required for a conclusion might differ radically depending on the envisioned fidelity of the inference. It may also be deductive, where confidence in the extrapolative conclusion is supposed to be promoted not only by the inductive support for the premises involved, but also by the deductive rigour of the inference being used.

Extrapolative inference may also differ in its envisioned general type of conclusion, e.g. whether a dichotomous conclusion is aimed for (will the effect of I in B be such-and-such or not), or whether the conclusion should be probabilistic in nature (e.g. the effect of I in B is more likely to be greater than ΔY than smaller, or the probability of the effect to be such-and-such is p, etc.).

Finally, extrapolation may differ importantly with respect to the assumptions necessary to licence the conclusion of interest with the envisioned confidence. Such

assumptions will typically concern the relationships that obtain between populations at the ontic level outlined above, e.g. whether populations are sufficiently similar at the level of the structure of mechanisms, as well as the functional form relationships, parameters, and variables relevant to the effects of interest.

Important differences here will also obtain with respect to the nature and amount of evidence used to underwrite these assumptions, and the degree of support it affords. Here, the type of evidence can be manifold. It may include observational, quasi-experimental, and experimental evidence; it can be evidence from the experimental, target, or indeed yet other populations or settings; it might be qualitative or quantitative, obtained by different methods, including empirical, quasi-empirical or non-empirical methods such as computer simulations, etc.

These different dimensions, ontic and epistemic, make clear that extrapolation is a highly diverse set of epistemic activities: different problems of extrapolation can differ radically in the severity of the challenges that they pose and different types of extrapolative inference can differ importantly in what they aim to achieve.

Despite these differences, there nevertheless seems to be a level of analysis at which extrapolations of these different kinds are alike. In what follows, I propose a general analysis of extrapolation that will form the basis for the critical and constructive contributions to be developed in subsequent chapters.

2.5.5 A (More) Comprehensive View of Extrapolation

The insights provided above help us put together a more comprehensive view of extrapolation. Recall Steel's characterization of extrapolation that I started from at the beginning of this chapter:

> "[…] one begins with some knowledge of a causal relationship in one population, and endeavors to *reliably* draw a conclusion concerning the relationship in a *distinct* population." (2008, 3)

We can now unpack this further: at the most general level, extrapolation is an inference ***I***, using evidence ***E*** obtained from system ***A***, that aims to infer a conclusion ***C*** about a target system ***B***, with the help of assumptions ***P***, background knowledge ***K***, and supplementary evidence ***S***, pertaining to the relation ***R*** between ***A*** and ***B***.

So far, the discussion has centred on the evidence ***E*** to be extrapolated from, the nature of the systems ***A*** and ***B***, including their relation ***R***, as well as the mode of inference ***I***, and the envisioned type of conclusion ***C***. This has already made clear how problems of extrapolations can differ radically in the kinds of challenges that they pose under variations in these dimensions, as well as how the aims of an extrapolation may differ significantly between cases. For instance, at one end of the spectrum, we may find extrapolations that aim to predict the qualitative effects of an intervention from a population A to a superpopulation B, where A is a large subset of B, A is sampled randomly from B, B is believed to be causally homogenous (supported by background theory and evidence), the intervention in A and the envisioned future intervention in B are identical under a detailed description, and the envisioned fidelity of the inference is low (e.g. it aims to predict whether there will be a non-zero average effect in B). At the other end of the spectrum we may find extrapolations where A and B are disjoint sets of individuals in different locations and times, are further justifiably believed to differ substantially along several causally relevant dimensions, the mechanisms governing the effects of interest are justifiably believed to be malleable, and the aim is to predict a precise causal effect magnitude of an intervention I' that is substantially different from I in A. Clearly, the background knowledge, assumptions and supplementary evidence required to licence the second kind of extrapolation will be significantly more extensive than in the first case, and there are good reasons to believe that this type of extrapolation is unlikely to be successful by any standard of success.

2.6 Outlook

The analysis provided above offers a useful background on what extrapolation is, the reasons why extrapolation is challenging, and the epistemic challenges involved in overcoming different types of problems of extrapolation that differ in the nature of the populations ***A*** and ***B***, including their relation ***R***, as well as the mode of inference ***I*** and the envisioned type of conclusion ***C***. What the discussion has also yielded is that extrapolation is a highly heterogeneous collection of problems and inferential activities that, beyond conforming to a general template, can exhibit important differences that need to be taken into account when devising and evaluating strategies for extrapolation.

This discussion hence forms a useful basis for the following chapters, where I will turn to the general questions of how to achieve successful extrapolation, what

assumptions ***P*** are needed to do so, and what supplementary evidence ***S*** and background knowledge ***K*** is needed to support these assumptions. *Chapter 3* will make the first steps here, by laying out the general assumptions that are shared by existing strategies for extrapolation, as well as characterizing strictures on what counts as *successful* extrapolation.

References

Ankeny, R. A., and S. Leonelli. (2011). "What's so special about model organisms", *Studies in History and Philosophy of Science Part A*, 42(2): 313-23.

Baron, R. M., and Kenny, D. A. (1986). "The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations". *Journal of Personality and Social Psychology*, 51: 1173-82.

Campbell, D. T. (1986). "Relabeling internal and external validity for applied social scientists". In W. M. K. Trochim (ed.), *Advances in quasi-experimental design and analysis*, pp. 67–77, San Francisco, CA: Jossey-Bass.

Cartwright, N. D. (2012). "Will this Policy work for You? Predicting Effectiveness Better: How Philosophy Helps", *Philosophy of Science*, 79 (5): 973-989.

— **(2013).** "Evidence, Argument and Prediction". In: EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings 2, Springer: Geneva.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press.

Craver, C. F., and Bechtel, W. (2007). "Top-down causation without top-down causes", *Biology and Philosophy*, 22: 547–63.

Dias, S., A. J. Sutton, N. J. Welton, and A. E. Ades. (2012). "Heterogeneity: Subgroups, Meta-Regression, Bias and Bias-Adjustment", NICE DSU Technical Support Document No 3. London: National Institute for Health and Clinical Excellence (NICE).

Elwert, F. (2013). "Graphical Causal Models". In: S.L. Morgan (ed.), *Handbook of Causal Analysis for Social Research*, pp. 245-73, Dordrecht: Springer.

Fuller, J. (2018). "The Confounding Question of Confounding Causes in Randomized Trials", *The British Journal for the Philosophy of Science*, axx015.

Holland, P. (1986). "Statistics and Causal Inference", *Journal of the American Statistical Association,* 81(396): 945-60.

Imai, K., D. Tingley, and T.Yamamoto. (2013). "Experimental designs for identifying causal mechanisms", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176: 5–51.

Keane, M. P. (2010). "A structural perspective on the experimentalist school", *Journal of Economic Perspectives*, 24(2): 47–58.

Khosrowi, D. (2019). "Trade-offs Between Epistemic and Moral Values in Evidence-Based Policy". *Economics and Philosophy*, 35(1), 49-78.

Kraemer, H. C., E. Stice, A. Kazdin, D. Offord, and D. Kupfer. (2001). "How do risk factors work together? Moderators, mediators, independent, overlapping, and proxy risk factors", *American Journal of Psychiatry*, 158: 848-856.

Kraemer H. C., G. T. Wilson, C. G. Fairburn, and W. S. Agras. (2002). "Mediators and moderators of treatment effects in randomized clinical trials". *Archives of General Psychiatry*, 59: 877-83.

Little, D. (1993). "On the Scope and Limits of Generalizations in the Social Sciences", *Synthese*, 97(2): 183-207.

Lucas, R. (1967). "Econometric Policy Evaluation: A Critique". In: K. Brunner, A. Meltzer (ed.), *The Phillips Curve and Labor Markets*, pp. 19-46, Carnegie-Rochester Conference Series on Public Policy, 1. New York: Elsevier.

Marsh, H. W., K. T. Hau, Z. Wen, B. Nagengast, and A. J. S. Morin. (2011). "Moderation". In: Little, T. D. (ed.), *Oxford Handbook of Quantitative Methods*. New York: Oxford University Press.

Muller, S. M. (2015). "Interaction and external validity: obstacles to the policy relevance of randomized evaluations", *World Bank Economic Review*, 29(1): 217-225.

Muller, D., C. M. Judd, and V. Y. Yzerbyt. (2005). "When moderation is mediated and mediation is moderated", *Journal of personality and social psychology*, 89(6): 852-63.

Rubin, D. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies", *Journal of Education Psychology*, 66: 688–701.

Shadish, W. R., T. D. Cook, and D. T. Campbell. (2002), *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford: Oxford University Press.

— **(2010)**. "A New Approach to Argument by Analogy: Extrapolation and Chain Graphs". *Philosophy of Science*, 77(5): 1058-69.

— **(2013)**. "Mechanisms and Extrapolation in the Abortion-Crime Controversy". In: Chao, Hsiang-Ke; et al. (eds.). *Mechanism and Causality in Biology and Economics*. pp. 185–206. Berlin/Heidelberg: Springer Science & Business Media.

Stuart, E. A., B. Ackerman, and D. Westreich. (2018). "Generalizability of Randomized Trial Results to Target Populations: Design and Analysis Possibilities", *Research on Social Work Practice*, 28(5): 532–37.

Subbaswamy, A, P. Schulam, and S. Saria. (2019). "Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport", *Proceedings of Machine Learning Research*, 89: 3118-27.

Sundell, K., K. Hansson, C. S. Löfholm, T. Olsson, L.-H. Gustle, and C. Kadesjö. (2008). "The transportability of multisystemic therapy to Sweden: Short-term results from a randomized trial of conduct-disordered youths", *Journal of Family Psychology*, 22(4): 550–60.

Thaler, R., and S. Benartzi. (2004). "Save More Tomorrow : Using Behavioural Economics to Increase Employee Saving", *Journal of Political Economy*, 112(1): 164-87.

Varadhan, R., and J. D. Seeger. (2013). **"**Estimation and reporting of heterogeneity of treatment effects. In: Velentgas, P., Dreyer, N. A., Nourjah, P., Smith, S. R. and Torchia,M. M. (eds.): *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide,* pp. 35–44. Rockville, MD: Agency for Healthcare Research and Quality.

Vivalt, E. (2019). "How Much Can We Generalize from Impact Evaluations?". Unpublished manuscript, ANU, Canberra. Retrieved from: http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf (accessed February 28, 2019)

Weinberg, C. (2007). "Can DAGs Clarify Effect Modification?", *Epidemiology*, 18(5): 569-72.

CHAPTER 3

Assumptions, Ideals, and Strictures

3.1 Introduction

As argued in *Chapter 1*, much of the promise of EBP initiatives hinges on our ability to successfully extrapolate to novel policy targets. Without this ability, the value and usefulness of effectiveness evidence would be significantly constrained as it would always only speak for the effectiveness of interventions where they have already been implemented. Clearly, for EBP to be successful we must be able to make some inferential leap from the available evidence to a conclusion about policy effects in novel targets that goes beyond what we already know. But this leap must also be justified, and the justification required for making it must not be overly demanding to come by. In the critical chapters to follow, I consider different strategies for extrapolation and argue that they fall prey to important problems when it comes to justifying the inferences that they enable. Some of these problems will be unique to each of the strategies on offer, but others are more general in nature.

In this chapter, I focus on the general problems. In doing so I begin, in *Section 2*, by spelling out what I consider to be important underlying assumptions that different strategies for extrapolation share; assumptions concerning how likely it is that populations are relevantly different and how causally relevant similarities and differences between populations bear on similarities and differences in causal effects. These assumptions have important bearing on several issues: 1) whether we should assume, by default, that populations are relevantly similar or different, 2) what effects we are justified to expect in the target when populations are relevantly similar, 3) what we can learn about likely differences in causal effects from investigating causally relevant similarities and differences between populations, and 4) whether we can still successfully predict causal effects in a target, despite causally relevant differences. I argue that the general assumptions made by existing strategies are largely compelling and allow us to sketch out a *guiding ideal* for extrapolation. This guiding ideal clarifies that, at least in principle, successful extrapolation is possible for a wide range of extrapolative queries and problems of extrapolation.

While this should make us optimistic that extrapolation, in general, is not an insurmountable problem, in *Section 3* I discuss some important strictures that compelling strategies for extrapolation must meet. Specifically, as we move from abstract general assumptions concerning the conditions under which extrapolation can, in principle, be achieved towards the practical challenges involved in overcoming concrete real-world problems of extrapolation, there are two important challenges that must be met. The first is that extrapolation strategies must not be overly epistemically demanding, i.e. they should not require supplementary evidence and knowledge that is impossible or otherwise extremely difficult (or costly) to obtain. The second challenge is that they must evade the *extrapolator's circle* (Steel 2008): the information about the target system required for an extrapolation must not be so extensive that we can identify the causal effect of interest based on that information alone.

I argue that these two strictures give rise to an important tension: on the one hand, the guiding ideal to be outlined below suggests that, in principle, and with enough information, we can correctly predict any causal effect in the target for a large class of extrapolation problems. The two strictures, however, make clear that there are important limits to 1) whether this is feasible in practice, and 2) whether, even if it is feasible, extrapolation can still serve its main function of being *ampliative*, i.e. offering *added epistemic value* in the sense that it provides sufficiently reliable information about causal effects in the target that goes beyond what we must already know about the target in order to reach a conclusion.

The main result of this discussion will be a more refined analysis of extrapolation that includes strictures on when extrapolation is *successful*. This refined analysis will figure as a general benchmark that will help with critically evaluating existing strategies for extrapolation in subsequent chapters.

3.2 No Extrapolation Without Assumptions

Successful extrapolation involves learning about the relationship ***R*** between experimental and target population, i.e. causally relevant similarities and differences, and how this relationship bears on the causal effects to be expected in the target. In some form or another, this will require that we make assumptions, i.e. assumptions about the likely form of ***R*** that we will encounter in real-world problems of extrapolation, as well as assumptions about how ***R*** bears on similarities and differences

in causal effects between populations. This is the main epistemic target of extrapolation in EBP: inferring whether causal effects in the target will be similar to those in the experimental population, or, if they are not, predicting how the effect in the target will differ from that in the experimental population.

In line with the discussion in *Chapter 2*, it is clear that extrapolation can come in different forms, and involve different aims. It is no surprise, then, that existing strategies for extrapolation also differ in the kinds of extrapolations that they enable. Some strategies, such as Steel's (2008), aim at specifying the conditions under which causal effects will have the same sign in the target as in the experimental population. Others, such as Cartwright's (2013b), aim at spelling out a range of conditions, including those ensuring that there is some non-zero effect for some individuals in the target and those ensuring that effect magnitudes will be (approximately) the same in the target as in the experimental population. Finally, there are also strategies, such as those offered by Hotz et al. (2005), Muller (2014; 2015), and Bareinboim and Pearl (2012) that aim to predict causal effect magnitudes in the target, including under conditions where causally relevant differences obtain between populations.

Despite differences in the kinds of inferences enabled, these strategies rest on related assumptions about how differences in the causal makeup of populations bear on differences in causal effects. In what follows, I reconstruct some of these assumptions at a general level. This will help sketch out a guiding ideal for how successful extrapolation of different kinds can be achieved in principle.

3.2.1 Populations Are More Likely to Be Different Than Similar

The first assumption concerns the general issue of whether we should, by default, assume that populations are causally relevantly different or similar. As the previous discussion suggests, causally relevant similarities and differences between populations can obtain in various respects and to various degrees. Although there is a wide spectrum of ways in which and degrees to which populations can differ importantly, it is at least not obvious whether we should by default assume that populations are relevantly similar or rather different.

Existing strategies for extrapolation often remain silent on this important issue (but see Cartwright 2013a; Steel 2008). Nevertheless, it seems plausible to think that they

are motivated by a substantive belief that populations in real-world extrapolation settings are often likely to exhibit causally relevant differences, or at least that, in light of potentially significant risks of error, it is prudent to assume that they do. Hence, these strategies seem to suggest that our default assumption (i.e. the assumption made before any details about the experimental and target population, including any details about their relation ***R***, are learnt) should be that populations are different, rather than similar and that this assumption should be entertained until we can offer compelling reasons to think otherwise, e.g. evidence of causally relevant similarities (cf. Cartwright 2013a, 107; Steel 2008 Ch.5; Fuller 2019; but see Petticrew and Chalmers 2011; Guyatt et al. 2008 for resistance in the context of evidence-based medicine).

There are several reasons to motivate this assumption. First, empirical evidence from prominent meta-analyses suggests that causal effects in social science settings, and in the fields targeted by EBP initiatives specifically, often differ significantly between individuals and between populations (see e.g. Vivalt 2019). While this evidence may be over- or underreporting actual heterogeneity[1], it at least gives us a clear indication that substantial heterogeneity in causal effects obtains in the kinds of settings that have been addressed by EBP so far. The ubiquity and strength of effect heterogeneity also suggest that underlying causally relevant differences, e.g. at the different levels distinguished in *Chapter 2*, may be responsible for this observed heterogeneity in effects (rather than, say, random measurement error).

Second, there are several plausible reasons for why we find such heterogeneity, and hence reasons for why we may expect it in future instances as well. For instance, the causal mechanisms governing the phenomena and effects of interest in social science contexts tend to be *less stable* than in other fields (Little 1993; Dahabreh 2018), e.g. in contrast to Evidence-Based Medicine and epidemiology where physiological mechanisms are often justifiably believed to be stable over time and homogeneous within populations (see e.g. Cartwright 2009, 8; Steel 2004, 58).

Moreover, in social science settings causal mechanisms are also more likely to differ importantly between individuals, particularly in cases where agents' psychological features and cognitive processes are involved in governing their behavioural response to interventions. For instance, students' response to educational interventions such as new

[1] Say, for instance, because there is selection bias involved in what kinds of populations are studied in the first place.

curricula may differ substantially as a function of domain-general but individual-specific features such as cognitive ability, or more specific features of their psychological makeup, which may bear importantly on their ability and willingness to convert new material into outputs relevant for assessment of learning. Similarly, economic agents' response to policy interventions often differs substantially between individuals, as their behaviours are determined to a significant degree by features that are specific to individuals and can vary importantly between them, such as preferences, utility functions, budget constraints, etc. In light of this, it seems that there are at least some reasons to think that it is prudent to assume that individuals, as well as whole populations, are likely to exhibit causally relevant differences, unless we can offer reasons to think otherwise.

This might not seem compelling to everyone, however. One important concern that might be raised is that assuming by default that populations are relevantly different can pose unnecessary obstacles to using effectiveness evidence for informing policy. A pertinent way to frame this objection would be to point out that we might sometimes have strong and compelling intuitions that two populations will be sufficiently similar to licence extrapolation, where these intuitions could be grounded in overt similarities, such as individuals or populations being members of the same general type, e.g. students in private schools in the south of England, and that type-membership makes it seem likely that individuals will respond similarly to one and the same intervention (see e.g. van Eersel et al. 2019 for an approach that draws on type-level similarities). However, as the objection goes, despite strong intuitions that populations are sufficiently similar to licence some form of naïve extrapolation, it might often be extremely difficult to demonstrate (empirically or otherwise) that they are relevantly *causally* similar. Type-membership does not guarantee causal homogeneity, unless types are defined along the right causal features, which cannot always be assumed. In such cases, assuming by default that populations, despite being members of some common type, are relevantly different and putting the burden of proof on those who wish to suggest otherwise, may substantially complicate the use of effectiveness evidence in EBP to an extent where we might consider this assumption to be too costly to be warranted.

This is a serious objection. Proposing strictures on how to use effectiveness evidence to make inferences about new populations should not make the use of such evidence

overly burdensome, or, indeed, preclude the use of such evidence in most cases. Ideally, we would handle clear-cut cases (if they exist) differently, of course. We could, for instance, shift the burden of proof or lower our standards of evidence as to what reasons count as compelling enough to believe that populations are relevantly similar. For instance, if an intervention type A has been learnt to be effective across many different settings P, Q, R, S, despite these settings being overtly highly dissimilar, and the same intervention type is envisioned to be implemented in a novel context T that is, as judged by the same overt features, highly similar to at least some of these settings, then we may have good inductive reasons to tip the balance in favour of believing, without further scrutiny of the causal makeup of these populations, that T will be sufficiently similar to P, Q, R, S, and hence that the effects of A in T will be similar to those in P, Q, R, S. The reason here could be, for instance, that A's effectiveness seems to be robust over (overt) differences in populations, so there are at least some reasons to believe that A might be similarly effective in a target population that does not stray too far outside the spectrum of the features exhibited by P, Q, R, S.

What is more, the non-epistemic consequences of error (i.e. inductive and epistemic risk, see Douglas 2009; Biddle 2013) can also differ importantly across extrapolation scenarios. Here, it seems plausible to think that in low-stakes scenarios, when the possible consequences of error are relatively mild, we may justifiably be more generous and allow ourselves to err on the side of mispredicting causal effects. This would be in contrast to being overly demanding when it comes to demonstrating that populations are sufficiently similar to licence extrapolation at all, potentially on pain of incurring substantial costs in terms of the foregone benefits that could have been achieved by successful intervention in the target.

While such contextual information can clearly make important differences to our assessment of what is required to licence an extrapolation, I maintain that the default assumption, before any contextual details are learned, should be that populations are different rather than similar. This is not as strong an assumption as it might initially seem. What is assumed here is just that it is more likely than not that causal effects will differ between settings, and potentially just mildly so. This, I think, is rather uncontroversial. Of course, differences might be minimal, but if we have strong reasons to believe that this is so, we may, on some occasions, use this as an overriding reason to shift the burden of proof. My point is merely that unless *some* such reasons can be

given, the default assumption should be that populations are different. If such reasons can be given, and giving them is easy, then all the better, because we have been appropriately cautious in considering the possibility that populations might be importantly dissimilar, and still did not need to do a lot of work in underwriting the assumption that they are not.

This seems superior to the alternative. If we are too permissive when allowing evidence to inform and justify policy action without supporting that populations are relevantly similar to licence extrapolation and subsequent intervention, then we run the risk of implementing interventions that will fail to be effective. It often seems prudent to be risk-averse, especially for large and costly interventions (including in terms of harms), such as universal basic income, minimum wage policies, and many development interventions, and when resources are limited, as they often are.

Moreover, it seems that being risk-averse when it comes to extrapolation is coherent with overt risk attitudes that motivate EBP approaches in the first place, i.e. rather than believing that an intervention will be effective in a target based on intuition, hope, armchair theorizing, or other controversial sources of justification, EBP methodology suggests that we need to have high-quality empirical evidence that the interventions of interest can be effective. It seems coherent with this practice to place more weight on the risk of implementing an intervention that will turn out to be ineffective or harmful in the target than on the risk of failing to implement an intervention that would have been effective. And there are further reasons to support this view.

First, it seems that there are generally more possibilities for two populations to differ in ways that diminish rather than increase the effectiveness of policy interventions.[2] For instance, if a target population differs from an experimental population at the level of the basic structure of causal mechanisms, this seems more likely to curtail the effectiveness of the intervention in the target rather than enhance it, because differences at this level will often mean that a causal pathway that governs the effects of interest in the experimental population will be disrupted in the target (rather than, say, substituted by yet other causal pathways that allow the intervention to be effective still).

[2] It might seem that I am committing to an asymmetry here: anytime an intervention is effective in A, but less so in B, then we could just as well say that differences between A and B make the intervention more effective in A, rather than less effective in B. There is no asymmetry, however, when experimental populations are selected according to features that make it likely that interventions are effective there.

Second, interventions are often tested in populations where there are at least some reasons to believe that they will be effective. But interventions are not necessarily deployed in such settings after their effectiveness has been demonstrated somewhere. This could be understood as a worry about selection bias: we prevalently select those populations into trials that exhibit features that are already (and perhaps more often than not correctly) believed to promote the effectiveness of the intervention in question.[3] Moreover, interventions that are candidates for extrapolation may also be biased towards those that have been tested in settings where they are more effective than they would be in others, due to the settings' exhibiting features that are favourable to the effectiveness of the intervention. Target settings for deployment might not exhibit these features, however, as they might not be selected on the same criteria, and indeed might be selected on other criteria that correlate with unfavourable realizations of features relevant to the effect of interest, such as being selected because the outcome of interest is in great need of improvement.

Third, even if policymakers are careful to select target populations on grounds of relevant similarities, and such selection is in fact successful, some causally relevant differences are still bound to obtain because studies usually involve idealized versions of interventions. Trial conductors, sponsors, and implementers typically have some (imperfect) understanding of what will make an intervention more or less effective and they will often be strongly incentivized to make the intervention a success. The same incentives might not apply when interventions are implemented in distinct settings, nor should it be taken for granted that implementers there will have a similarly sophisticated understanding of how to best deliver the intervention, or be similarly skilled in doing so (see also Muller 2015). So, it seems more likely that interventions, as actually implemented in distinct targets, will exhibit features that bear unfavourably rather than favourably on their effects.

Hence, the assumption that causally relevant differences between populations are likely to obtain seems well supported, and strategies for extrapolation hence have an important function to perform: if it is almost always important to explicitly demonstrate that populations are sufficiently similar, or otherwise take into account their differences, then almost all real-world use of effectiveness evidence will require conscientious extrapolation efforts.

[3] Further, issues surrounding publication bias would only aggravate this concern, see e.g. Ioannidis (2005).

Let me proceed to discuss two further assumptions, which I take to be central to proposing strategies for overcoming problems of extrapolation. They concern details of the relation ***R*** that obtains between experimental and target populations, i.e. causally relevant similarities and differences at different levels and to different degrees, and how these details bear on differences and similarities in causal effects.

3.2.2 Same Causes Imply Same Effects

The first assumption is that any two individuals or populations who are identical with respect to all causally relevant features that bear on the effect of interest, and who experience the same intervention (at some arbitrarily fine-grained level of description), will experience, at least in expectation, the same causal effects.[4] I say in expectation, because some variables involved in the mechanisms that govern the effects of interest might be random variables, where individual and aggregate causal effects can consequently differ over a range of values depending on how these variables are realized at the time when the effects of interest are produced. In expectation, however, causal effects should be randomly distributed around a mean with some variance that is induced by the random variables involved in their production. Hence, in expectation, and relevant distributional assumptions given, the expectation of the causal effect will be just the mean of that effect over repeated realizations.

This assumption is important because it ensures that if we learn that experimental and target populations do not differ in any causally relevant respects, and our search for such differences was, in fact, exhaustive, then we are justified to believe that the effect of interest in the target will be (approximately) the same as that in the experiment.

Importantly, this assumption does not imply that the more similar two populations are in causally relevant respects, the more similar the causal effects experienced by these populations will be (as, for instance, Hume ([1739] 1975) suggests). Such a 'convergence theorem' would require a host of additional, and substantially stronger assumptions. The reasons for this are manifold, but one extreme example would be a case where the magnitude of a causal effect depends non-monotonically on the value of a moderating variable MO, where the causal effect is a negative function of MO for values $MO < mo$, but jumps to a large value at $MO = mo$, so even minor differences in

[4] See Frigg and Votsis (2011) on historical mentions of this assumption, including by Hume ([1739] 1975).

MO imply significant differences in causal effects. Here, for a large region of MO, the causal effect would depend negatively on MO and effects would be more dissimilar the more similar the populations are with respect to MO. This is despite the fact that it would still hold true that causal effects would be identical if MO were identical between populations.

So believing that, if populations are not identical but highly similar beyond some arbitrary threshold of similarity, then the causal effect of interest is also likely to be similar, will require further assumptions, such as that minor differences between populations in, say, the distribution of some moderating variable, do not induce major differences in causal effects, or that causal effects are a monotonic function of moderating variables. Such assumptions might sometimes be justified, perhaps even often, but they should generally not be taken to hold as assumptions of convenience and without any support provided by empirical demonstration or sufficiently strong background theory and investigations of plausibility.

3.2.3 Differences in Effects Imply Differences in Causes

The second important assumption reinforces this relationship between differences in causally relevant features that figure in ***R*** and differences in causal effects. It says that any true, systematic, and significant difference in causal effects between individuals or between populations is brought about by some causally relevant difference in ***R***, e.g. a difference at any of the three levels identified in *Chapter 2*.[5] Again, this abstracts away from differences in causal effects that are induced by random variables (hence the emphasis on systematic), or measurement error (hence the emphasis on true).

This assumption is important because it makes clear that if 1) we learn that two populations are similar in potentially many putatively causally relevant respects, 2) we predict causal effects to be broadly the same on this basis, and 3), despite our best efforts, the causal effect in the target turns out to deviate substantially from our prediction, then there is an explanation for this deviation: We must have missed some causally relevant difference to which the observed difference in effects can be attributed. This makes clear that causal effects do not just differ randomly between populations, and hence secures the ideal that if we do everything right, and learn about

[5] See e.g. Russell ([1927] 1992, 255) for a mention of this assumption.

all potentially causally relevant features, and these features turn out to be the same, then we are justified in believing that causal effects will be the same.

Importantly, these two assumptions do not imply the converse relation, i.e. that individuals who experience the same (or similar) causal effects will exhibit causally relevant similarities (or even be identical). One and the same causal effect can be multiply realized by various different underlying arrangements of causal structure, functional form, parameters, and variables. At best, homogeneous causal effects between many individuals and between many populations could be taken to suggest that causally relevant similarities *might* obtain at some or all of these levels, but the latter cannot be unambiguously inferred from the former.

3.2.4 A Guiding Ideal

The assumptions outlined above suggest two things: first, real-world problems of extrapolation are likely to involve important obstacles that need to be explicitly addressed (rather than assumed away). Second, these obstacles are not in principle insurmountable. Specifically, the second and third assumption together licence an important corollary: if same causes imply same effects, and differences in effects imply differences in causes, then, either we learn populations to be similar in all relevant respects, in which case we can extrapolate straightforwardly, or we learn populations to be different. In that case, any true, systematic difference in causal effects between individuals or between populations can, at least in principle, and at least up to some threshold of accuracy, still be predicted using knowledge about causally relevant similarities and differences and how they bear on similarities and differences in the effects to be extrapolated.

To be sure, successful prediction at some arbitrarily high level of accuracy will often be impossible, e.g. when outcomes are produced probabilistically. But at least in expectation, i.e. averaged over repeated attempts, accurate prediction within some margin of error should be possible when all causally relevant differences are known, and how they bear on differences in causal effects in the experimental and target populations is correctly accounted for, including how interactions among such differences bear on these effects (see e.g. Fuller 2018).

We are now in a position to articulate an important *guiding ideal* for overcoming problems of extrapolation: if we knew all the causally relevant features that bear on a causal effect of interest, acquired all the information pertaining to how these features bear on that effect, and also learnt how these features are realized in the experimental and target populations, then we could successfully predict (within some margin of error) causal effects in the target population in a wide range of cases.

This is indeed the view that many proponents of existing strategies for extrapolation seem to take. Bracketing epistemic considerations about how to acquire and use information about causally relevant similarities and differences, they tell us that, in principle, extrapolation of different kinds can be successful, and they subsequently aim to spell out the abstract conditions under which this is the case. As Marcellesi puts it most confidently, having clarified the conditions under which extrapolation is feasible in principle means that "[…] the problem [of extrapolation] has been solved" (2015, 1309).

This makes the importance of the assumptions underlying the guiding ideal clear: if they did not hold then there could be problems of extrapolation that are in principle impossible to overcome by considering information about causally relevant differences and similarities. Moreover, there could be heterogeneity in causal effects that could not be attributed to and explained in terms of differences at any of the three levels outlined in *Chapter 2*. If such heterogeneity would remain unexplained not just in practice, but also in principle (and sufficiently frequently), this would seem to pose a serious problem for attempts to offer strategies for extrapolation. Conversely, attempts to offer such strategies need to entertain these assumptions in some form. It is hence not surprising that the *guiding ideal*, including its assumptions, in some form or another, underpins many and even appears explicitly in some of the strategies for extrapolation to be discussed.

Having such a guiding ideal in place is surely helpful, as it can help us to orientate ourselves in the epistemically less-than-perfect world we inhabit, i.e. a world where we do not usually know what causally relevant features bear importantly on the effects of interest, or how these features are realized in the experimental and target populations. What the guiding ideal suggests, in the first instance, is that if we possess or acquire the right kind of knowledge about the experimental and target populations, and use this

knowledge in the right way, then we can overcome a large class of problems of extrapolation.

Despite this, there are also reasons to remain sceptical about the usefulness of this ideal. The concern is not that it is misleading. At least for now, I will not take issue with whether the conditions under which extrapolation can succeed outlined by existing strategies are adequate. Rather, my main concerns are epistemic ones: even if existing strategies for extrapolation get the conditions right under which we can, in principle, correctly predict causal effects in a target, this does not imply that they are helpful for successfully overcoming concrete real-world problems of extrapolation. There is still a significant *epistemic* problem to be solved. Let me elaborate this concern more fully.

3.3 Two Strictures on Extrapolation

Despite the positive outlook afforded by the guiding ideal, two further important challenges remain in the way of successfully overcoming problems of extrapolation. These challenges arise not from whether problems of extrapolation can be overcome in the abstract, but from how much supplementary evidence ***S*** and background knowledge ***K*** is needed to overcome them in practice.

To elucidate these challenges, it is useful to once again draw a distinction between *ontic* and *epistemic* levels of extrapolation, where the ontic level concerns the kinds of entities, their features, and relationships that are involved in producing the causal effects we are interested in, i.e. experimental and target populations, the individuals that constitute them, their features that are causally relevant for the effects of interest, and relationships of similarity and difference that obtain between such features. The epistemic level, on the other hand, concerns our knowledge of and ability to learn about these entities, features, and relationships. The ontic/epistemic distinction makes clear that there is an important difference between the relationship ***R*** that is 'out there' and in fact holds between an experimental and target population; the assumptions ***P*** about this relationship that are required for an extrapolative inference; and the part of ***P*** that is (or can be) in fact supported by background causal knowledge ***K*** and supplementary evidence ***S***. Ideally, ***S*** and ***K*** will be jointly sufficient to validate our assumptions ***P***. Clearly, the more ***S*** and ***K*** manage to support ***P***, other things being equal, the closer we will come to successfully overcoming problems of extrapolation. At the same time,

however, the more extensive the conjunction of ***S*** and ***K*** *must* be to adequately justify the assumptions ***P*** demanded by specific extrapolation strategies, the more empirically demanding such strategies will be, and potentially undesirably so. In the limit, ***P*** would require us to assume all there is to assume about ***R***, and to fully support ***P***, ***S*** and ***K*** would need to encompass all there is to know about ***R***. This would be over-demanding. Realizing this yields two general challenges.

3.3.1 Overdemandingness

The first challenge is straightforward: an attractive strategy for extrapolation should not be epistemically over-demanding. If this were the only type of strategy feasible, then so be it, but it would seem preferable to have a strategy whose demands are realistically satisfiable. For instance, if justified extrapolation would require us to learn the physical causal microstructures underpinning the social phenomena of interest, and issues of causally relevant similarity and difference would need to be settled at this level, including details on how a physical causal basis realizes the social-level events of interest, and how doing so involves and conforms to accepted physical laws, this would be over-demanding and undesirable.

More generally (and bracketing, for now, the role of ***K***), overdemandingness concerns cases where supplementary evidence ***S*** is needed to support the assumptions ***P*** required for extrapolation, but acquiring this evidence would involve things that are extremely costly, difficult, or even impossible to learn, such as individual causal effects (which are typically considered in-principle unobservable magnitudes, cf. Rubin 1974; Holland 1986) or other causal features of the populations of interest that cannot (principally or realistically) be learnt from observational or experimental procedures. Moving away from such extreme cases, what counts as overly demanding may, of course, vary importantly from case to case, and there are no general strictures to be put in place at this stage. Nevertheless, when discussing extant strategies for extrapolation in the following chapters I will repeatedly emphasize important ways in which these strategies may be considered epistemically over-demanding, i.e. they require knowledge of causally relevant similarities and differences that is difficult to come by and, moreover, often demand confidence in having exhausted the relevant respects in which important differences between populations can obtain.

This, by itself, would be a rather thin criticism, however, as one may object that (at least some of) the strategies for extrapolation I will consider were perhaps never intended to overcome concrete problems of extrapolation on their own or were not intended to enable all kinds of extrapolative inference. Spelling out the abstract conditions under which extrapolation can be successful and addressing concrete problems of extrapolation are clearly substantively related. But perhaps, as long as abstract strategies still have some general things to say on the latter, these are different enough aims to warrant offering distinct proposals addressing each of them. While there may still be a pressing need for complementary empirical strategies to help acquire the supplementary evidence ***S*** that is required for underwriting extrapolation, we should perhaps not expect both the general recipes as well as all the concrete details for how to do the messy work of justifying our inferences from a single, overarching strategy for extrapolation. What is more, as perhaps already suggested by my arguments in *Chapter 2*, real-world problems of extrapolation might be too heterogeneous to permit a single, unified strategy that offers concrete enough guidance to help overcome a significant fraction of real-world extrapolation problems.

These points are well taken. What I will argue in the following chapters, however, is not only that existing strategies for extrapolation are epistemically over-demanding. My criticisms go deeper in the sense that even if there were off-the-shelf empirical strategies to obtain the supplementary evidence ***S*** required by existing strategies for extrapolation, acquiring this evidence faces a second important challenge.

3.3.2 The Extrapolator's Circle

This second challenge, called the *extrapolator's circle,* originally due to LaFollette and Shanks (1996), has more recently been brought to the fore by Steel (2008) and adds more concrete strictures on how epistemically demanding a strategy for extrapolation may be before it is overly demanding.

In a nutshell, the idea is as follows: if there are causally relevant differences between populations, which is likely in many social science and EBP scenarios, and we want to successfully extrapolate despite such differences, then we must somehow learn what differences there are and whether and how these differences bear on the effect of interest. As elaborated above, this will usually proceed against a background of a causal effect estimated in an experimental population as well as a conjunction of background

knowledge $\boldsymbol{K}$ and supplementary evidence $\boldsymbol{S}$. However, it is important to recognize that, irrespective of how challenging it is to obtain such resources, the conjunction of $\boldsymbol{S}$ and $\boldsymbol{K}$ needed for supporting specific assumptions $\boldsymbol{P}$ required by concrete strategies should not be so extensive that it allows us to answer the causal query of interest based on these resources *alone*. If this were the case, it would make our experimental evidence $\boldsymbol{E}$ redundant to answering our causal query.

This is clearly undesirable. We might argue that it would turn the problem of extrapolation into an altogether different sort of inferential problem, e.g. reasoning from background knowledge and piecemeal evidence pertaining to the causal makeup of the target to the effects of some intervention there. Or we might stick with Steel and say that, while we are still in the business of extrapolating, the experimental result is rendered redundant to the extrapolative conclusion. Either way, falling prey to the extrapolator's circle would undermine much of the promise that EBP holds, i.e. that causal effects learned in some population A can be informative for predicting causal effects of the same or similar interventions in other populations B, C, D, etc. Principally, the prevailing hope in EBP is that we can build libraries of evidence pertaining to the causal effects of different interventions, where such libraries can help us, to varying degrees of accuracy, predict what will happen if we implement these or other, similar interventions in novel policy targets. If the only way to make use of such evidence were to learn so much about the target populations that the experimental evidence became redundant to answering our questions, then why bother building libraries in the first place?

This worry becomes more pressing still when considering that acquiring supplementary evidence $\boldsymbol{S}$ about the target might sometimes involve implementing the intervention of interest there, rather than just learning something about the target from observational data, and that this can come at the risk of harming agents in the target (e.g. implementing an intervention that had positive effects in a study A, but induces significant harm in B). If intervention in the target were required, then effectiveness evidence would bring little epistemic value to the table beyond telling us about the effects of interventions where they have already been implemented, and perhaps giving us some hope that they *might* be effective in other places but falling short of supporting or even fully warranting that they will be. In such cases, the justificatory burden involved in implementing interventions in other places would be carried entirely by

supplementary evidence ***S*** that is distinct and unrelated to the primary effectiveness evidence ***E*** that we started from, and where ***S*** can sometimes only be acquired by implementing the intervention of interest in the target.

Similar to Steel, then, I take the extrapolator's circle to be a crucial challenge that any attractive strategy for extrapolation should be able to overcome, and preferably for a large class of cases. Not only should such strategies avoid being overly epistemically demanding concerning the conjunction of ***S*** and ***K*** required to justify extrapolation, but they should also steer clear of the extrapolator's circle, and widely so.

At this point, one might ask why it is important to consider the extrapolator's circle as a challenge for *strategies* for extrapolation. As suggested above, these strategies were perhaps not intended to provide complete recipes for extrapolation including for how to empirically support the assumptions that they involve. While the extrapolator's circle would still be a problem to be considered when extrapolating, it would apply to empirical strategies involved in providing supplementary information that is pertinent to an extrapolation, but perhaps not to the general strategies that tell us, first and foremost, which assumptions are needed to enable an extrapolative inference.

Yet, while it is true that the extrapolator's circle presents a challenge for empirical strategies concerned with providing support for assumptions involved in extrapolation, it also affects strategies for extrapolation if, by requiring certain kinds of assumptions that are in need of empirical support, they effectively demand evidence that is difficult to acquire without falling prey to the extrapolator's circle. Put differently (and bracketing once more the role of ***K***), demanding ***S*** is at best undesirable and disappointing if it were extremely difficult or impossible to acquire ***S*** without falling prey to the extrapolator's circle, and at worst a significant shortcoming on the part of strategies for extrapolation if there are either alternative ways ***S'*** to support the assumptions ***P*** that they involve, but where these remained unacknowledged by proponents of the strategies, or if there were yet other kinds of assumptions ***P'*** that could be licenced by yet other means ***S''*** and that would be suitable for reaching the same kinds of extrapolative conclusions. This is in contrast to Marcellesi (2015, 1315), who argues that the extrapolator's circle is not relevant to abstract analyses of the conditions under which extrapolative inference can proceed successfully but only to empirical strategies used for supporting their assumptions. Unlike Marcellesi, I maintain

that the extrapolator's circle is relevant also to abstract analyses and general strategies for extrapolation that require such assumptions in the first place.

For the remainder of this thesis, I acknowledge the extrapolator's circle as an important challenge for strategies for extrapolation. At the same time, I also want to make some further suggestions for how to improve our understanding of the underlying problem highlighted by the extrapolator's circle, and generalize it beyond the specific construal offered by Steel.

3.3.3 It's a Bind, Not a Circle

First, terminologically, the extrapolator's circle could be criticized for failing to be a circle proper. What happens when it is triggered is often not that we must already know $\boldsymbol{C}$ to infer $\boldsymbol{C}$, but rather that $\boldsymbol{S}$ and $\boldsymbol{K}$ jointly permit inferring $\boldsymbol{C}$, thus making $\boldsymbol{E}$ *redundant* to $\boldsymbol{C}$. Hence, going forward, I will refer to this inferential challenge as the *extrapolator's bind*.[6] The bind generalizes beyond the circle. It can accommodate cases where $\boldsymbol{C}$ must be known to infer $\boldsymbol{C}$, or $\boldsymbol{C}$ is trivially learned in the process, such as when we need to implement an intervention in the target to learn what its effects will be. But it is also more general, in that it captures cases where other resources, such as $\boldsymbol{S}$ and $\boldsymbol{K}$, displace $\boldsymbol{E}$ in inferring $\boldsymbol{C}$. The bind also captures two nuances at once: it characterizes a *problem*, first and foremost. But once the importance of this problem is recognized, this also makes the bind *normative*: it is a binding stricture on what we may and may not do when aiming to extrapolate successfully.

Second, Steel seems to assume that the extrapolator's circle is an all-or-nothing affair (Steel 2008, 78, 85, 86, 99). Both his own as well as LaFollette and Shanks' original formulation (1996, 157) suggest that it is triggered whenever we *know* the answer to our extrapolative query based on evidence from the target alone, so the experimental result is not relevant to answering our causal query anymore. But surely, there can be gradual variations on this situation, i.e. cases where the experimental result is rendered almost redundant to our conclusion, but not entirely.

One way of thinking about this is in terms of the *sensitivity* of an extrapolative conclusion $\boldsymbol{C}$ (say the quantitative prediction of a causal effect in the target) with respect to changes in the experimental result E to be extrapolated from. Understood in

[6] I am indebted to Finola Finn who has suggested this term to me.

this way, we might say that the *degree of relevance* of an experimental result ***E***, and hence the degree to which we manage to evade the extrapolator's bind, would be proportional to the degree of sensitivity of the extrapolative conclusion with respect to changes in the experimental result. Other things being equal, the more sensitive ***C*** is with respect to changes in ***E***, the more ***C*** hinges on, and is informed by ***E***, and hence the more *relevant* ***E*** is to ***C***. Lower levels of sensitivity, on the other hand, suggest that ***E*** plays a less important role and that ***C*** hinges relatively more on ***S*** and ***K***.

There are two problems with this way of thinking about relevance, however. The first relates to what we might call the *weight* of evidence (see Peirce 1878; Keynes 1921). Here, the idea is that ***E*** can be relevant to ***C*** in at least two different ways: it can change the *content* of our conclusion, say from ***C*** to ***C'***, and it can provide more or less *support* for one and the same conclusion ***C***. The first is captured by the idea of sensitivity above, i.e. the changes induced in a conclusion ***C*** as a response to changes in the evidence ***E***. However, the weight for ***C*** that ***E*** affords needs to be considered, too. Here, changing ***E*** (or subtracting or adding it from our evidence base) can yield important changes to how confident we are in ***C***, even though it does not change the content of ***C*** as such. This is the case, for instance, if we have yet other kinds of support ***S*** for the same conclusion ***C***, and ***E*** may hence add to the weight of the evidence in favour of ***C***, but does not change the content of ***C*** as such. In these cases, where ***E*** reinforces (perhaps significantly) what we would already believe to be the case from other sources, an account of relevance that only takes issue with changes in the content of ***C*** would say that adding or subtracting ***E*** from our evidence base is irrelevant to ***C***, despite the fact that ***E*** could still potentially drastically change our confidence in ***C***, and hence remain highly relevant to it.

This, of course, must be recognized by a richer account of the extrapolator's bind that draws on evidential relevance. We might hence say that a second, complementary way of spelling out relevance is with respect to changes in the weight of evidence for an extrapolative conclusion ***C***. If adding or subtracting ***E*** makes a larger difference to the weight in favour of ***C***, then, other things being equal, the more relevant ***E*** is to ***C***. Conversely, if ***E*** makes no difference to the weight of the evidence for ***C***, other things being equal, then it is irrelevant. This could be the case for instance, if ***S*** and ***K*** already warrant ***C*** beyond some relevant saturation threshold of confidence α, so that ***E*** would

not make a difference to whether we feel sufficiently confident in ***C*** one way or another.[7]

Of course, the way in which evidence, from a study population as well as from other sources, including particularly the target, bears on an extrapolative conclusion by means of changing its content and the weight in its favour is likely to be interactive. This means that there will likely be many cases where, by adding a token of evidence to our evidence base, we may not only change the content of the conclusion ***C*** (what it asserts about the target), or only the weight of the support in favour of ***C***, but both at the same time. For instance, when ***E*** contravenes the conclusion about the effect of interest in the target that would have been reached by considering only evidence ***S*** from the target but not the study, then this may plausibly change both the content of our conclusion (say, that an effect is positive rather than zero), e.g. from ***C*** to ***C'***, as well as the weight in favour of this conclusion (there was previously no weight in favour of our modified conclusion ***C'***).

For the present purposes, it is not particularly important to consider such interactions in more detail. What is more, while at least some of the above intuitions might also be considered to lend themselves to formalization in a Bayesian framework, my goals in this thesis are not importantly furthered by a formal treatment of evidential relevance, which is why I will not attempt such formalization here. For now, it is enough to note that in assessing the relevance of experimental evidence ***E*** for an extrapolative conclusion ***C*** we must consider both how ***E*** bears on the content of ***C*** as well as the weight in its favour.

A second potential problem for my suggestion to think about the extrapolator's bind in terms of relevance as outlined above is that the very nature of the extrapolative conclusion of interest can itself play an important role in determining the relevance of the evidence obtained from an experiment in supporting the conclusion. For instance, if ***C*** is highly general in nature, or highly abstract, such as when we are only interested in answering whether an intervention on X is positively causally relevant for Y in the target

[7] One might worry that there are important perspectival questions here about what evidence comes first: the experimental result E, or the supplementary evidence S. For instance, it could seem odd to say, e.g., that E is irrelevant to C, because there is additional evidence S invoked to infer C but where S was produced after E, which would render E irrelevant to C. I am not too concerned by this, as I am assuming that E alone is insufficient by itself to infer C with the desired level of confidence – this is what serious problems of extrapolation are all about. Hence, whether or not S was available before or after E, and what role K plays, is immaterial to assessing whether relying on S to infer C, which could not be done from E alone, would render E largely irrelevant to C in the senses outlined here.

(but not, say, in the magnitude of the effect), then this may itself bear importantly on how relevant ***E*** is to ***C***. For instance, the more general the type of conclusion we are interested in, the less relevant an experimental result will be to it in terms of potentially changing its content, other things being equal. Here, even major variations in the magnitude of a causal effect in a study population might induce rather small changes in the content of a qualitative conclusion, or perhaps no changes at all. This could seem counterintuitive, as the relevance (or rather irrelevance) of ***E*** to ***C*** would not seem to be driven by the fact that it is made redundant by other sources of support for ***C*** from the target. Indeed, even these other sources of support would seem to be rendered less relevant by this, too.

These implications are not problematic, however. They merely help us to recognize that context matters in determining how likely it is that we will fall prey to the extrapolator's bind. Clearly, if ***C*** is more easily reached, say because it is more general and hence less demanding to support, then this will, other things being equal, make it more likely that we fall prey to the extrapolator's bind. This is because even relatively incomplete knowledge about the target may be sufficient to reach ***C*** based on information about the target alone, unaided by ***E***. Keeping the evidence ***E*** from an experiment constant, as well as the support ***S*** obtained from other sources, including the target, then changing the nature of ***C*** in such a way that it is easier to support will simply make it more likely that ***E*** is rendered redundant to reaching ***C*** by ***S***. This hence preserves the way that ***S*** and ***E*** compete for relevance to ***C***, and hence preserves the key problem that the extrapolator's bind seeks to highlight.

Taking the above concerns into account, we can now formulate more precisely what it takes to evade the extrapolator's bind. An attractive strategy for extrapolation should steer clear of the extrapolator's bind *as best as it can*, i.e. by ensuring that the experimental result remains relevant, and potentially highly relevant, to the extrapolative conclusion of interest. Relevance, in turn, has two facets. One concerns the content of the conclusion, the other concerns the weight speaking in its favour, and both are important. Evading the extrapolator's bind should hence strike some reasonable balance on both desiderata. Relevance for ***C***'s content is important, but not if ***E*** puts little weight in favour of ***C***. Similarly, weight is important, but not if ***E*** does not, or indeed could not, bear on the content of our conclusion at all, and we would have inferred ***C*** regardless, although perhaps with slightly less confidence. What mixture of

these two aspects of relevance is adequate for a specific extrapolation will of course, in large part, hinge on specifics about the case, so not much more can be said at this stage other than that both should be considered.

A second conceptual refinement I wish to add to the extrapolator's bind relates to Steel's assumption that the extrapolator's *circle* is only triggered when information from the *target* is sufficient to answer the causal query of interest. However, it does not seem necessary that the information that renders $\boldsymbol{E}$ redundant to $\boldsymbol{C}$ comes from the target. It would be similarly disappointing if our causal query could be answered by only (or only by) considering evidence from populations other than the experimental population, which may or may not include the target and may even include information about the experimental population that is not supplied by the experiment itself. Here, too, $\boldsymbol{E}$ would be rendered irrelevant to $\boldsymbol{C}$, and a strategy for extrapolation demanding supplementary evidence from such sources would be as unsatisfactory as one that demanded such information to come only from the target.[8]

3.3.4 What's Successful Extrapolation?

With the above refinements in place, let me add some general strictures to my working analysis of extrapolation. Recall that I have characterized extrapolation as an inference $\boldsymbol{I}$**,** using evidence $\boldsymbol{E}$ obtained from an experimental population $\boldsymbol{A}$ to infer a conclusion $\boldsymbol{C}$ about a target population $\boldsymbol{B}$, with the help of assumptions $\boldsymbol{P}$ pertaining to the relation $\boldsymbol{R}$ between $\boldsymbol{A}$ and $\boldsymbol{B}$, as well as background knowledge $\boldsymbol{K}$ and supplementary evidence $\boldsymbol{S}$ that help support $\boldsymbol{P}$.

We can now extend this analysis by spelling out conditions for what constitutes *successful* extrapolation. These conditions help put important strictures on the assumptions $\boldsymbol{P}$ required by strategies for extrapolation, as well as the epistemic demands involved in producing supplementary evidence $\boldsymbol{S}$ that they require for support.

[8] In some such cases, we might also conclude that the experiment was simply poorly designed, e.g. when experimentalists could have used a more informative experimental design.

Based on the above, an extrapolative inference $\boldsymbol{I}$ aiming for a conclusion $\boldsymbol{C}$, based on experimental evidence $\boldsymbol{E}$, assumptions $\boldsymbol{P}$, supplementary evidence $\boldsymbol{S}$, and background knowledge $\boldsymbol{K}$, is **successful** if the following conditions hold[9]:

1) (***Informativeness***) Some conclusion $\boldsymbol{C}$ of the desired kind is inferred.
2) (***Justification***) $\boldsymbol{C}$ is adequately justified, i.e. our confidence in $\boldsymbol{C}$ that is warranted by a combination of $\boldsymbol{E}$, $\boldsymbol{P}$, $\boldsymbol{K}$, and $\boldsymbol{S}$ exceeds some threshold $\boldsymbol{\alpha}$.
3) (***Accuracy***) $\boldsymbol{C}$ is accurate, relative to some standard of accuracy $\boldsymbol{\beta}$.
4) (***Relevance***) $\boldsymbol{C}$ is inferred in such a way that $\boldsymbol{E}$ remains relevant to $\boldsymbol{C}$ beyond some threshold $\boldsymbol{\gamma}$.

The first condition helps ensure that $\boldsymbol{C}$ speaks to what we want to know about the target and not to some other, potentially related question. As we will see, some strategies for extrapolation are limited in the kinds of conclusions that they can yield, so they will be successful or unsuccessful depending on whether their abilities to provide certain kinds of conclusions match with those we are interested in.

The second condition requires that $\boldsymbol{C}$ is adequately justified. This is supposed to ensure that $\boldsymbol{C}$ is not obtained by sheer luck or accident, but that it enjoys sufficient support, and is arrived at by means of a sound reasoning process. As the standards for what counts as adequate justification will plausibly differ from case to case (such as when the stakes involved in drawing a mistaken conclusion differ), a threshold α can be used to capture both the gradual nature of justification as well as the idea that we may have sharp or fuzzy standards for what level of justification is needed to justify subsequent action (e.g. intervention in the target).

The third condition requires that what $\boldsymbol{C}$ asserts about the target is accurate with respect to what is (or will be) the case there.[10] This captures the idea that it is not enough to have a well-justified conclusion that speaks to queries of interest to us, but that ultimately turns out to be radically mistaken. Accuracy, here, will, of course, be context-dependent again. It could mean, for instance, that a causal effect predicted by

[9] One might wonder why there is no fifth condition pertaining to the *validity* of the inference schema used. This, of course, is an important ingredient of successful extrapolative *inference.* I will bracket it here from explicit consideration, however, as I will, for the most part, grant that the inference schemas supplied by the strategies that I will examine are either valid, or invalid but strongly compelling, if adequately justified.

[10] We might also be interested in the *precision* of the conclusion. I will assume that issues surrounding precision are captured by condition 1), however.

$\boldsymbol{C}$ to be positive in the target indeed turns out to be positive, $\boldsymbol{C}$ correctly predicts the magnitude of a causal effect, $\boldsymbol{C}$ correctly instructs us to perform a co-intervention on a moderating variable to achieve a specific outcome distribution in the target, etc. So accuracy can come in different forms, and we may hence wish to spell out varying standards of accuracy β relative to our purposes. Of course, it is also important to recognize that, unlike some other conditions, accuracy can often only be determined after the fact. This does not pose any special problems to more general assessments of whether strategies for extrapolation meet this desideratum, however, as, if they repeatedly and consistently fail to provide accurate conclusions, then this is still informative about how successful the application of these strategies might be in future instances.

Finally, the fourth condition captures the extrapolator's bind. It maintains that successful extrapolation requires that $\boldsymbol{E}$ remains relevant to $\boldsymbol{C}$ beyond some threshold of relevance γ. γ is merely a conceptual placeholder, of course, and it might be difficult to operationalize relevance in such a way that measuring it and determining a meaningful threshold γ is practically feasible. Spelling out relevance in more detail would surely be interesting but is beyond the scope of the current project. The major conclusions to be developed in the following chapters remain largely untouched by lack of a more explicit treatment, however, and it seems enough to note that it seems plausible to think that some threshold of relevance that is above full-blown irrelevance of $\boldsymbol{E}$ to $\boldsymbol{C}$ seems desirable in many real-world extrapolations. RCTs, for instance, do not usually come cheap, and if our best available extrapolation strategies would standardly make $\boldsymbol{E}$ obtained from RCTs almost entirely irrelevant to $\boldsymbol{C}$, then this would seem undesirable.

Considered together, then, successful extrapolation requires minimally that an extrapolative conclusion of the envisioned kind is reached; that the conclusion is justified to some sufficient degree; that it is accurate to some sufficient degree; and that the experimental evidence we are extrapolating from remains relevant to it to some sufficient degree. Moving beyond the necessary, and taking the gradual nature of 2)-4) into account, (highly) successful extrapolation requires a good mixture of these desirable attributes and if any of them fail to be realized, extrapolation will be unsuccessful.

Such failures may come in different degrees and in different forms. For instance, a failure of 1)-3) might be called a failure to extrapolate successfully, full stop

(potentially with different weights). A complete failure of 4), however, in the sense that *E* is almost or entirely irrelevant to *C*, may lead to an even stronger conclusion. We might say that not only does successful extrapolation fail, but, particularly when 1)-3) are indeed satisfied, there is a special kind of failure going on: one fails not only to extrapolate successfully, but one fails to extrapolate *at all*, since *E* does not relevantly figure in inferring *C* anymore.

One important point to note at this stage, which has remained implicit so far, is how the above analysis of successful extrapolation tells us something about how attractive a strategy for extrapolation is. So far, what seems to be provided by the analysis is first and foremost a clarification of what it means to successfully extrapolate in any concrete instance. So how does this touch upon the more general issue of what we want a *strategy* for extrapolation to provide us with?

The answer is that the analysis works 'bottom-up', and in a context-sensitive way. We start from single instances of extrapolation and then, based on how specific strategies for extrapolation handle specific instances of extrapolation, or types of extrapolation, proceed towards more general conclusions about the attractiveness of these strategies. This is advantageous as it seems likely that most serious candidate strategies for extrapolation will be able to achieve some instances of successful extrapolation (and perhaps consistently for certain kinds of extrapolation), but fail in other (kinds of) cases. So success is piecemeal, likely to be heterogeneous, and one failed instance of extrapolation does not make for an entirely failed strategy. But if the circumstances under which specific strategies for extrapolation are prone to failure are important, systematic, and general enough, then this can still licence relatively broad conclusions about their attractiveness. Importantly, we can also *predict* the success of an extrapolation, as well as of a strategy more generally, at least with respect to relevance. If a strategy for extrapolation, in virtue of the assumptions it makes and the support they require, routinely requires extensive support ***S*** and ***K*** that would clearly render ***E*** irrelevant to ***C***, then, even before endeavouring to make an extrapolation, we can predict that it will fail to be successful. Finally, the fact that relevance is a part of success also yields a more general intuition: if it is foreseeable that a strategy is bound to yield unsuccessful extrapolations in virtually every instance (of a kind of extrapolation), we may conclude that it is entirely unsuccessful (for this kind of extrapolation).

There are two additional features of the above analysis that deserve brief emphasis; they concern how its conditions are related. First, it seems plausible to think that *accuracy* and *justification* may interact in important ways. For instance, the higher the desired standard of accuracy β, other things being equal, the more justification will typically be required. Conversely, for lower standards of accuracy, e.g. when the stakes are low, we might be more lenient when it comes to justifying our inference. A second, more interesting type of interaction concerns the relation between *accuracy* and *justification* on the one hand and *relevance* on the other. Here, it seems that there is a *trade-off* between them. Increasing accuracy can and must often be achieved by adding supplementary resources to justify an inference, e.g. because the assumptions required to enable more sophisticated inferences, and more accurate conclusions, often require more extensive support. Yet, this will typically come at the price of sacrificing relevance, as when adding more and more supplementary resources, in particular resources that say something about the effects of interest in the target on their own, we run the risk of displacing the relevance of E to C. Beyond suggesting that it might be interesting to explore such interactions in more detail, I will not expand in more detail on such relationships here, as this would not seem to importantly further the aims of this project.

In sum, the above strictures highlight the relation between extrapolation, successful extrapolation, and the extrapolator's bind. Specifically, while extrapolation simpliciter only requires that evidence $\boldsymbol{E}$ from an experiment is used to infer an extrapolative conclusion $\boldsymbol{C}$, *successful* extrapolation, among other things, requires that E remains relevant to C, so the extrapolator's bind is now built into our analysis of successful extrapolation.

3.4 Conclusions and Outlook

The discussion offered in this chapter allows us to make important progress in the way of spelling out what general kinds of assumptions strategies for extrapolation will need to involve, as well as what kinds of desiderata they should meet to be attractive. They should, ideally, help us infer, on the basis of $\boldsymbol{E}$ and a conjunction of background assumptions $\boldsymbol{P}$, and background knowledge $\boldsymbol{K}$ and supplementary evidence $\boldsymbol{S}$ pertaining to $\boldsymbol{R}$, an action-guiding, ampliative conclusion $\boldsymbol{C}$ about the causal effects of interest, where *ampliative* means that the conclusion should go beyond what we already know

about the target by virtue of ***S*** and ***K***. This might not seem like a tall order for now, but as I will argue throughout the following chapters, existing strategies for extrapolation are liable to fall well short of this desideratum in many cases.

In light of the discussion provided in this chapter, the vulnerability of these strategies to fall prey to the extrapolator's bind should not be surprising. Extrapolation involves learning and accommodating information about causally relevant similarities and differences between experimental and target populations. But as the assumptions and guiding ideal outlined in this chapter make clear, in the limit, accurately predicting a causal effect in a target despite potential causally relevant differences between populations might require one to know about all relevant differences and similarities, as any unknown difference may hamper successful extrapolation. Yet, because some of these differences and similarities are at least extremely difficult to learn without falling prey to the extrapolator's bind, extrapolation with certainty and/or maximal accuracy will not only remain an elusive ideal, but also undesirable, since it undermines success.

Moving from the ideal to the practically feasible, it is clear that any attractive and useful strategy for extrapolation must stop well short of these extensive requirements. It also suggests that any such strategy must tell us a rich story about the inferential leap that will necessarily persist between what we (must) learn for the purpose of successful extrapolation and what we are interested in inferring on the basis of what we have learnt. This is something that extant strategies have not been telling us much about: what are good ways of closing, as much as possible, the gap that will persist between what we are in a position to learn, at acceptable cost, and without falling prey to the extrapolator's bind, and what we aim to infer. They provide (potentially adequate) accounts of the conditions under which successful prediction of causal effects in the target is possible in principle, as well as what assumptions are needed to help infer the conclusions of interest. But they do not tell us how to acquire information about the experimental and target population to support these assumptions, and in a way that helps us *successfully* extrapolate, i.e. extrapolate without falling prey to the extrapolator's bind.

References

Bareinboim, E. and J. Pearl. (2012). "Transportability of causal effects: Completeness results", In: Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12), Menlo Park, CA.

Biddle, J. (2013). "State of the field: transient underdetermination and values in science". *Philosophy of Science*, 44: 124–133.

Cartwright, N. D. (2009). "How to do things with causes", *Proceedings and Addresses of the American Philosophical Association*, 83(2): 5–22.

— **(2013a).** "Knowing What We Are Talking About: Why Evidence Doesn't Always Travel". *Evidence and Policy: a Journal of Research, Debate and Practice*, 9(1): 97-112.

— **(2013b).** "Evidence, Argument and Prediction". In: V. Karakostas, and D. Dieks (eds.), EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings. Cham, Switzerland: Springer International Publishing Switzerland.

Dahabreh, I. (2018). "Randomization, randomized trials, and analyses using observational data: A commentary on Deaton and Cartwright", *Social Science & Medicine*, 210: 41-44.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.

Frigg, R., and I. Votsis. (2011). "Everything you always wanted to know about structural realism but were afraid to ask", *European Journal of Philosophy of Science*, 1(2): 227-76.

Fuller, J. (2018). "The Confounding Question of Confounding Causes in Randomized Trials", *The British Journal for the Philosophy of Science*, axx015.

— **(2019).** "The myth and fallacy of simple extrapolation in medicine", *Synthese*, online first, https://doi.org/10.1007/s11229-019-02255-0.

Guyatt, G., D. Rennie, M. O. Meade, and D. J. Cook. (2008). *Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice*. 2nd ed. New York: McGraw-Hill.

Holland, P. (1986). "Statistics and Causal Inference", *Journal of the American Statistical Association,* 81(396): 945-60.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer. (2005). "Predicting the efficacy of future training programs using past experiences at other locations", *Journal of Econometrics*. 125: 241–70.

Hume, D. ([1739] 1975). In: L. A. Selby-Bigge, and P. H. Nidditch (eds.), *A treatise of human nature*. Oxford: Clarendon Press.

Ioannidis, J. (2005). "Why most published research findings are false". *PLoS Medicine*, 2(8): e124.

Keynes, J. M. (1921). *A Treatise on Probability*. London: Macmillan.

LaFollette, H., and N. Shanks. (1996). *Brute Science: Dilemmas of Animal Experimentation*. New York: Routledge.

Little, D. (1993). "On the Scope and Limits of Generalizations in the Social Sciences." *Synthese*, 97(2): 183-207.

Marcellesi, A. (2015). "External Validity: Is There Still a Problem?". *Philosophy of Science*, 82(5): 1308-17.

Muller, S. M. (2014). "Randomised trials for policy: a review of the external validity of treatment effects". Southern Africa Labour and Development Research Unit Working Paper 127, University of Cape Town.

— **(2015).** "Interaction and external validity: obstacles to the policy relevance of randomized evaluations", *World Bank Economic Review*, 29(1): 217-25.

Peirce, C. S. (1878). "The Probability of Induction", *The Popular Science Monthly, Illustrations of the Logic of Science*, XII.

Petticrew, M., and I. Chalmers. (2011). "Use of research evidence in practice", *Lancet*, 378(9804): 1696.

Rubin, D. B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology,* 66(5): 688–701.

Russell, B. (1927). *The analysis of matter*. London: George Allen & Unwin.

Steel, D. (2004). "Social Mechanisms and Causal Inference", *Philosophy of the Social Sciences* 34(1): 55-78.

— **(2008).** *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.

Vivalt, E. (2019). "How Much Can We Generalize from Impact Evaluations?". Unpublished manuscript, ANU, Canberra. Retrieved from: http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf (retrieved Feb. 28, 2019).

CHAPTER 4

Argument-Based Extrapolation

4.1 Introduction

The previous chapters have laid the foundations for a critical assessment of different strategies for extrapolation. In this chapter, I begin to draw on these resources, focusing on Nancy Cartwright's contributions to the literature on extrapolation in EBP, in particular her *Argument Theory of Evidence* (Cartwright 2013a).

Cartwright's *Argument Theory* is a theory of *evidential relevance* and was developed against the background of her sustained line of criticism against naïve extrapolation from RCT results (Cartwright 2009; 2011; 2012; Cartwright et al. 2009; Cartwright and Stegenga 2011; Cartwright and Hardie 2012; Deaton and Cartwright 2017). Here, Cartwright argues that experimental evidence of the effectiveness of an intervention in A, on its own, is only ever evidence of and for itself (2013b). In order for an experimental result $\boldsymbol{E}$ to speak to questions about the effects of an intervention in other populations, places, and times B, something beyond $\boldsymbol{E}$ itself is needed, i.e. something that establishes the evidential relevance of $\boldsymbol{E}$ to questions pertaining to B.

Cartwright's Argument Theory tells us that in order to render evidence $\boldsymbol{E}$ from A relevant to questions about distinct targets B, we need valid and sound arguments. Framed in terms of the analysis provided in *Chapter 3,* these are arguments where $\boldsymbol{E}$ as well as supplementary knowledge and evidence pertaining to the relation $\boldsymbol{R}$ that holds between A and B figure as premises, and which are aimed at inferring a conclusion $\boldsymbol{C}$ that answers our causal queries about a target.

Beyond Cartwright's general emphasis that extrapolative inference should proceed in terms of valid and sound arguments, she also offers a concretization of this abstract proposal in terms of a specific *effectiveness argument*, which is used to illustrate what kinds of premises would be sufficient to infer a conclusion about the effectiveness of an intervention in a novel target. This construal provides more detail about the specific nature of the assumptions (read: premises) that might be involved in extrapolation and what kinds of additional evidence could help establish the relevance of RCT results to answering questions about novel targets.

In what follows, I offer a critical discussion of these proposals. In *Section 2*, I provide a broad sketch of the Argument Theory and highlight what I consider to be several of its key virtues. I also expand on some of the commonalities and differences between the Argument Theory and other strategies for extrapolation to be discussed later.

Section 3 proceeds to offer some critical remarks. I begin with two concerns targeting the specific effectiveness argument provided by Cartwright. While it is best understood as a mere illustration of how the Argument Theory can work, it usefully exemplifies what problems we might encounter when extrapolating with the help of arguments. As I will argue later, these problems will also carry over to more sophisticated arguments, potentially informed by the resources provided by other strategies for extrapolation. My arguments should hence not be taken to suggest that the effectiveness argument is Cartwright's final word on matters of extrapolation, nor that the persuasiveness of the Argument Theory stands and falls with that of the effectiveness argument.

Following my critical discussion of the effectiveness argument, the interested reader can turn to *Appendix 1*, where I take a short detour and focus attention on one of the main concepts that Cartwright invokes in specifying the effectiveness argument: *causal support factors*. I argue that, while useful and important, the concept of support factors is in need of some conceptual refinement, including distinctions between different kinds of support factors, and clarification concerning the relationship between support factors and other kinds of variables that can induce causally relevant differences discussed in *Chapter 2*. I provide an updated and comprehensive analysis of support factors to help clarify these issues.

In *Section 4*, I offer a more general discussion of the Argument Theory. I argue that while the specific effectiveness argument, understood as an illustration of the Argument Theory's capabilities, is unlikely to provide a recipe for successful extrapolation, the Argument Theory still has important merits. It may not provide a working strategy for extrapolation from start to finish, but can be useful as a general framework in which to scrutinize extrapolative inference including existing and yet to be developed proposals for how to achieve such inference. It can hence figure as a useful background against which the critical examinations of other extrapolation strategies in subsequent chapters can proceed.

4.2 The Argument Theory

As anticipated above, Cartwright's Argument Theory is primarily what it says on the box: a theory of evidence, or *evidential relevance*. As such, it is intended to elucidate general questions pertaining to the issue of when, i.e. under what conditions, and how, i.e. in what particular ways, something can be evidence for something else. The presentation of this theory, however, proceeds in the more specific context of when and how experimental evidence of the causal effects of interventions in some study population is evidentially relevant for claims concerning the causal effects of these interventions in novel targets. Finally, beyond these general contributions, Cartwright's exemplary effectiveness argument provides us with a concrete illustration of how a specific argument form may help us establish such claims.

At the most general level, Cartwright's Argument Theory hence seems to provide us with three things at once: 1) a general theory of evidential relevance, 2) when applied to extrapolation specifically, a general framework for articulating and scrutinizing extrapolative inference, and 3), as exemplified by Cartwright's effectiveness argument, a concrete strategy for extrapolation. As will become clear in *Section 5*, this latter understanding is somewhat contentious, as both Marcellesi (2015) and Cartwright (in personal communication) conceive of the Argument Theory only in the first two senses. I will not focus on the first sense here, as I am concerned with issues of extrapolation. 2) and 3) will hence be my primary focus. For the moment, let me proceed on the assumption that Cartwright's Argument Theory pertains to issues of extrapolation in at least two ways:

First, according to the Argument Theory, justified extrapolation, i.e. the activity of using evidence of causal effects as evidence pertaining to causal conclusions about some distinct target, should be (or is at least helpfully) understood as proceeding in terms of valid and sound arguments*,* i.e. arguments that entail (some kinds of) conclusions about the causal effect of interest in the target. If an extrapolative inference is not cast in terms of such an argument, then either justification for the conclusion would remain implicit, which is taken to be undesirable, or no justification at all would be involved, meaning that extrapolation would be based on hope, and fail to be successful in at least one of the senses elaborated in *Chapter 3*. Likewise, If an extrapolative inference could not be cast in terms of an argument, then no valid/compelling inference can take place. Second, illustrating how such arguments

could proceed in practice, Cartwright provides her exemplary effectiveness argument. Let me expand on this argument in more detail.

4.2.1 The Effectiveness Argument

According to Cartwright (2013a, 14), the weakest construal of an effectiveness argument is as follows, where X is the intervention variable, Y the outcome, and A and B denote the populations from which and to which experimental results are extrapolated:

P1: X can play a causal role in the principles that govern Y*'s* production in A.

P2: X can play a causal role in the production of Y in B *if* it does so in A.

P3: The support factors necessary for X to make a positive contribution are present for at least some individuals in B.

C: Therefore, X can play a causal role in the production of Y in some individuals in B.

Here, P1 is established by means of a separate argument. This argument involves an experimental result, as well as additional premises that establish the validity of this result, e.g. premises that mirror the typical identification assumptions involved in RCTs, such as that randomization was successful; that there was no differential post-randomization or post-treatment attrition between treatment and control groups; that there were no spillover effects; that blinding of participants, administrators, and evaluators succeeded, etc. If these (and other) assumptions are satisfied, it follows that a significant mean difference in the outcome Y between treated and untreated units establishes that X can play a causal role in the production of Y in A.

This alone, of course, does not entail that X can play a causal role in the production of Y in any place, time, or population other than A. This is Cartwright's cautionary message. To establish any such conclusion about a distinct target B, additional premises are required, and evidential support for these premises is needed to justify the inference.

P2 and P3 are supposed to do the first part of this job. They specify sufficient conditions on the relation $\boldsymbol{R}$ between two populations for an intervention on X to be effective in the target. P2 is broadly concerned with the causal mechanisms that govern the production of the outcomes of interest in the two populations. It asserts that these

mechanisms need to be related in such a way that if X is causally relevant for the production of Y in A then it follows that X is causally relevant for the production of Y in B. In essence, P2 minimally requires that the causal mechanisms governing the production of Y in A and B are identical in the sense that X is causally relevant for Y in *some* way, i.e. there is some causal relationship or pathway between X and Y in both populations, such that at least under some background conditions, intervening on X *can* induce changes in Y.

P3 specifies additional constraints on $\boldsymbol{R}$. Specifically, while P2 only guarantees that intervening on X *can* induce changes in Y in both populations, P3 is supposed to ensure that this is actually the case in the target, at least for some individuals. Specifically, P3 requires that the so-called *causal support factors* necessary for an intervention on X to actually induce changes in Y are realized for at least some individuals in B. This requires some elaboration.

There are two important and closely connected ingredients that figure centrally in Cartwright's contributions to issues of extrapolation: the concept of *causal support factors* and the view of causes as *INUS conditions*. The latter is adopted from Mackie (1965) and amounts to the substantive view that causes rarely operate on their own to produce some outcome Y. Instead, they are best understood as ***I***nsufficient but ***N***on-redundant parts of ***U***nnecessary but ***S***ufficient conditions for Y. To produce some outcome Y, causes typically operate in tandem. Together, conjunctions of causes form sufficient conditions for an outcome Y, sometimes called (sufficient) *causal complexes* or *causal pies*[1], where the constituent causes may interact or may need to be co-instantiated in certain ways to jointly bring about Y, but where none of the constituent parts is sufficient for Y by itself. Nor is any constituent part of a causal complex ultimately necessary for Y, as long as there are alternative arrangements of factors in the same causal complex that would be sufficient to bring about Y, or when there are alternative causal complexes constituted by yet other factors, which are jointly sufficient for Y as well.

Adopting this general view of causes as INUS conditions helps Cartwright emphasize that the kinds of causal variables we are typically interested in intervening with are not 'pure' causes that will have some effect on the outcomes of interest

[1] *Causal pies* are a metaphor frequently used in epidemiology (Cartwright and Stegenga 2011, 302).

irrespective of context. Rather, whether or not an intervention on X brings about some outcome Y is always a matter of whether other INUS conditions that, together with X, are sufficient for Y, are also present. These additional factors are what Cartwright calls *support factors*.

Importantly, support factors should not be understood to take the back seat in producing outcomes. What counts as a support factor and what counts as the causal variable that an intervention is supposed to manipulate will often be a matter of our particular epistemic or pragmatic interests. So, causes and causal support factors are not necessarily relatively more or less privileged candidates for intervention. They can often figure as equals in causal complexes, where only interest dictates what is the target variable of an intervention and what is a support factor necessary for, or conducive to this intervention making a suitable contribution to the production of an outcome.

With these clarifications in place, P3 in the effectiveness argument specifies how the causal support factors necessary for an intervention to yield its intended effect need to be instantiated in the target. They need to be instantiated in such a way that the intervention of interest makes some contribution to the outcome. Minimally, this requires that the requisite support factors are suitably realized for at least some individuals. If this is the case, then an intervention on X can succeed in bringing about changes in Y for at least these individuals, and hence yield changes in the average of Y across the target population.

In sum, P2 and P3 specify conditions on the relation $\boldsymbol{R}$ between experimental and target populations that help ensure that an intervention that is effective in A will also be effective in B. These conditions are somewhat milder than those required by other strategies (Hotz et al. 2005; Steel 2008; Bareinboim and Pearl 2012), but this potential advantage for Cartwright comes at the cost of weaker results. Let me expand on this and other, related concerns, starting with the scope of the effectiveness argument.

4.3 Scope, Assumptions, and the Extrapolator's Bind

4.3.1 Scope

My first concern is that the scope of the effectiveness argument is too narrow. In *Chapter 2* I have distinguished different kinds of extrapolative queries, which can now help us map the scope of the effectiveness argument more precisely.

The effectiveness argument is limited to addressing two kinds of extrapolative queries. The first are *efficacy queries*, i.e. queries asking whether an intervention makes any contribution at all for any individual in the target. This is guaranteed as long as P2 is satisfied and the minimal conditions of satisfaction for P3 are met, i.e. at least some individuals in the target exhibit realizations of support factors sufficient for the intervention to make a contribution to the outcome for these individuals. So minimally, qualitative extrapolative queries of the form "will an intervention on X make *some* contribution to Y in the target?" can be addressed by the effectiveness argument.

The effectiveness argument can also handle *matching effectiveness queries*, i.e. queries asking whether the effect of X on Y in B will be identical (or otherwise highly similar) to the effect measured in A. Affirming this question will require some modification of P3, so that P3 demands that the support factors necessary for the intervention to make a contribution to the outcome are distributed in such a way in the target that the intervention can make the same (or otherwise highly similar) contribution as in A. Standardly, this will require that the distribution of support factors in the target is the same as in the experimental population.[2] If this is the case, it permits the conclusion that the ATE induced by the intervention will be the same (or otherwise highly similar) in B as in A.

This is about as far as the effectiveness argument can take us with respect to addressing extrapolative queries of different kinds. And this is not very far, nor does it cover a wide range of important extrapolative queries, including particularly queries that ask "what is the effect in the target, given the effect in the experiment and given that there are causally relevant similarities and differences between populations?". This is an important shortcoming, as I consider this type of query to be typical in EBP. The reason is that in many EBP scenarios experimental and target populations are likely to exhibit causally relevant differences, so there is little hope of getting lucky and encountering situations where all we need to do is clarify issues of similarity or identity in causal mechanisms and assess the suitability of support factor distributions in the target in either of the two ways permitted by the effectiveness argument.

However, as explained above, the effectiveness argument only covers extreme cases where support factors are either distributed in such a way that the intervention is

[2] Or, in cases where effects are linear in mediators and moderators, that the means of the distributions are the same.

effective for some agents in B, or in the same way as in A. Yet, the former may not yield conclusions that are sufficient for licencing policy action, as it may not be enough to infer that some agents will experience positive effects (while permitting, as Cartwright acknowledges, that other agents might be harmed by the intervention; see also Weinberger 2014). The latter is only useful in extremely rare cases as it is often plausible to assume, as argued in *Chapter 3*, that populations exhibit at least some causally relevant differences. An effectiveness argument that only addresses these two narrow classes of cases seems consequently too limited in scope to be useful for the general purpose of addressing problems of extrapolation routinely encountered in EBP. Here, successful extrapolation is often not merely a matter of providing support for similarities in mechanisms and suitable support factor distributions, but the aim will be to obtain a conclusion about causal effects in the target despite, and taking into account the effects of, causally relevant differences. As suggested in *Chapter 3*, these cases do not necessarily pose insurmountable problems, and a failure of the effectiveness argument to handle such cases should not be taken to suggest that one should refrain from extrapolating, but only that the effectiveness argument is not suitable for this purpose.

As I will explain in the chapters to follow, other strategies for extrapolation offer ways to answer such queries (e.g. those proposed by Bareinboim and Pearl 2012; Hotz et al. 2005). In essence, the idea underlying these approaches is that at least for some kinds of causally relevant differences between populations, in particular differences in the distributions of moderating variables that bear on effect magnitudes, the effect of interest in the target can still be correctly predicted as long as we can account for how such differences bear on differences in these effects.

The effectiveness argument, in contrast, lacks the resources to do this. While it specifies conditions on causal support factors (of which moderating variables might be considered an instance; more on this in *Appendix 1*) and how they need to be distributed in A and B, it does not offer conclusions for cases where the distributions of support factors or other causally relevant features differ between populations and we wish to take such differences into account when inferring a conclusion about the magnitude of the effect in B. While there is some flexibility in Cartwright's effectiveness argument pertaining to the exact specification of P3, this flexibility is constrained to two extremes, and there remains an important lack of guidance concerning how to handle

cases where populations differ in causally relevant respects, including in how support factors are distributed, and how to obtain quantitative predictions of causal effects in such cases.

Let me turn to a second concern, which takes issue with important ambiguities in the premises figuring in the effectiveness argument, as well as how empirically supporting these premises can raise concerns about the extrapolator's bind.

4.3.2 Causal Assumptions and the Extrapolator's Bind

Setting issues of scope aside, the effectiveness argument seems largely compelling. However, on closer inspection, its premises are ambiguous and in need of concretization. As I will argue, this concretization may involve yet stronger assumptions. This is a problem, as validating these assumptions may require extensive causal knowledge of the target. This raises the important concern that in using the effectiveness argument to address real-world problems of extrapolation, the knowledge about the target required for supporting its premises might be so extensive that obtaining it would allow us to answer the causal query of interest based on this knowledge alone, thus rendering the experimental result redundant to the conclusion of the argument. Let me revisit P2 and P3, in turn, to explain why they are ambiguous.

Recall that P2 asserts that X can play a causal role in the production of Y in B *if* it does so in A. How does the relation $\boldsymbol{R}$ between A and B need to look like in order to satisfy this conditional? As suggested above, minimally, the causal mechanisms governing the production of Y in A and B must involve X in such a way that, in principle, interventions on X can induce changes in Y in both populations.

However, this condition can be satisfied in at least three different ways. First, the causal mechanisms might be identical in both populations[3], in which case it would follow that whatever intervention can be effective in A can also be effective in B. Second, the effects of X on Y may be governed by different causal pathways in the two populations, each involving a different and non-overlapping set of moderating and mediating variables and support factors of the X-Y-effect. Third, there can be intermediary cases where there is a partial overlap in what variables are involved in the

[3] i.e. at the level of abstraction relevant to the causal queries of interest, but not, of course in their microphysical underpinnings.

causal pathway(s) from *X* to *Y*, where some but not all of the moderating and mediating variables and support factors are identical between populations.

A similar ambiguity surrounds P3, which asserts that the support factors necessary for *X* to make a positive contribution are present for at least some individuals in the target. Again, we can interpret P3 in at least three different ways:

> **P3.1** The support factors necessary for *X* to make a contribution to *Y* *in the target* are present for at least some individuals there.
>
> **P3.2** The support factors necessary for *X* to make a contribution to *Y* *in the experimental population* are present for at least some individuals in the target.
>
> **P3.3** The support factors necessary for *X* to make a contribution to *Y* *in the experimental and target population* are present for at least some individuals in the target.

In short, what is unclear is whether the support factors invoked in P3 are support factors for the effect of interest in the experiment, in the target, or in both.

These ambiguities are important because nothing in the effectiveness argument up to P3 ensures that the same support factors are involved in producing the effects of interest in both populations. For instance, due to cultural and institutional differences, the effects of a social policy might be positively moderated (or supported, in the terminology of support factors) by informal institutions such as trust in one population and by a strong legal system in another. This can pose a problem for the effectiveness argument, as P3.1 and P3.2 would leave open whether the support factors for the *X*-*Y*-effect are the same in the target as in the experiment.

The first construal only ensures that, whatever the support factors for the *X*-*Y*-effect in the target are, they are realized for at least some individuals there. The second construal ensures that the support factors necessary for the *X*-*Y*-effect *in the experiment* are realized in the target as well. This could be despite the fact that the support factors required for the *X*-*Y*-effect *in the target* might not be the same as (or only partially overlapping with) those in the experiment. Sticking with the example, it might not be enough, for instance, that there are high levels of trust in a target population when the support factor relevant for the effects of interest there is whether there is a strong legal system. Only the third construal would evade this problem, as it assumes that the

support factors for the *X*-*Y*-effect are indeed the same in both populations, and asserts that they are realized for at least some individuals in the target.

The two ambiguities outlined above are importantly related. If P2 is understood to assert that the causal mechanisms governing the effects of interest are exactly the same in both populations, then the ambiguity surrounding P3 does not matter. In virtue of identity in mechanisms, P2 would guarantee that the support factors will be the same in the experiment and in the target, and it is enough to learn that they are realized for some agents in the target to conclude that the intervention will be effective at least for these agents.

However, if the causal mechanisms are not identical between populations, but only partially so, or if they are entirely different, this conclusion no longer follows on all construals of P3.

P3.1 would still yield this conclusion; it just asserts that, *whatever* the mechanism governing the *X*-*Y*-effect in the target, the support factors for the intervention *in the target* are realized for some individuals *in the target*, so the intervention will be effective for these individuals. P3.3 would not yield the conclusion, at least not unless we further assume that the support factors play the same qualitative roles in both populations, which puts further, heretofore implicit, constraints on how mechanisms must be related. P3.2, too, would not yield the conclusion. It would require the additional assumption that the support factors in both populations are indeed the same and that they are involved in producing the effects of interest in the same way in both populations.

The most plausible combinations of the above construals of P2 and P3 hence seem to be the following. First, P2 could be taken to say that mechanisms in both populations need to be identical, and any construal of P3 would be enough to yield the conclusion about the target. Alternatively, P2 could ensure only that *X* is causally relevant for *Y* in both populations in *some* way. But then P3 would need to ensure that the support factors are either the same in both populations (P3.3) or, if they are not, that the support factors necessary for the intervention to be effective *in the target* are realized for some individuals there (P3.1).

Let me expand on how these alternatives can raise important concerns about the extrapolator's bind. The general concern here will be that empirically supporting either

set of premises can require so much causal knowledge about the target that the conclusion to be reached by the effectiveness argument could also be reached based on information about the target alone.

Before we move from the ontic level of specifying constraints on the relation ***R*** between populations to the epistemic level of empirically (or otherwise) supporting assumptions ***P*** pertaining to ***R***, some qualifications are needed to make more precise what, exactly, the assumptions ***P*** demand. In accordance with the analysis developed in *Chapter 2*, we can distinguish between differences and similarities/identities between populations at three levels: 1) the basic structure of the causal mechanisms governing the outcomes of interest, i.e. the features that determine *whether X* is causally relevant for *Y*, 2) the functional form of causal relationships and the parameters that figure in these mechanisms and that capture *how X* is causally relevant for *Y* (e.g. what is the marginal effect of a unit increase in *X* on *Y*), and finally 3) the distributions of variables that figure in the causal mechanisms, such as the distributions of support factors. With this in mind, let us take a look at the first case, where P2 asserts that populations are identical in causal mechanisms.

P2 only requires that populations are identical at the first two levels, but remains open to (some) differences at the third level. This is easy to see: the weakest construal of the effectiveness argument permits that the support factors necessary for an intervention to be effective are only realized for some individuals in the target, so it permits that the support factors are differently distributed. However, at least on the first reading of the effectiveness argument, for this to yield the conclusion that an intervention will be effective in the target, we still require that these support factors are also support factors of the effect in the target, not just in the experiment.

Moreover, the particular way in which these support factors are involved in the production of the effects of interest must be the same or similar. Specifically, if we are interested in quantitative conclusions, such as ensuring that an intervention will have such-and-such effect for at least some individuals in the target, we need to ensure that the causal mechanisms in both populations are identical up to the level of the parameters and functional form relationships involved in the causal pathways from *X* to *Y*, and especially those governing the interaction between support factors and the treatment variable. Likewise, if we are interested in qualitative conclusions, the same qualitative relationships must hold between support factors and treatment variables in

producing the outcome of interest. In both cases, this can only be guaranteed by placing yet further constraints on $\boldsymbol{R}$ with respect to the parameters and functional form that govern the individual causal relationships constituting the mechanism that governs Y in both populations. They need to be qualitatively or quantitatively identical (or, in the latter case, at least highly similar).

All of these assumptions are demanding and supporting them can raise concerns about the extrapolator's bind. To support P2, we will need to learn something about the mechanisms governing the outcomes of interest in both populations. As Cartwright recognizes, even learning the mechanism governing the X-Y-effect in the experiment will often be difficult (see e.g. Cartwright 2013b, 100). Importantly, experimental data from RCTs themselves will not shed light on questions concerning the structure of causal mechanisms or the path-specific parameters and functional form involved in governing an effect, nor will they tell us anything about what support factors were involved in producing the effect, how they were distributed in the experiment, or how this distribution played a role in bringing about the observed effect.

How do we acquire the information that is required for elucidating these issues? In many realistic cases, where strong background theory speaking decisively to these issues is not available off-the-shelf, supplementary analyses need to be carried out. For instance, issues concerning the structure of causal mechanisms can be clarified by methods such as *process tracing* (Beach and Pedersen 2016), *causal discovery* from observational data (Spirtes et al. 2000), and *qualitative comparative analysis* (Beach and Pedersen 2019). With some understanding of the structure of causal mechanisms in place, econometric methods can be used to estimate parameters and functional form for path-specific effects. Finally, no particularly sophisticated method is needed to measure variable distributions, at least for observable variables, but measuring the right variables, i.e. support factors and causally relevant moderators and mediators, is of course crucial, and a great deal of understanding of the causal mechanisms governing the outcomes of interest will be needed to do so (cf. Cartwright 2013a, 15; see also Muller 2013, ms.; 2014; 2015).

Even if all of these steps were successful, however, in order to support the assumptions outlined above, we will often still need to compare what we have learned about the experimental population with what is the case in the target. In the language of the analysis provided in *Chapter 2*, we need to get a handle on the relation $\boldsymbol{R}$ between

experimental and target populations, i.e. instances of similarities and differences in the causal features that matter for the production of the effects of interest.

This is where things will often get thorny in practice. Even relatively complete knowledge of the causal makeup of an experimental population will often only provide us with hypotheses about what similarities and differences between populations might be important, but will not, by itself, help settle issues about whether such similarities and differences obtain. For that, very often, a look at the causal makeup of the target will be required. The crucial challenge here will be to obtain such information without falling prey to the extrapolator's bind.

To see just how difficult this can be, consider again P2, which requires that the intervention of interest on X can, at least under some conditions, induce changes in Y. Learning this can be easier and more difficult. In the best-case scenario, strong background theory or causal generalizations are available. Such generalizations would need to assert the causal relevance of X for Y in a broad range of cases, and the target must be uncontroversially understood to be among those cases. Think for instance about generalizations such as "microfinance availability can help people out of poverty", "distributing free bed nets can help decrease malaria infection", or "reducing class sizes can increase student performance".

If the scope of such generalizations is wide enough to cover the target, they can help us support that X can be causally relevant for Y in the target without requiring a detailed look at its causal makeup. All we need to do is affirm that the target is among the cases covered by the generalization, e.g. because it is a member of some well-understood type of population, which ensures, or makes it otherwise highly likely, that the outcomes of interest there are governed by the same causal mechanisms as in the experimental population (see e.g. Beach and Pedersen 2019, Ch.4 for methodological suggestions for how causally similar populations may be identified by type-membership).

However, well-supported generalizations with well-defined extensions are rare in social sciences. What is more, the generalizations that are needed for supporting P2 would also need to be significantly more precise than the toy generalizations suggested above. It may not be enough to learn that X can be causally relevant for Y in the target in *some* way, as otherwise it may remain unclear whether its effects on Y are governed by the same pathways in the same way and including the same support factors, or whether there are unanticipated differences in these respects that may pose obstacles to

successful extrapolation (see e.g. Beach and Pedersen 2019, Ch.4 and 8 for comments on tensions between generality and precision in building causal generalizations).

What, then, should we do in more realistic cases where generalizations covering both populations are either not available, or not sufficiently informative? Information obtained from the target itself may be a good guide for supporting P2, but learning whether X can be causally relevant for Y in the target from scratch can easily raise concerns about the extrapolator's bind. This is most obvious in cases where the intervention of interest has not yet been experienced in the target. Here, the concern is that if we cannot observe the mechanism governing the effects of interest 'in action', it will be difficult to learn features of this mechanism that are needed to compare it to the mechanism in the experimental population.

How, for instance, could we provide an assessment of whether a job market training programme can be efficacious in increasing employment outcomes in a novel target by increasing applicants' CV quality and interview skills, if there are neither observational data nor experience reports about the sorts of changes in applicants' CV quality and interview skills that the intervention would seek to induce or the envisioned changes in the employment outcomes of interest.

We might get luckier in cases where the intervention variable of interest regularly experiences natural, endogenous changes in the target. In these cases, the same kinds of analyses that can be used to learn something about the causal mechanisms in the experimental population could also be used in the target in order to acquire information pertinent to answering our questions about causally relevant similarities and differences. If, say, a microfinance intervention seeks to increase welfare outcomes through increasing spending on durable goods, it might be easier to tell whether this intervention may in principle be effective by means of statistical and econometric analyses of natural (co)variation in spending on durable goods and the welfare outcomes of interest, as well as (co)variation in other suspected moderating and mediating variables and support factors. Such analyses could help us tell (although not conclusively) whether the effects of interest may be transmitted along the same pathways and affected by the same moderating and mediating variables and support factors. Similarly, process tracing and qualitative comparative analysis may also be helpful tools for this purpose (Schmitt and Beach 2015; Beach and Pedersen 2016; 2019; Beach 2017).

However, even if we get lucky enough to have large and informative datasets that

include relevant variation in the variables of interest, for the results of such analyses to be useful in supporting our assumptions we will always need to make the further substantive assumption that the causal features learnt by such analyses are invariant under the intervention of interest. Sticking with the example, naturally occurring differences in spending on durables may not have the same effects on welfare outcomes as differences that are induced by a microfinance programme. This is just a familiar concern about *structure-altering interventions* (cf. Steel 2008, ch.8; Lucas 1976), where we must assume, for instance, that the basic structure, functional form, and parameters of the mechanisms governing household welfare outcomes do not change with respect to *how* variation in spending on durables is induced. If, for instance, individuals would spend their money radically differently depending on whether it is obtained as a microloan or obtained through wage labour, then natural variation obtained from populations where individuals do not (yet) take out loans, but earn their endowment through labour, might be a poor guide for learning what will happen if they were to be exposed to microfinance products.

Importantly, trying to lay such concerns about structure-altering interventions aside will require an even deeper look into the causal makeup of the target, including specifically information pertaining to whether the intervention will be structure-altering or not, and if so how. Such information, almost by definition, will require that the intervention of interest be implemented in at least a sample from the target. This in turn, however, would clearly fall prey to the extrapolator's bind as we might then trivially learn the effect of interest in the target. While we might still need to reason from a small sample of the target to a conclusion about the target population as a whole, this would turn our extrapolation problem into the potentially somewhat easier problem of generalizing from a sample to a larger super-population. In any case, it seems that the result obtained from the experimental population is rendered largely redundant to clarifying whether the intervention will be effective in the target.

Similar concerns apply to P3, i.e. issues of whether the support factors necessary for *X* to make a contribution to *Y* are in fact instantiated in the target. The general problems here should be clear by now. If P3 is understood as saying that the support factors necessary for *X* to make a contribution to *Y in the experiment* are realized in the target, this is not enough to infer a conclusion about the target, and we need to further assume that the same support factors are involved in the production of the effects of interest and

in the same way. This just triggers largely the same concerns about the extrapolator's bind as discussed above. The same is true if P3 demands that the support factors necessary for *X* to make a contribution to *Y* in the experimental *and* target population are realized in the target as well. Here, too, extensive causal knowledge spanning both populations may be needed.

What deserves separate discussion is the construal P3.1 according to which the support factors necessary for *X* to make a contribution to *Y in the target* are realized there. This construal makes my concerns about the extrapolator's bind even more vivid. Arguably, learning what support factors are required for *X* to make a contribution to *Y* in the target requires causal knowledge pertaining to the target, and very often knowledge that needs to be obtained *from* the target. This is largely analogous to the above concerns. However, it seems that the construal P3.1 is special in the sense that validating it empirically would make the other premises of the effectiveness argument redundant to its conclusion in an even more straightforward way. The reason is that in affirming P3.1, we already presuppose that *X* can make a contribution to *Y* in the target, so neither P1 nor P2 is needed to infer the conclusion of the effectiveness argument. If we learn which support factors are necessary for *X* to make a contribution to *Y* in the target, and these factors are indeed present in the target for some individuals, as P3.1 asserts, it follows that the intervention can be effective for these individuals from P3.1, and the knowledge used to support it, alone.

Again, while knowledge from the experimental population may still be relevant for inspiring hypotheses about what might be support factors of the effect of interest in the target, we still need to validate whether these candidates are in fact support factors, and this will require information pertaining to the target, and potentially pertaining to the target alone. So while the experimental evidence and other information from the experimental population may be relevant for the *discovery* of candidate premises about support factors in the target, they are not relevant for the *justification* of these premises.

This makes clear that in the absence of strong background theory or generalizations that are justifiably believed to cover both the experimental and the target population, and where validating the premises of the effectiveness argument mostly demands that we assert that the target is among the cases covered by the available generalizations, supporting the effectiveness argument's premises is likely to fall prey to the extrapolator's bind; at least unless some persuasive strategy is provided that helps evade

this problem.

So far, I have raised some critical concerns about the scope of the effectiveness argument, as well as the epistemic issue of validating its premises pertaining to identities in causal mechanisms (including issues of structure, functional form, and parameters) and how support factors are distributed in the target.

I will now turn to more constructive contributions. One involves a detour and concerns Cartwright's concept of causal support factors, and how this concept extends importantly beyond the standard conceptions of moderating and mediating variables discussed in *Chapter 2*. In *Appendix 1* to this chapter, I will engage in some conceptual gardening by proposing different ways of understanding support factors, as well as of conceptually integrating them with moderating and mediating variables. Readers who are in a hurry, however, may proceed to the next section straight away. There, I briefly anticipate some objections and proceed to a positive outlook on what the Argument Theory contributes to addressing problems of extrapolation in practice.

4.4 What's the Argument Theory after all? Objections, Replies, and a Way Forward

The previous discussion suggests that the effectiveness argument is importantly limited in scope as well as difficult to support empirically without falling prey to the extrapolator's bind. In this section, I anticipate two potential objections to these criticisms, which raise broader questions about what kinds of contributions Cartwright's Argument Theory makes.

First, Marcellesi (2015), in a paper arguing that the problem of extrapolation (or 'external validity') has been solved, anticipates the concerns about the extrapolator's bind I have developed above and argues that Cartwright's Argument Theory remains untouched by such concerns because it is not supposed to offer a *method* for how to extrapolate, but rather a general *analysis* of the abstract conditions under which successful extrapolation is feasible. Since the latter is not intended to offer concrete guidance for how to extrapolate, it remains unaffected by the extrapolator's bind, as the bind only pertains to epistemic demands that in fact obtain when a specific method is adopted and used. So while concerns about the extrapolator's bind might apply to the

effectiveness argument as discussed above, they would leave the Argument Theory largely untouched.

I have two replies to this objection. The first is that Marcellesi's interpretation of Cartwright's contributions is lacking in support. Many aspects of how Cartwright presents her contributions speak in favour of a more practice-oriented interpretation on which the Argument Theory, at least as concretised in the form of specific effectiveness arguments, is (among other things) intended to inform users of effectiveness evidence about how to extrapolate results to their intended domain of application.[4] On this reading, the effectiveness argument is not merely an illustration of what the Argument Theory can tell us about extrapolation at the most abstract level, and one that may be easily sacrificed in light of the arguments developed here.

My second reply is that if Marcellesi's interpretation were right, it would seem to trivialize the Argument Theory. Claiming that the problem of extrapolation has been 'solved' by an abstract account that, when concretised, may provide us with effectiveness arguments whose premises are difficult to support without falling prey to the extrapolator's bind, misses the point of what extrapolation is ultimately about: *overcoming* the real-world epistemic difficulties involved in extrapolation. Marcellesi's interpretation of Cartwright's account would undermine its contributions towards addressing real-world problems of extrapolation, and would leave us with a compelling but not practically useful contribution pertaining to extrapolative inference, i.e. that we cannot justifiably reach extrapolative conclusions about causal effects in a target unless we can validly infer such conclusions by means of *some* argument.

I maintain that the extrapolator's bind remains a serious challenge for the Argument Theory. It does so as long as it is not further elaborated how the premises that support the conclusions of effectiveness arguments, whether in the form of the particular argument supplied by Cartwright, or indeed any other argument, can be established in a way that does not trigger concerns about the extrapolator's bind.

The second objection I want to consider here is closely related to the first and argues that my characterization of what the Argument Theory aims to contribute is perhaps too limited and that my verdict about its limitations in scope is too hasty. Specifically, one may object to my concerns about scope that the specific effectiveness argument

[4] This is in line, for instance, with how the effectiveness argument figures in Cartwright and Hardie's (2012) *Evidence-Based Policy: A Practical Guide to Doing it Better.*

provided by Cartwright is merely the most minimal characterization of what an effectiveness argument could look like, and that there are other conceivable characterizations of such arguments that may be able to address the kinds of extrapolative queries that I worry cannot be adequately addressed by the exemplary effectiveness argument. On this characterization, the Argument Theory is open to other kinds of effectiveness arguments, too, including perhaps arguments that allow us to answer a broader range of causal queries.

I agree that other arguments, perhaps at some price to tractability, could be able to address more intricate causal queries, including those concerning cases where there are causally relevant differences between populations. At the same time, this understanding of the Argument Theory would still leave open just what form these more sophisticated arguments would take. My concern here is that while it may be feasible to develop other valid effectiveness arguments that characterize the conditions under which extrapolation can proceed successfully even in the presence of causally relevant differences between populations, these arguments may need to draw to a significant extent on substantive theory borrowed from elsewhere, e.g. causal graph or econometric theory that also underlies other strategies for extrapolation such as those by Bareinboim and Pearl (2013), Hotz et al. (2005), and Muller (2014; 2015). It seems that there are only so many ways in which one can characterize the conditions under which we can predict causal effects in a target despite causally relevant differences. Indeed, at least in some cases, the results provided by these strategies seem to converge already.

So if we were to construct effectiveness arguments that could help us extrapolate in the presence of causally relevant differences as well, my concern would be that these arguments are bound to invoke assumptions and underlying theories similar to those used by other strategies, in order to arrive at results similar to those that these other strategies already provide. The latter, arriving at the same or similar results, would, of course, be a virtue of our newly constructed, more sophisticated effectiveness arguments. The former, however, invoking similar assumptions and characterizing similar conditions on the relation ***R*** between populations under which extrapolation in such-and-such a way is feasible, raises an important concern. If we were to draw heavily on the substantive causal graph or econometric theory underpinning other strategies for extrapolation, it would seem that it is this theory, instead of the Argument Theory itself, that would be doing the important work of characterizing the conditions

under which, as well as the specific ways in which, extrapolation is feasible. While the Argument Theory would still add particular and characteristic emphasis on the importance of making the assumptions required for the application of such theories explicit, i.e. preferably in the form of valid and sound arguments, it is unclear what contribution the Argument Theory would make towards overcoming problems of extrapolation aside from adding such emphasis.

My proposal for reconciling the concerns outlined above is to suggest that the Argument Theory should be understood as aiming to provide an account of how to *underwrite* extrapolative inference, i.e. an account not of how to construct valid arguments but of how to render them sound, and of what challenges are involved in doing so. This is specifically in contrast to proposing a specific method for how to extrapolate, at which the exemplary effectiveness argument is unsuccessful, or conceiving of the Argument Theory as an abstract proposal that does not offer much towards overcoming real-world problems of extrapolation, as per Marcellesi. On my understanding, the Argument Theory would not aim to offer any *specific* method for extrapolation (e.g. in the form of a specific argument), but rather focus on the ways in which, given different methods, and given a variety of effectiveness arguments that can be constructed by drawing on such methods, one should go about supporting the assumptions that these arguments involve.

Some contributions towards offering such an account are already provided by the Argument Theory: First, it explicitly models the activity of providing support for the assumptions involved in effectiveness arguments as an *inductive* endeavour. This is important, as it seems plausible to think that the extrapolator's bind can be best evaded if support for assumptions such as similarity of experimental and target populations is not understood as an all-or-nothing package deal, where 'all' is too demanding, and 'nothing' is insufficient to get extrapolative inference off the ground. Instead, and specifically with a view towards evading the extrapolator's bind, it seems plausible to think that supporting evidence that helps build a basis for extrapolation must be gathered and used in a cumulative and incremental fashion, where the support for each assumption comes in parts and each part can offer support to different degrees, and in different ways. In virtue of its emphasis that support for premises comes inductively, and in degrees, the Argument Theory already offers a useful starting point.

The Argument Theory, in its current form, also provides an important cautionary

message. Conscientious extrapolation requires that one arrives at the desired conclusion by way of valid inference, by making the assumptions and the evidence required to support them explicit and doing so preferably in the form of explicit, valid, and sound effectiveness arguments. This is not just important for cases where extrapolation can proceed successfully. Even if successful extrapolation is infeasible, the Argument Theory can help us recognize why, by emphasizing just how challenging conscientious and rigorous extrapolative inference can be. In doing so it can also help instil a sense of epistemic humility in producers and users of evidence that have not yet appreciated the intricacies involved in extrapolation. This cautionary message is similar to comments offered by Bareinboim and Pearl (2013) on their graph-based approach. They argue that one of the distinctive advantages of using graphical causal models is that it makes explicit what causal knowledge we require for extrapolation, and in the process of this, helps realize that we may often possess very little of such knowledge. This, too, is supposed to make us more sensitive to the limits of our ability to justify extrapolation. In a somewhat cynical twist, this may be just one of the very insights that these strategies are supposed to supply: extrapolation is a difficult endeavour, and in some cases it will remain insurmountably difficult.

An updated Argument Theory augmented by the discussion here can add more nuance to this cautionary message: even when extrapolation is feasible by means of adequately supported effectiveness arguments, successful extrapolation requires that the conclusion of interest be reached without falling prey to the extrapolator's bind. The Argument Theory can alert us to the possibility that this is not always feasible, and tell us why this is so: because the information from the target used to support the premises of an effectiveness argument can render its premises redundant to its conclusion.

Finally, a third important contribution that the Argument Theory makes is conceptual. The concept of *causal support factors* provides an intuitively graspable way of thinking about potential obstacles to extrapolation. As suggested in *Appendix 1*, this concept may be further supplemented by a more comprehensive analysis. This analysis of support factors, and how they relate to concepts invoked in the broader literature (specifically moderators and mediators) can offer additional resources that so far remain unavailable on other approaches such as Bareinboim and Pearl's (2013), Muller's (2014; 2015) and Steel's (2008) which 1) do not offer the ability to represent sufficient cause scenarios (or at least not without introducing further complications, e.g. on

Bareinboim and Pearl's account), 2) focus predominantly on moderators (Hotz et al. 2005; Muller 2014; 2015) or 3) simply bracket cases where these types of variables play an important role (Steel 2008). Of course, to be useful for overcoming concrete problems of extrapolation, we also need a guide for identifying support factors (further building, e.g., on the ideas developed in Cartwright and Hardie 2012), for learning about how they bear on effect magnitudes, and for using information about how they are realized in experimental and target populations respectively; all with a view towards evading the extrapolator's bind. I will return to some suggestions pertaining to these issues in *Chapter 8*.

For now, it is enough to note that by drawing on the resources already provided by the Argument Theory, as well as perhaps complementing it in some respects suggested here, it can be useful in at least three important ways: 1) it can help scrutinize existing strategies for extrapolation with a view towards the assumptions they involve, 2) it can complement and unify their conceptual arsenal by providing a more comprehensive analysis of causal support factors, and 3) it can provide a general framework for thinking about how to justify extrapolation in a way that evades the extrapolator's bind.

4.5 Conclusions

I have argued that Cartwright's effectiveness argument, understood as an illustration of how the Argument Theory may be applied, is limited in scope. It can only address a highly restricted class of extrapolative queries and does not have the resources to address what I consider to be the most important class of queries, i.e. those pertaining to causal effects in the presence of causally relevant differences between populations. Moreover, even for the limited class of queries that the effectiveness argument can address, once its premises are concretised, it remains unclear whether these premises can be supported without falling prey to the extrapolator's bind. Despite the effectiveness argument being perhaps best understood as a mere illustration of the Argument Theory, it was a useful exercise to consider how even relatively simplistic arguments can already raise important concerns about the extrapolator's bind. More ambitious extrapolations, including extrapolations addressing the more important class of queries pertaining to causal effects in the presence of causally relevant differences, will require more complicated arguments drawing on substantive theory that can help us accommodate and adjust for causally relevant differences between populations. These,

in turn, will involve even stronger assumptions that amplify concerns about the extrapolator's bind.

As I have argued, recognizing these limitations does not undermine the usefulness of the Argument Theory. The Argument Theory, by itself, may not provide theoretical resources and empirical strategies to construct and justify more sophisticated effectiveness arguments; these might need to be obtained from elsewhere. However, if and when available, the Argument Theory can be a useful tool for scrutinizing extrapolative inference that draws on these resources.

Following this positive outlook, the subsequent chapters can be broadly understood as an attempt to put the Argument Theory to use for critically evaluating other strategies for extrapolation. In line with my proposal for how we could understand the Argument Theory going forward, these critical investigations can be understood as an attempt to consider how the resources provided by other strategies might be used to construct more broadly useful and sophisticated effectiveness arguments. In the spirit of the Argument Theory, my main aim will be to make the assumptions that they involve explicit and to scrutinize these assumptions with a view towards whether and how they can be empirically supported without falling prey to the extrapolator's bind.

References

Bareinboim, E. and J. Pearl. (2012). "Transportability of causal effects: Completeness results", In: Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12), Menlo Park, CA.

— **(2013).** "A general algorithm for deciding transportability of experimental results", *Journal of Causal Inference*, 1: 107-134.

Beach, D., and R. B. Pedersen. (2016). *Causal case studies: Foundations and guidelines for comparing, matching and tracing*. Ann Arbor: University of Michigan Press.

— **(2019).** *Process-Tracing Methods - Foundations and Guidelines*. 2nd edition. Ann Arbor: University of Michigan Press.

Beach, D. (2017). "Process-tracing methods in social science", Oxford Research Encyclopedia of Politics. Retrieved 18 February, 2019 from http://politics.oxfordre.com/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-176

Cartwright, N. D. (2009). "Evidence-Based Policy: What's to Be Done About Relevance", *Philosophical Studies*, 143(1): 127-36.

— **(2011)**. "Predicting 'It will work for us': (Way) beyond statistics". In: F. R. Phyllis McKay Illari, and Jon Williamson (Ed.), *Causality in the Sciences*. Oxford: Oxford Scholarship Online.

— **(2012)**. "Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps", *Philosophy of Science*, 79: 973-89.

— **(2013a).** "Evidence, Argument and Prediction". In: V. Karakostas, and D. Dieks (eds.), EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings. Cham, Switzerland: Springer International Publishing Switzerland.

— **(2013b).** "Knowing what we are talking about: why evidence doesn't always travel", *Evidence & Policy*, 9(1): 97-112.

Cartwright, N. D., A. Goldfinch, and J. Howick. (2009). "Evidence-Based Policy: Where Is Our Theory of Evidence?", *Journal of Children's Services*, 4(4): 6-14.

Cartwright, N. D., and J Hardie. (2012). *Evidence-Based Policy: A Practical Guide to Doing it Better*. Oxford: Oxford University Press.

Cartwright, N., and J. Stegenga. (2011). "A Theory of Evidence for Evidence-Based Policy" *Proceedings of the British Academy*, 171: 289–319

Hotz, V. J., G. W. Imbens, and J. H. Mortimer. (2005). "Predicting the efficacy of future training programs using past experiences at other locations", *Journal of Econometrics*. 125: 241–70.

Lucas, R. (1967). "Econometric Policy Evaluation: A Critique". In K. Brunner, A. Meltzer (eds.), *The Phillips Curve and Labor Markets*. pp. 19-46. Carnegie-Rochester Conference Series on Public Policy. New York: Elsevier.

Mackie, J. L. (1965). "Causes and conditions", *American Philosophical Quarterly*, 2: 245–64.

Marcellesi, A. (2015). "External Validity: Is There Still a Problem?". *Philosophy of Science*, 82(5): 1308-17.

Muller, S. M. (2013). „External validity, causal interaction and randomised trials: the case of economics", Unpublished manuscript.

— **(2014).** "Randomised trials for policy: a review of the external validity of treatment effects". Southern Africa Labour and Development Research Unit Working Paper 127, University of Cape Town.

— **(2015).** "Interaction and external validity: obstacles to the policy relevance of randomized evaluations", *World Bank Economic Review*, 29(1): 217-25.

Schmitt, J., and D. Beach. (2015). "The contribution of process tracing to theory-based evaluations of complex aid instruments", *Evaluation*, 21(4): 429–47.

Spirtes, P., C. N. Glymour, and R. Scheines. (2000). *Causation, Prediction, and Search*. 2nd edition. MIT Press, Cambridge, MA.

Steel, D. (2008). Across the boundaries: Extrapolation in biology and social science. Oxford: Oxford University Press.

Weinberger, N. (2014). "Review: Evidence-Based Policy: A Practical Guide to Doing it Better", *Economics and Philosophy*, 30: 113-20

APPENDIX 1

What are Support Factors?

In *Section 3* I have discussed causal support factors as an important ingredient of Cartwright's contributions towards clarifying the conditions under which some kinds of extrapolation are feasible, as well as highlighted their role in the effectiveness argument. I have also suggested that the effectiveness argument involves substantive assumptions not only about support factors, but also about moderating and mediating variables. As discussed in *Chapter 2*, moderating variables are frequently invoked when modelling situations where the causal effect of X on Y depends on the value of another variable W. There, I have also emphasized that, although so far unrecognized, some kinds of mediating variables can play similar roles. Moderating and mediating variables are important for extrapolation because differences between populations in their distribution can pose obstacles to extrapolation, much like causal support factors as Cartwright discusses them. What remains unclear so far is how these different types of variables relate to each other. There are several questions to be clarified, including: are all moderating variables support factors? Are all support factors moderating variables? Are there different types of support factors, and if so, which? If there are different types, are all of them equally important for extrapolative inference?

In this appendix, I aim to make progress on these questions. I begin from an abstract conception of support factors and subsequently concretise it in order to focus attention on a number of important distinctions. This exercise in conceptual gardening will be useful for building a more comprehensive and detailed analysis of support factors and how they relate to moderating and mediating variables.

The abstract conception I begin with, and the one that Cartwright seems to endorse (see e.g. Cartwright and Stegenga 2011, 301), is the following:

W is a support factor for the causal effect of X on Y if there is a causal complex C that is sufficient for the production of Y and X and W are INUS conditions in C.

This captures the key intuitions that 1) causes are best understood as INUS conditions, 2) interventions rarely work on their own, and 3) support factors play an important role in extrapolating from one population to another, which may exhibit different realizations/distributions of support factors. At the same time, this abstract

conception still glosses over various more fine-grained distinctions, which I elaborate below.

First, support factors can be understood as *enabling conditions* in a causal complex C that is sufficient for Y. Specifically, on this understanding a binary support factor $W \epsilon [0; 1]$ needs to be instantiated ($W = 1$) for an intervention on X to have any effect on Y. This can be modelled as:

$$Y = \beta * X * W \; ; \; W \epsilon [0; 1]$$

Here, the causal effect of X on Y will be zero at $W = 0$ and β otherwise.

Second, we can imagine cases where there are larger ensembles of support factors in the causal complex C. Consider an n-ensemble of binary support factors $W_n \epsilon [0; 1]$ such as usually found in toy examples from epidemiology that characterize sufficient cause scenarios, i.e. cases where *all* components of a causal complex C need to be jointly realized for an intervention on X to be efficacious in inducing changes in Y. We can characterize this by

$$Y = \beta * X * \min\{W_1, W_2, \dots, W_n\}$$

So the effect of X on Y will be zero if any $W_n = 0$, and β otherwise.

Third, an important variation on these all-or-nothing construals is to understand support factors as a *threshold* INUS conditions for Y. Here, we may have a continuous support factor W, where the effect of X on Y will be zero at any level of W below a threshold λ, and β otherwise. We can model this with the help of an auxiliary variable Z as:

$$Y = \beta * X * Z \; ; \; Z = \begin{cases} 0, & W < \lambda \\ 1, & W \geq \lambda \end{cases}$$

We can also think of further, arbitrary variations where only some, but not all, factors in a causal complex need to be realized (or realized above/below some threshold level) for an intervention on X to be effective. Here, specific combinations of factors might be able to enable the effects of X on Y in different ways, e.g. W and Z might be sufficient for X to be effective, but only W in the absence of Z would also require P, Q, R, and S.

These three ways to understand support factors cover the conception of support factors in the traditional sense of *enabling conditions* that help model sufficient cause

scenarios where X is one among several distinct components of a causal complex C for Y, and where changes in X do not effect changes in Y unless other components of the causal complex are realized in specific ways (e.g. realized at all, or realized above some threshold level).

A fourth important sense in which to understand support factors is to understand them as *moderating variables* in the traditional sense discussed in *Chapter 2*. Here the idea is that W is a support factor in the sense that different levels of W induce different magnitudes of the causal effect of X on W, so that, for instance, higher values of W can help one and the same intervention on X make a larger contribution to Y than lower values of W. Importantly, this characterization would not conform to the abstract conception that I started from, where support factors are *necessary* for X to make a contribution to Y. This is because understanding support factors as moderating variables W means that they might not be *necessary* for X to make a contribution to Y in any strict sense (for instance, there might be no interesting way in which W can be 'absent', e.g. age). Still, W may need to be realized at or above certain levels for X to make a particular contribution to Y and can, in this sense, be an important prerequisite for an intervention to be effective in a particular way. This can be modelled in different ways. For instance, a fully interactive case can be modelled as:

$$Y = \beta * X * W$$

Here, both X and W are moderators of each other's effects on Y and hence on equal footing in producing changes in Y. In this case, W can still be necessary for X to make a contribution to Y, because any such contribution would be precluded at $W = 0$.

This is also true for variations on this fully interactive form where a causal effect of X on Y varies non-linearly over W. Here, generally:

$$Y = \beta * X * f(W)$$

for some $f(W)$ that is non-linear in W, such as when:

$$Y = \beta * X * \gamma^{W}$$

Whether W (or a specific value of W) is necessary for X to make a contribution to Y changes in *partially interactive* cases. Here, there is a baseline effect of X on Y that obtains no matter the level of W and Y is co-determined by this baseline effect as well as an additional, W-specific contribution that is produced in interaction with W, such as

when

$$Y = X * (\beta + \gamma W)$$

Here, it is important to recognize that W is not *necessary* for X to make a contribution to Y since there is a baseline marginal effect of X on Y, β, that will obtain no matter the value of W. Strictly speaking, this case would hence fail the abstract conception of support factors as INUS conditions espoused by Cartwright, which requires that W is *necessary* for X to make a contribution to Y.[1]

It is unclear what proportion of real-world cases are best modelled in such a way, or whether it is more common to find fully interactive cases (including non-linear variants thereof). But to the extent that the above cases are not entirely unrealistic, it may be useful to extend our understanding of support factors and say that support factors are not always INUS conditions, but extend beyond these to include cases where a support factor is not necessary but still important for whether an intervention on X yields (certain kinds of) changes in an outcome Y.

This extended view would also help capture an important feature of fully interactive moderating variables. Above, I have considered fully interactive moderating variables as support factors in the enabling-conditions sense because they have the capacity to curtail an effect of X on Y, e.g. if $W = 0$ in $Y = \beta * X * W$. But W is not only important in this sense, but also, and perhaps more significantly, in the sense that it has important bearing on the magnitude of an effect. Put differently, a good part of the importance of moderators for issues of extrapolation is constituted not by their regularly exercised ability to completely suppress causal effects of X on Y, but by importantly determining the magnitude of the effect of X on Y. To capture the importance of this ability, we may hence extend our conception of support factors to capture variables that exhibit this feature, including variables that only exhibit this feature, without being in any important or regularly exercised sense necessary for a causal effect.

So far, I have offered two broad conceptions of support factors: understanding them as *enabling conditions*, i.e. conditions that, in different ways, enable interventions on X

[1] Mackie's (1965) formulation of the INUS concept states that an INUS condition is W '*nonredunant*' instead of 'necessary'. The former simply means that W and its contribution to a sufficient causal complex for Y cannot be trivially substituted by any other factor, except in cases where there are yet other sufficient conditions for Y that involve all the same conditions except W and W is replaced by some other condition Z. This is covered, however, by the fact that both of these causal complexes themselves are unnecessary but sufficient for Y. We may hence think of non-redundancy as necessity. Cartwright, too, seems to subscribe to this understanding (Cartwright and Stegenga 2011, 301).

to make a contribution to Y; and understanding support factors as different varieties of *moderators*. Both conceptions have a crucial commonality, namely that support factors are at least relevant, and at most necessary, for whether X can induce *changes* in Y.

Highlighting this is useful for pointing out an important vagueness in the general terminology surrounding causes as INUS conditions. They are characterized as INUS conditions *for* something (e.g. contributions, as on Cartwright's account). But what is that 'something', specifically? In the above cases, it pertains to *changes* in an outcome variable with respect to *changes* in an intervention variable. Support factors in both senses are support factors that interact (partially or fully) with the causal effect of X on Y; they can enable and suppress it dichotomously, and they can modify it gradually (including partially).[2]

However, there is a third sense in which we can understand support factors, namely as factors W that help achieve specific *levels* of an outcome (rather than specific changes in such levels), but do not interact with X, so different realizations of W do not induce different changes in Y induced by given changes in X. On this understanding, W is simply an additively separable cause of Y. Such a cause can be an INUS condition for a specific *level* of Y, for instance, if there are levels of Y that cannot be achieved by any intervention on X under all possible realizations of W.

To give an example, consider the effect of class size on student achievement (I will expand on this example in more detail shortly). Let us assume that the smaller a class, the more pronounced the effects of teaching on student achievement. Let us also assume that this effect is limited by the lower bound of X where there is only one student in a class. Even at this lowest bound, where, let us assume, the effect of an additional hour of teaching on student achievement is most pronounced, the student's educational outcomes might still be even further improved by additional interventions on other variables that contribute to her achievement level Y, such as teacher quality, supplementary teaching, parental support, and others. Conversely, certain levels of Y might be infeasible if variables such as class size and teacher quality do not simultaneously assume certain ranges of values.

Importantly, however, despite two or more variables such as class size and teacher

[2] Cartwright and Stegenga (2011, 306) ask and seek to clarify the same question. They do so differently, however, by expanding on the distinction between support factors for dichotomous and multi-valued *effects*. I am interested in the distinction between support factors for *levels* of an outcome vs. *effects* on an outcome.

quality jointly contributing to the achievement of a particular level of an outcome, this does not mean that there is any interaction between them. So the marginal causal effects of class size on student achievement, i.e. the changes in Y with respect to changes in X, do not depend on the values of teacher quality W. This gives us a *non-interactive* conception of INUS conditions. On this understanding, different parts of causal complexes join together to produce certain *levels* of an outcome, and some of these levels may be unattainable for any intervention on some part of the causal complex (e.g. any intervention on X), unless other parts of the complex (e.g. W) assume suitable values as well. But over and above the necessity of certain kinds of joint contributions to achieve a specific *level* of an outcome, there is no interaction between the constituent INUS conditions of the causal complex C for Y, and the individual contributions (i.e. marginal effects) of these causes on the outcome do not depend on the values of other variables in the causal complex. The causal pie, as it were, is a mere sum of its ingredients, unlike on interactive conceptions, where the pie is more than that and cannot be neatly decomposed (when the rules of composition are unknown).

The distinction between interactive and non-interactive support factors can have important ramifications in practice. Cartwright invokes various test cases to illustrate the importance of support factors for purposes of extrapolation, specifically by illustrating what happens when these support factors fail to be in place in the target. A case repeatedly used by Cartwright is that of 'Project STAR', an educational intervention implemented in Tennessee to identify the effect of reducing class size on student performance (Bohrnstedt and Stecher, 2002). In this example, an intervention that decreases class sizes is demonstrably effective in increasing student performance in a study population in Tennessee. Yet, when this intervention is implemented in another population in California, it fails to bring about an analogous change in levels of student performance. One of the reasons that are typically cited for this failure is that the causal effect of class size on student performance depends on background characteristics of the setting, such as teacher quality. Specifically, implementing the intervention on a large scale in California involved general equilibrium effects that curtailed the success of the intervention. Decreasing class sizes increased the number of classes, and given a fixed teacher supply schedule that is downward sloping in teacher quality, demanding more teachers led to decreasing average teacher quality for employed teachers. According to Cartwright, teacher quality is one of the support factors that need to be in place for the intervention on class sizes to make the envisioned contribution to improving student

performance outcomes. As average teacher quality decreased by expanding the total number of classes, the intervention was not effective in increasing student performance when implemented in California.

So on Cartwright's account, teacher quality figures as an INUS support factor in at least one causal complex that is involved in the production of the outcome of interest. As Cartwright argues, and following the effectiveness argument, extrapolating from Tennessee to California can only be successful to the extent that 1) the treatment can play the same causal role in both populations and 2) that the support factors, such as teacher quality, are distributed similarly in both populations (e.g. the means of the distribution must be similar). Provided that we assume the first condition to hold, the failure of the intervention in California is then attributable to a failure in the similarity of the support factor distributions. Due to general equilibrium effects, post-intervention average teacher quality was not distributed in the right (read: similar) way in California to yield a similar effect of class-size reduction on student performance there.

It is important to recognize that understanding teacher quality as a support factor in the interactive or non-interactive sense can have different implications for how the effectiveness of the class size programme in California should be evaluated. To appreciate this, it is important to distinguish between different senses of effectiveness. One is in terms of outcomes: here, we ask whether a certain outcome of interest has been realized, say a specific level of student achievement. The other is in terms of contributions to an outcome. Here, we ask whether certain contributions have been made to an outcome, e.g. specific changes to student achievement.

If teacher quality were an interactive support factor, it would be correct to conclude that the intervention failed to be effective in the California setting in both senses of effectiveness. The general equilibrium effects on average teacher quality decreased teacher quality to such an extent that the marginal effects of the change in class size were zero. In terms of outcomes, too, the desired outcome has not been achieved and the intervention remains ineffective.

In the non-interactive case, this story changes. Here, there are two counteracting contributions to the outcome, both induced by the intervention. The first would be a positive contribution. For instance, at least some post-intervention classes would still plausibly end up with high-quality teachers. At least for these classes, it would seem reasonable to think that students benefited from smaller classes, and the intervention

was effective in the sense of making a positive contribution for students in these classes. However, there was also a second, negative contribution to the outcome. This negative contribution obtains as a result of the general equilibrium effects on teacher quality, which in turn makes a negative contribution to student achievement for those post-intervention classes with lower than pre-intervention quality teachers. Taken together, these two counteracting contributions would yield a net effect of zero and hence would yield the conclusion that the intervention was ineffective in the outcomes sense. But that does not mean that the intervention was ineffective in the contribution sense. It was effective, but in an unfortunate way that failed to yield the envisioned level of the outcome.

It is important to recognize that by inspecting pre/post-intervention observable differences in student achievement, as it was performed in the key evaluation of the California class size intervention by Bohrnstedt and Stecher (2002), it is difficult to disambiguate between the interactive and non-interactive ways in which teacher quality could have figured in the mechanism governing student performance. But despite empirical challenges, the distinction remains an important one. For one, a non-interactive understanding of teacher quality allows that the intervention had important distributive effects on student performance. Some students, those in post-intervention high-quality teacher classes, may have benefitted, while other students, those in post-intervention low-quality teacher classes, may have been made worse off. The interactive case, by contrast, would simply suggest that the intervention did not induce any (differential) contributions to the outcome in any population. This is how the distinction is important for policy evaluation.

The distinction between interactive and non-interactive support factors also has important implications for prediction and intervention design. Here, it is important to recognize that the California class size example is a special one. General equilibrium effects that obtain when scaling up interventions are a real concern, but not all interventions of interest in EBP are scaled up in such a way as to raise such concerns. In cases where such concerns are plausibly believed to be irrelevant, the situation is entirely different. Here, there will often be a marked difference between identifying a variable as an interactive or non-interactive support factor. In the former case, the support factor, if unsuitably distributed in the target, may completely suppress the contribution of an intervention to the outcome. Not so in the non-interactive case. Here,

unhelpful distributions of support factors may bear importantly on *where* the effects of interest materialize (e.g. at lower levels of the outcome variable), but will not bear on *whether* these effects materialize, or indeed on the magnitude of these effects.

Likewise, co-interventions on interactive support factors (such as decreasing class sizes and simultaneously investing in teacher training programmes) can drastically increase effect sizes yielded by a given intervention, but not if they are performed on non-interactive support factors. Again, all they will do is change *where* the effects of X on Y are materialized, but they will not change whether these effects obtain or indeed the magnitude of these effects. The causal pie remains a mere sum of its parts, and no disproportionately bigger pie may be expected if we meddle with the proportion of its ingredients. This suggests that the distinction between interactive and non-interactive support factors is an important one in many cases, both for evaluation as well as for informing prediction and optimal intervention design.

Let me briefly expand on two further ways that I consider to be important additions to a more comprehensive conceptual arsenal of what kinds of variables and causal arrangements may be considered to fall under the umbrella heading of support factors.

One interesting variant of support factors is what I call a *switching variable*. By this, I mean variables S that determine the relative marginal effects of changes in X on Y transmitted along multiple pathways. Essentially, they can be thought of as moderators of moderators. To give an example, consider a case where there are two pathways between an intervention and an outcome, where one pathway is moderated by W_1 with parameter γ_1 and the other is moderated by W_2 with parameter γ_2. *Figure 1* illustrates:

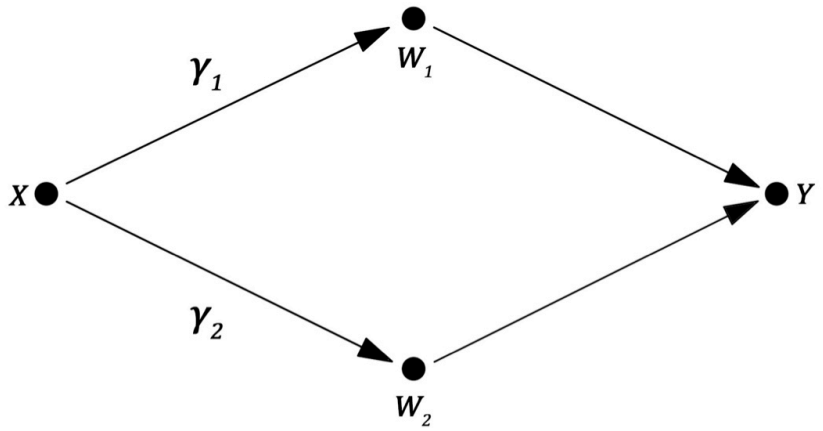

Figure 1: Two mediating paths, with path parameters determined by a switching variable

A switching variable S (not encoded in the graph) is a variable that induces changes

in γ_1/γ_2. For instance, consider a microfinance programme intended to increase private investment in durable goods in households of the rural poor. Let X be household endowment, the (indirect) intervention on X is making microloans available, and the outcome of interest Y is the share of private investment in durable goods. Suppose further that the effect of X on Y is mediated by two different variables: W_1 is the availability of affordable durable goods, e.g. energy-efficient cooking stoves, and W_2 is the availability of cheap perishable commodities. Let's suppose that W_1 affects the marginal effects of X on Y positively, and W_2 does so negatively. A switching variable S can be thought of as a variable that meddles with the relative salience or importance of W_1 and W_2. At constant values of these variables, S may meddle with the path-specific parameters that they are involved in. For instance, a co-intervention on S could be a public awareness campaign emphasizing the importance of investment in durables and how returns on such investments may help increase household wealth in the long-term. S can hence be an important interactive support factor for the effectiveness of an intervention as even at one and the same distribution of W_1 and W_2, it is important that these moderators are involved in the right way in producing the effects of interest, and the right values of γ_1 and γ_2, i.e. those needed for X to make suitable contributions, may need to be achieved by co-interventions on S.

Second, in contrast to the static conceptions of support factors outlined above we can also think of support factors as *dynamic support factors*. Dynamic support factors are support factors that, over time, have the capacity to sustain an effect of X on Y, or, alternatively, to erode such an effect. *Figure 2* offers an abstract example:

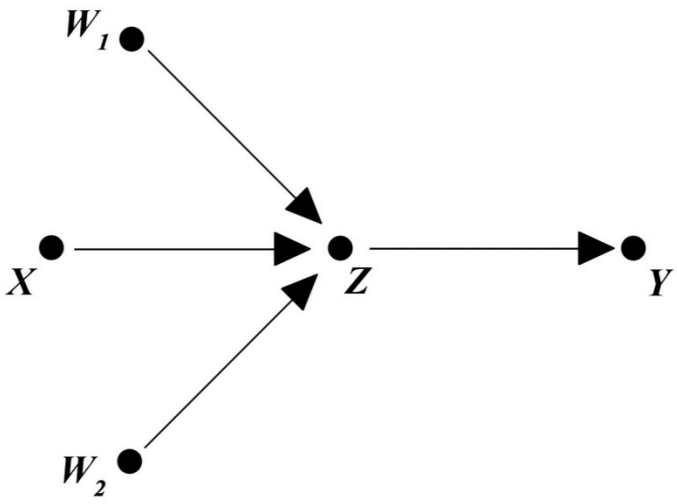

Figure 2: Dynamic support factors

Here, two variables, W_1 and W_2 co-determine a mediator Z that mediates the path between X and Y. W_1 can gradually, and over time re-set the value of Z to some level Z_0 (e.g. its pre-intervention level), thus undermining the persistence of the changes in Y induced by X through Z.

To give a concrete example, suppose that Y is smoking behaviour; X is smokers' awareness of the deleterious health consequences of smoking; an intervention on X is a public awareness campaign to attract smokers' explicit attention to these consequences; Z is a black box mediator that encodes agents' reasoning processes about the relative utilities of short-term gratification and long-term health consequences (which we assume is negatively relevant for smoking); and W_1 are the acute short-term cravings to smoke.

A public awareness campaign may be able to increase X, which induces changes in Z, i.e. how relative utilities are perceived, computed, and used as a basis for making smoking-related decisions. Z, in turn, induces changes in smoking behaviour Y. Yet, W_1 may be a constant variable, that, over time, will re-set Z to its pre-intervention level Z_0 in accordance with some function of W_1 and time. Thus, over time W_1 will re-set Y to its pre-intervention level and hence erode the persistency of the changes in smoking behaviour induced by the public awareness campaign. W_1 is hence negatively relevant for the effect of interest, as it has the capacity to undermine these effects gradually over time. We can now think of another variable, W_2, say support by friends and family or an incentive scheme that rewards non-smoking, which plays the opposite role of W_1. At the right levels it can cancel out the negative erosive effects of W_1, and hence help sustain the effects of X on Y induced through Z by keeping Z stable and resistant to the erosive changes induced by W_1. W_1 can hence be understood as a negative dynamic support factor, and W_2 as a positive dynamic support factor.

There might be yet other interesting senses in which we can think about support factors. For answering the questions posed at the beginning of this section, the conceptions outlined above suffice. First, in answering what different kinds of support factors there are, they help clarify that support factors can come in many different forms: interactive and non-interactive, static and dynamic, gradual and dichotomous. They also highlight that support factors can be support factors *for* different things: for specific levels of an outcome, for specific contributions to some outcome, or for maintaining an existing contribution to an outcome.

Second, clarifying the relation between support factors and moderating and mediating variables, it seems that moderating and mediating variables can be understood as one variant of support factors. So all moderating and mediating variables can be understood as support factors, but not all support factors are moderating or mediating variables. As suggested above, the comprehensive analysis of support factors developed here can both accommodate moderators and mediators, as well as extend beyond them significantly.

Third, not all support factors are important for (all kinds of) extrapolation. Non-interactive support factors only matter for predicting outcome distributions, but not for predicting effect sizes since they do not interact with interventions. Moreover, support factors of any flavour that do not regularly exercise their ability to modify causal effects do not pose relevant obstacles to extrapolation and may be disregarded. On the other hand, it is important to note that particularly in fully interactive cases, support factors are important targets for co-interventions. Depending on the relative costs and benefits of intervening on certain variables (which may differ between populations if these variables are differently related to the outcome), thinking more broadly about support factors can be helpful in identifying important co-interventions that may be necessary for achieving certain outcome distributions or effect sizes, and may be useful for realizing these distributions of effects by means of entirely different interventions than those that were the initial object of our extrapolation efforts.

Building on Cartwright's work, the more comprehensive analysis of support factors provided above makes important conceptual contributions to the literature on extrapolation by offering resources that remain unavailable on other accounts of extrapolation, specifically those to be discussed in the chapters to follow. These resources are not only useful at the level of abstract and theoretical results. One of the great advantages of thinking about support factors, at the most general level, and barring the details of more specific conceptions, is that it provides a graspable and intuitive way for policy analysts and policy makers to consider a diverse range of variables that may induce important obstacles to successful extrapolation. This is in contrast to the concepts of moderators and mediators, which are more technically involved and conceptually limited in the kinds of causal arrangements they can capture.

This more fully cultivated conceptual garden is not essential to the arguments developed in subsequent chapters, but nevertheless makes interesting contributions of

its own by facilitating the integration of central concepts figuring in characterizations of problems of extrapolation and in discussing strategies for extrapolation.

References

Bohrnstedt, G. W., and B. M. Stecher (eds.) (2002). "What we have learned about class size reduction in California", Sacramento, CA: California Department of Education.

Cartwright, N. D., and J. Stegenga. (2011). "A Theory of Evidence for Evidence-Based Policy" *Proceedings of the British Academy*, 171: 289–319

Mackie, J. L. (1965). "Causes and conditions", *American Philosophical Quarterly*, 2: 245–64.

CHAPTER 5

Mechanism-Based Extrapolation

5.1 Introduction

In this chapter, I consider Daniel Steel's mechanism-based strategy for extrapolation (Steel 2008). Steel's aim is to offer a strategy that can help us decide whether claims of qualitative causal relevance can be justifiably extrapolated from an experimental population to a target by using knowledge of similarities and differences in the mechanisms that govern the production of the outcomes of interest in both populations.

Steel begins by arguing that there are two crucial challenges that any persuasive strategy for extrapolation must evade. The first is that since causally relevant differences between experimental and target populations will almost invariably obtain, a persuasive strategy for extrapolation should tell us how extrapolation can proceed successfully despite such differences. Steel calls this the *problem of difference* (2008, 85).

The second challenge, discussed at length in *Chapter 3*, is the *extrapolator's circle.* The basic idea of mechanism-based approaches to extrapolation is to compare the mechanisms in the experimental and target populations with respect to whether they are sufficiently similar. However, while this idea is intuitively plausible, a persuasive strategy for extrapolation must at the very least avoid requiring full knowledge of the mechanisms that operate in the target, since obtaining such knowledge would threaten to render learning about the effect of interest in the target *from* the experiment redundant. So what Steel envisions for a useful strategy for extrapolation is that it can help decide whether an experimental result can be extrapolated given only *partial* information about the mechanisms in the target (2008, 87). I will continue to refer to this challenge as the *extrapolator's bind*, as it is more general and accommodates Steel's concerns about the circle.

Steel offers his *mechanism-based* strategy as a way to extrapolate claims of causal relevance that can evade both of the above challenges. More specifically, Steel argues for the following general procedure, called *comparative process tracing* (CPT): first, learn the mechanism in the experimental population by means of *process tracing*, a

method for causal model construction and evaluation that proceeds by comparing empirical consequences of a causal model against qualitative or quantitative data (see e.g. Schmitt and Beach 2015, Beach and Pedersen 2016). Second, *compare* the mechanisms in the experimental and target populations at stages where they are most likely to differ significantly (2008, 89). By "most likely to differ significantly", Steel means those stages at which differences in mechanisms are most likely to present obstacles to extrapolation, such as in a causal chain $X \rightarrow A \rightarrow B \rightarrow Y$ where the relationship between A and B might be disrupted in the target, or B may be absent, although, if realized, would permit transmission of causal effects from X to Y.

This strategy, by itself, is of course not sufficient to evade the extrapolator's bind, as learning the mechanisms in the experimental and target populations in order to compare them with respect to causally relevant similarities and differences might still require one to learn about the full mechanisms in both populations.

In order to avoid this problem, Steel argues that comparisons of similarity or difference of mechanisms are not necessary for *every* stage of the mechanisms at which differences could curtail the effectiveness of X on Y in the target, but rather only at so-called *downstream bottleneck* stages.

For instance, consider the following sketch of a mechanism borrowed from Steel's running example, which focuses on extrapolating the carcinogenicity of a substance called Aflatoxin B1 (AFB1) from animals to humans (to be explained in more detail shortly). Here, the mechanism in animals is as follows, with C being exposure to AFB1, and Y being the cancer outcome:

$$C \rightarrow X \rightarrow A \rightarrow Z \rightarrow B \rightarrow E \rightarrow Y$$

Suppose that the stages at which the mechanisms governing how animals and humans metabolize AFB1 are most likely to differ are Z and E. Then, if one is interested in extrapolating the effect of C on Y, it may be sufficient to compare the mechanisms at E, a descendant of C and the nearest causal ancestor to Y. If the mechanisms are similar or identical at this stage, in the sense that the effect of changes in C is transmitted down to E, then extrapolation of claims of causal relevance can proceed successfully even in the absence of full knowledge about the mechanism in the target. In a nutshell, the reasoning is that, if the experimental and target populations are relevantly different at any stage upstream of E and downstream of C, then variation

induced in C will not transmit down to E. Conversely, if variation induced in C does transmit down to E, either the mechanisms are similar at the intermediate stages between C and E, or they are dissimilar but the dissimilarities are not relevant since they do not block the path from C to E. According to Steel, this reasoning allows one to avoid learning about all intermediate stages of the C-Y-mechanism, and hence evades the extrapolator's bind.

In addition to the problem of difference and the extrapolator's bind, there are two further challenges that Steel considers.

The first is *causal dissonance*. Consider the case where the effect of X on Y is transmitted through two variables, P and Q, that mediate the effect on Y on two parallel causal pathways, such that X is positively causally relevant for P and Q, P is positively causally relevant for Y, and Q is negatively causally relevant for Y. To illustrate, consider the following diagram encoding the causal relationships between supplementary teaching X and students' performance on tests Y:

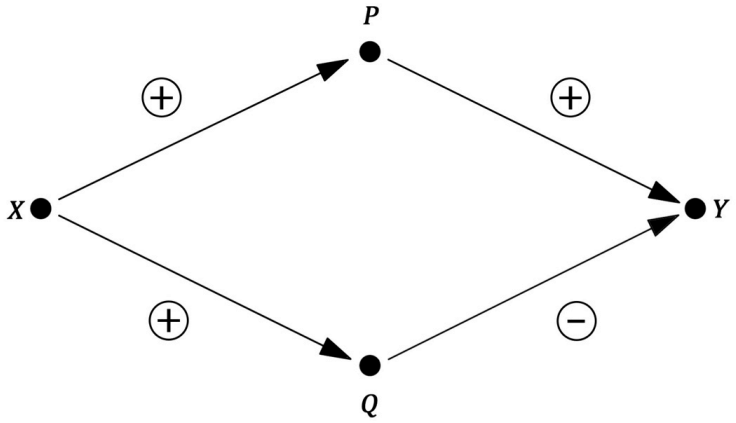

Figure 1: Two mediated paths from X to Y, one positive, one negative

Here, supplementary teaching X increases students' performance on tests Y by increasing unobserved ability P, which in turn is positively relevant for performance Y. In addition, on a parallel mediated pathway, receiving supplementary teaching also leads to students being stigmatized by peers for being in need of such teaching, represented by Q, which is negatively relevant for performance because stigmatization renders students less confident in their abilities.

Assume now that we wish to extrapolate the positive causal relevance of supplementary teaching from an experimental population in which the pathway involving stigmatization is absent, i.e. the arrow between X and Q is severed, to a target

where this connection is present. Then, the negative causal relevance of stigmatization in the target is an obstacle to extrapolating the effect of interest.

This is important because remaining ignorant of the mediated effect through Q might lead to mistaken predictions. Moreover, absent knowledge about functional form relationships as well as the signs (and magnitudes) of the parameters associated with the arrows that connect P and Q with Y, and how these parameters are distributed among individuals in the target, we will be unable to decide whether X is positively causally relevant for Y in the target.

Acknowledging this problem, Steel argues that the positive causal relevance of X for Y in the target can only be decided if the causal relationships $X \rightarrow P$, $X \rightarrow Q$, $P \rightarrow Y$ and $Q \rightarrow Y$ in the target are *positively consonant*, that is, if the signs of the parameters associated with all edges on the pathways through P and Q are positive. In the example above, this is not the case since $Q \rightarrow Y$ has a negative sign. So in this case, qualitative causal relevance cannot be decided unless we know the signs of the parameters for all causal relationships. More formally, according to Steel, successful extrapolation of positive causal relevance claims requires that there is no subpopulation Φ_i of the target population B such that the relative frequency φ_i of Φ_i in B is positive and X is a negative causal factor for Y in Φ_i (2008, 112). In other words, positive consonance requires that "[…] X is not negatively relevant to Y for *any* combination of mechanisms found in the population" (2008, 112; emphasis added).

The second important obstacle for Steel's CPT is that learning about similarities at downstream stages of mechanisms can only be useful if one is reasonably confident that the mechanism in the target is not *bypassed* around these stages. Recall the supplementary teaching example again and suppose that the mechanism learnt in the experimental population is $X \rightarrow P \rightarrow Y$. In this case, even though we might be able to learn that the relationships $X \rightarrow P$ and $P \rightarrow Y$ are instantiated in the target, it may still be possible that there is a parallel and unknown causal pathway from X to Y through Q, which has negative causal relevance for Y. In such a case, depending on the signs of the parameters associated with the relationships $X \rightarrow P$, $X \rightarrow Q$, $P \rightarrow Y$, and $Q \rightarrow Y$, and how they are distributed in the target, it may be possible that the joint effect of X on Y is positive in the experimental population but negative in the target.

Steel argues that the success of his proposed strategy hinges on knowledge of whether the downstream stages at which mechanisms are compared are so-called *bottlenecks*, i.e. stages "[…] through which *any* influence upon the outcome must be transmitted" (2008, 90; emphasis added). The supplementary teaching example above is a case in which no such bottleneck exists, since there is no immediate causal parent of Y that is a descendant of all pathways connecting X and Y such that comparisons of similarity at this stage could, under the favourable conditions of positive causal consonance, obviate the need to learn about the parallel pathways through P and Q respectively.

Against the background of these two important caveats, Steel offers his extrapolation theorem, specifying when claims of positive causal relevance can be extrapolated. Steel's theorem says that this is possible when 1) there is a nonempty subset of individuals from the target population B of interest for which the mechanisms connecting X and Y are not disrupted, 2) all mechanisms connecting X and Y are positively consonant for all individuals in B, and 3) the mechanisms exhibit a downstream bottleneck stage (2008, 113). Under these conditions, learning that the mechanisms in both populations are similar at a downstream bottleneck stage allows one to infer that X will be positively causally relevant for Y in the target.

With this result in place, let me proceed to a critical discussion of Steel's strategy. In *Section 2*, I elaborate on several general concerns about the epistemic demands imposed by Steel's strategy as well as its scope. In *Section 3* I turn to a more important concern, which is that Steel's strategy experiences difficulties in evading the extrapolator's bind in cases routinely encountered in EBP. To this end, I develop a distinction between two kinds of extrapolation, *attributive* and *predictive*. The former is the kind that Steel's strategy specifically targets and can help us to successfully overcome; the latter is common in EBP, and substantially more difficult to overcome using Steel's strategy. *Section 4* offers a detailed example to illustrate the distinctive difficulties involved in predictive extrapolation. *Section 5* concludes.

5.2 General Concerns

5.2.1 Epistemic Demands and Scope

The first concern I want to discuss is that Steel's strategy imposes highly demanding lower bounds on the knowledge required to decide whether experimental results can be extrapolated. Specifically, these demands concern 1) identifying the stages at which differences in mechanisms are most likely to pose obstacles to extrapolation and which of these stages are *downstream stages* of the mechanism, 2) determining whether mechanisms are *consonant*, and 3) determining whether downstream stages are not bypassed, i.e. whether they are *bottlenecks*.

First, in order to identify stages at which differences in mechanisms are most likely to present obstacles to extrapolation and which of these stages are *downstream* stages, one will typically need to have extensive understanding of the mechanisms in both populations. For instance, in Steel's running example concerning the carcinogenicity of AFB1 (to be discussed more fully shortly), researchers had extensive experimental evidence concerning between-species differences in the carcinogenic effects induced by AFB1 exposure, as well as an understanding of how differences in the metabolic mechanisms present in different species accounted for these differences in effects. This rich background of experimental evidence and consequent understanding of mechanistic differences between species allowed researchers to determine the stages at which mechanisms are most likely to differ with respect to how AFB1 is metabolized. This provided a background against which Steel's CPT could proceed effectively, as the available understanding of the mechanisms significantly constrained the range of stages at which mechanisms would need to be compared in order to support between-species extrapolation of claims of causal relevance.

While experimental evidence and subsequent understanding of mechanisms of this kind may be more frequently available in epidemiology, biochemistry, molecular biology, and life sciences more generally, it seems that this is less frequently the case in social science contexts where researchers often have little grasp of the details of the mechanisms that connect an intervention variable, such as household endowment, to outcomes, such as private investment, welfare levels, or health indicators for children in households of the rural poor. This is especially the case if important parts of these mechanisms involve agents' decision-making and are hence governed by psychological processes. To be sure, it is not uncommon for social science researchers to have so-

called *programme theories*, *theories of change*, or *logic frames* which describe details of how an envisioned intervention is *supposed* to be effective, including variables believed to be important for mediating their effects (see e.g. White 2009; Pawson and Tilley 1997; 2001; Pawson 2013; Davey et al. 2017). Yet, while there are arguably cases where such theories are well-supported and provide adequate representations of important mechanistic features, in many other cases such theories only characterize how envisioned interventions are *hoped* to be effective, but where these theories are based on potentially incomplete knowledge or indeed mistaken beliefs about the true underlying causal mechanisms that will ultimately govern the effectiveness of these interventions.

The second way in which Steel's strategy is epistemically demanding concerns the requirement that mechanisms must be *positively consonant* for all individuals in both populations. This is both difficult to learn as well as unlikely to be the case in many social science contexts (Vivalt 2019). For instance, the effects of microfinance programmes are known to vary considerably between households, both quantitatively as well as qualitatively (Banerjee et al. 2017). In virtually all contexts where the welfare effects of policies are a concern, it is plausible to assume that individuals' responses to interventions differ not only in magnitude but also in sign; in Steel's terminology, the mechanisms are likely to be causally *dissonant*. This not only undermines the applicability of Steel's extrapolation theorem, which hinges on the assumption of consonance, but also threatens the ability of Steel's strategy to evade the extrapolators' bind. More specifically, in order to determine whether mechanisms are consonant in the target, one may need to learn about *all* parts of the mechanisms that connect treatment and outcome. This makes it likely that the effects of interest may be identified on the basis of such knowledge alone, thus rendering evidence from the experimental population redundant to our conclusion.

The third concern about epistemic demands is that even if the condition of causal consonance were satisfied and could be supported without falling prey to the extrapolator's bind, this alone may often not be enough to decide whether an experimental result can be extrapolated. In addition to being reasonably confident that mechanisms are similar at relevant downstream stages, we need to be confident that the downstream stages are not bypassed either.

For instance, cash transfers X may increase household expenditures on goods Z that increase children's nutritional health Y in an experimental population. Yet, while these

aspects of the mechanism may also be instantiated in a target population, obtaining evidence of such mechanistic similarity may not be enough for successfully extrapolating claims of causal relevance. For instance, in the target population, cash transfers might be causally relevant for other activities that adversely affect children's health outcomes such as cigarette consumption, where these adverse effects may importantly depend on background conditions such as cigarette prices or social norms concerning smoking in the proximity of children. As these background conditions may conceivably vary between different settings, this means that the mediated and moderated effects may significantly differ between populations. This makes it possible that even though the target may share *all* the features of the mechanism in the experimental population, it may nevertheless exhibit other, *extraneous* causal relationships that bypass the downstream stages of interest and thereby, if unaccounted for, create obstacles to successful extrapolation.

In light of this, it is important to note that Steel's strategy does not only *put emphasis* on exploiting the fact that upstream differences do not matter for extrapolation so long as mechanisms are similar at downstream bottlenecks, but crucially hinges on whether such bottlenecks exist at all and fails to offer the desired shortcut for extrapolating causal claims whenever this is not the case.

While Steel gives explicit consideration to these worries, by arguing that background theory and intuitive considerations of plausibility can help rule out alternative specifications of mechanisms in the target that could obstruct extrapolation, he does not discuss how extrapolation should proceed in cases where such information is unavailable, other than suggesting that naïve extrapolation may be used as a fall-back strategy in such cases (called *simple induction* by Steel; 2008, 96). Steel explicitly objects to this strategy in several other places, however, on the grounds that it is unreliable whenever there are reasons to suspect that mechanisms are likely to differ between populations (e.g. 2008, pp.80).

So, particularly in social science scenarios where sophisticated mechanistic background theory is often unavailable (such as in economics), the preconditions required for applying Steel's strategy may simply not be satisfied.

This is in line with Steel's acknowledgement that the difficulty in addressing problems of extrapolation in social science contexts often starts with lacking any knowledge of the mechanism in the experimental population apart from knowing that

there is *some* mechanism connecting intervention and outcome variables. This concern is particularly pressing in EBP where the predominant aim has been to focus on the effects of causes, rather than the causes of effects (see e.g. Heckman 1992), which means that there is traditionally little concern for learning about mechanisms. Yet, at the same time, and perhaps ironically, it seems that mechanisms are often of little concern because it is neither particularly controversial that variables such as household endowment are causally efficacious for outcomes such as children's nutritional health in *some* way, nor are claims of causal relevance the primary epistemic target in EBP. Instead, social scientists, policy analysts, and evaluators working on problems of extrapolation are often interested in extrapolating *quantitative* causal effects.

This brings me to the second general concern with Steel's strategy, which is that it focuses on a specific class of extrapolative queries, i.e. qualitative rather than quantitative ones.

5.2.2 No Quantitative Extrapolation

Steel's strategy is intended to elucidate how learning about similarities and differences between populations can help decide whether an intervention is causally relevant for some outcome in a target population. This moves significantly beyond Cartwright's effectiveness argument in that some causally relevant differences can be overcome by Steel's strategy, whereas the effectiveness argument, at least on some interpretations, requires full-fledged identities in causal mechanisms. Yet, despite these advantages, Steel's strategy still falls short of elucidating how differences and similarities between populations bear on differences in the *magnitude* of the causal effects of an intervention. It only tells us that similarities and identities are not required at all stages of causal mechanisms, and permits qualitative extrapolative conclusions in the presence of some causally relevant differences. To be sure, Steel's strategy was never intended to facilitate extrapolation of quantitative causal claims, so it might seem odd to highlight that it cannot address such issues. Yet, since one of the main aims of the present project is to investigate the usefulness of different extrapolation strategies for extrapolations encountered in EBP specifically, this feature of Steel's strategy is still important to mention, rather than criticize.

With this in mind, the concern here is similar to that offered concerning the effectiveness argument: the applicability of Steel's strategy is restricted to a narrow

class of extrapolation problems. Even under favourable epistemic conditions, it is only sufficient to decide whether claims of qualitative causal relevance can be extrapolated. Yet, for many social science purposes, specifically those often pursued in EBP, this is insufficiently informative, since claims of positive causal relevance are not enough to decide whether an intervention in the target will be sufficiently effective; sufficiently cost-effective; more (cost-) effective than alternative interventions. Neither can claims of causal relevance help us decide, moreover, whether an intervention has desirable welfare consequences with regard to the distribution of treatment effects or how an intervention might need to be modified and tailored to the specific circumstances of the target for it to achieve some desirable effect. Hence, a large and important class of queries relevant in EBP remains unaddressed by Steel's proposal.

This relates to a third general concern, which is that Steel's strategy focuses on a specific class of mechanisms that does not seem to be the predominant kind of mechanism encountered in many social science contexts.

5.2.3 Social Mechanisms

Steel's running example concerning the carcinogenicity of AFB1 in animals and humans involves a mechanism that proceeds from start to finish through a single path with multiple mediators. In social science settings, however, it is sometimes, and perhaps often, plausible to assume that outcomes of interest are co-determined by an intervention variable through multiple pathways that involve various mediating and moderating variables. This renders the application of Steel's strategy more difficult. For instance, even the simplistic case of supplementary teaching outlined above does not seem amenable to be addressed by Steel's strategy because there are no downstream bottleneck stages at which comparison of mechanisms could proceed. This is the case whenever there are multiple causal pathways connecting treatment and outcome and there is no immediate causal parent of the outcome that is a proper descendant of *all* pathways that connect treatment and outcome. In such cases, using CPT to decide whether causal claims can be extrapolated requires comparison of mechanisms at multiple stages, and at least one for each pathway connecting the intervention and outcome variables. This may be significantly more burdensome than meeting the relatively mild epistemic demands in Steel's AFB1 example where there is only one downstream stage.

A second general concern about mechanisms encountered in social science contexts such as development economics is that it is often unclear whether the mechanisms between an intervention variable, such as household endowment, and an outcome, such as a household welfare measure, are *homogeneous* between units in a population at any level of description more detailed than that X is causally relevant for Y. For instance, the economic decision problems that households in development contexts face with respect to consumption, investment, and savings behaviours, and consequently the mechanisms that connect variables such as household endowment and welfare indicators, are likely to involve highly heterogeneous disjuncts of variables that can vary between households in at least three ways: whether they are involved at all in the mechanisms; the ways in which they are involved; and in the levels at which these variables are realized before an intervention (see e.g. Garcia and Wantchekon 2010; Vivalt 2019).

For instance, two individuals *i* and *j* that face identical economic constraints and have identical information and preferences over outcomes may still engage in different economic behaviours depending on how economically sophisticated they are. While *j*, in virtue of prior education, may be able to calculate the expected utilities of different courses of action with reasonable accuracy and subsequently choose the action that is expected to maximize her subjective utility, *i* might be unable to do so with similar sophistication, and hence engage in different choice behaviours. This toy example suggests that specifically those parts of mechanisms that operate at the psychological level of choice behaviour may often be highly heterogeneous both within and between populations with no straightforward way to distinguish individuals with respect to their individual-specific mechanisms.

This poses a distinctive challenge for the applicability of Steel's strategy. Steel's strategy is premised on the assumption that mechanisms are *consonant* within both study and target populations. Yet, while this seems plausible in the epidemiological case that Steel discusses, and more broadly in biomedical and life sciences, it seems less so in social science contexts. Even if X is causally relevant for Y in *some* way in every individual in a population, information on mechanisms for almost all individuals might be necessary in order to decide the sign of the aggregate effect. This may not only be difficult to obtain but may already be sufficient to learn an effect of interest. In short, the worry is that heterogeneity in mechanisms within and between populations may

raise the lower bound for mechanistic knowledge required for using CPT even higher than it already is, to the point where such knowledge might need to be obtained for a large number of individual-types that are distinguished by their type-specific mechanisms.

This relates to a third concern, which is that *homogeneity* in mechanisms is an important feature that helps CPT yield extrapolations that generalize over broad domains.

In Steel's AFB1 case, it is plausible to assume that while there are often important *between*-species differences in mechanisms that matter for the carcinogenicity of AFB1, such differences rarely obtain *within* species. This allows researchers to extrapolate from relatively few animal experiments to a very broad target, i.e. humans in general. Yet, as suggested above, in social science contexts it is often not plausible to assume that experimental and target populations are internally homogeneous with respect to some causal query. This presents a problem for the generality of the conclusions that can be obtained from Steel's strategy. Specifically, extrapolating successfully from one human population to another does not necessarily support extrapolation to any other human population beyond the target. Unlike in biomedical sciences, where such generalizations can often be supported by identifying further targets as members of the same biological type, e.g. rats, humans, etc., which can be supported by appeals to mechanistic homogeneity at the type-level, social science contexts often do not allow such additional inference in a straightforward way. This is because human populations frequently differ in their structural, institutional, social and psychological makeup, so type-level generalizations are often difficult. For additional inferences to yet other targets, a separate extrapolative inference might hence need to be entertained, which will require a separate set of comparisons of mechanisms with respect to similarities and differences.

So while the persuasiveness of Steel's strategy as a method for extrapolating claims of causal relevance across the boundaries of species is supported by the fact that, for instance, metabolic mechanisms are often relevantly homogeneous within biological species, such support may not be present in other domains, where extrapolation hence not only becomes more tedious, but also becomes substantially limited in its reach.

Finally, the fourth concern about mechanisms in social science is that they are recognizably more difficult to observe than those encountered, for instance, in biology,

biochemistry, and epidemiology. In the spirit of the process tracing literature, by *observed* I mean, in the first instance, that one is able to observe distinctive or characteristic marks or symptoms of these mechanisms (see e.g. Salmon 1984 ch. 4; Beach 2017). For example, the presence of certain metabolites in the excretions of rats may strongly indicate that the metabolism of a certain compound proceeded in a specific, rather than some other way. To the extent that background theory is sufficiently strong to indicate that a suspected mechanism, and only that mechanism, tends to bring about certain kinds of observable consequences, then observing such consequences may provide a strong basis for abductive inference to the presence of the mechanism in question.

In addition, it may sometimes be possible to observe mechanisms more directly, such as in biochemistry where researchers can often study metabolic mechanisms *in vitro*. In Steel's running example of AFB1 carcinogenicity, such relatively direct observations of mechanisms can be obtained from experiments at different levels, e.g. on whole-cell systems such as precision-cut liver slices; at the cellular level on hepatocyte cultures; and at the sub-cellular level on hepatic microsomes (see IARC 1993). Importantly, such studies can be performed both on tissue and cell samples from animal models as well as on human cell systems and cultures, thus allowing not only relatively immediate observation of mechanisms in the experimental population but also observations of aspects of relevant mechanisms obtained directly from in vitro samples of the target population.

It is clear that such observations make it significantly easier to compare mechanisms with respect to similarities at important stages compared to social science settings, where it is often unclear whether mechanisms can be observed in such straightforward ways even in the experimental population. While there are instances where this seems possible, e.g. when agents' self-reports of how and why they (expect to) behave in response to a certain hypothetical intervention can figure a reliable guide to predicting their behaviours (I will say more on this later), there are of course many other cases where acquiring evidence with bearing on mechanistic features is not possible, or at least likely to be unreliable.

For instance, examining the mechanism by which students convert various kinds of teaching inputs into performance on standardized tests is a case where students' self-reports are unlikely to elucidate important features of the mechanism by which they

come to acquire and use knowledge for passing tests. In such cases, using CPT may not be applicable as obtaining detailed understanding of the mechanisms of interest may not be feasible even in an experimental population, let alone the target.[1]

The concerns outlined above do not, of course, provide a general reason to reject Steel's strategy for extrapolation in social science, as the question of whether any given setting of inquiry will exhibit the features that give rise to them will depend much on the specific context, and many real-world cases may be substantially less troublesome than the worst-case scenarios outlined above. Yet, there is a second, and I believe more general, concern about Steel's strategy, which is that it applies only to specific kinds of extrapolation, and experiences substantial difficulties in evading the extrapolator's bind in other cases, which I consider to be prevalent in EBP. The next section will offer an overview of this concern, with a concrete and detailed example following in *Section 4*.

5.3 Two Kinds of Extrapolation

The problem of extrapolation that Steel considers in developing his strategy focuses on a specific kind of extrapolation, which I call *attributive extrapolation*. I argue that this is not the kind of extrapolation that is typically of interest in social science contexts, including in EBP. Specifically, I argue that the ability of Steel's strategy to persuasively evade the extrapolator's bind does not extend to this latter kind of *predictive extrapolation* and that the extrapolator's bind hence remains a serious obstacle for it.

Before I proceed to my argument, let me expand in more detail on Steel's running example concerning the carcinogenicity of AFB1 in animals and humans. I begin with a brief historical exposition that highlights important and distinctive characteristics of the conditions under which the extrapolation from animals to humans proceeded. Against this background, I then develop my distinction between attributive and predictive extrapolation, and flesh out my argument for why Steel's strategy is not successful in evading the extrapolator's bind in predictive extrapolation cases.

---

[1] To appreciate these challenges, consider the extensive literature on so-called *educational production functions,* where economists and econometricians have long attempted to estimate reliable models of student learning for evaluative and predictive purposes (see e.g. Hanushek 1979 and Todd and Wolpin 2003, cited in Muller 2013 ms.)

5.3.1 Aflatoxin B1 Revisited

Aflatoxins, a class of spoilage mould metabolite, are now known as a contaminant of a variety of foods such as peanuts, grains, corn, as well as animal feeds (Wogan 1966). Over the course of roughly two decades of animal experiments, the carcinogenicity of Aflatoxins, including a particularly potent type called Aflatoxin B1 (AFB1), were established in a variety of animal species including rats, mice, hamsters, and others. These experiments also clarified that the carcinogenicity of AFB1 varied significantly between species (Wogan 1992; Gold et al. 1992), and supplementary in vivo and in vitro experiments helped attribute these differences to differences in metabolic mechanisms between species. The remarkable potency of Aflatoxins as carcinogenic and toxic agents spurred interest in the question of whether humans are similarly susceptible to these effects.

By the mid-1980s this question had enjoyed significant attention from epidemiologists who were successful in producing extensive observational evidence from case-control studies in human populations with high prevalence of liver cancer (hepatocellular carcinoma; henceforth HCC) that were suspected to have been exposed to AFB1, as well as evidence from prospective cohort studies in populations that were known to be exposed to AFB1 through diet (Wogan 1999). In these populations, high odds ratios for HCC and significant increases in relative risk of HCC were estimated for individuals that had been exposed to AFB1 (IARC 1993). However, causal attribution of HCC incidence to AFB1 exposure was complicated by the fact that the populations of interest also exhibited high background rates of Hepatitis B virus (HBV) infection, which was already known to be a potent cause of HCC in humans. This made it more difficult to unambiguously, and *causally* attribute HCC to AFB1 exposure. Hence, aside from more carefully designed observational studies that made attempts to disambiguate the covariance structure between HCC, HBV, and AFB1 exposure, one of the crucial steps in successfully causally attributing HCC to AFB1 exposure was to establish the existence of a causal mechanism that could underpin the observed associations. As the mechanism governing carcinogenicity of AFB1 in animals became gradually better understood through in vitro and in vivo experiments on parts of the hypothesized mechanism, this prompted researchers to investigate whether the same or similar mechanisms that governed the production of carcinogenic metabolites of AFB1 in animals were also present in humans.

This proceeded by comparing the mechanistic evidence obtained from in vitro and in vivo studies on animals with evidence from in vitro studies on human liver slices, liver cell cultures, and individual liver cells. These studies established that AFB1 metabolism proceeded similarly in humans as in rats, as specific metabolites that were known to be the proximate cause of HCC in rats were also obtained from human liver samples exposed to AFB1. As in vivo studies on rats had determined that rats are highly susceptible to carcinogenic effects of AFB1, these similarities in metabolic mechanisms provided a strong basis for the conclusion that AFB1 exposure is also a potent cause of HCC in humans (Hengstler et al. 1999; but see Reiss 2010 who questions whether the extrapolation indeed evaded the extrapolator's circle).

With this sketch in place, let me proceed to highlight what I consider to be important and distinctive features of Steel's example. These features underlie my distinction between *attributive* and *predictive* extrapolation.

5.3.2 Attributive and Predictive Extrapolation

Extrapolating the causal relevance of AFB1 for HCC from animals to humans proceeded against a background where observational evidence of an association between AFB1 exposure and HCC prevalence spanning various human populations was already available. Here, both the suspected cause of interest, AFB1 exposure, as well as its suspected effect, HCC, were jointly realized and observed in at least some humans. Against the background of these observations, the inferential target of the animal-human extrapolation was hence not to answer the question of whether AFB1 was a cause of HCC in humans *simpliciter*, but rather, in the first instance, whether it was a cause of *observed* HCC in humans that were known to have been exposed to AFB1. So, given that one had observed individuals who exhibit a particular outcome, HCC, the question was whether AFB1, *rather than,* e.g., HBV infection, or *in addition to* HBV infection, was a cause of this outcome. In more general terms, the extrapolative query of interest to researchers was whether X is a cause of Y, *rather than* or *in addition to* Z, in population B, if it is so in A. Given that one observes the outcome of interest, and that one believes that the outcome is caused by *something*, the aim is hence to successfully *attribute* causal relevance for the outcome to the suspected cause of interest rather than others. I call this type of extrapolation *attributive extrapolation*.

This activity is importantly different from the activity that researchers in social science contexts typically engage in, and specifically in EBP. Here, the question of interest is often whether X will bring about changes in Y in a target population where neither the intervention on X nor the requisite changes in Y, qua potential causal effects of X, have yet been observed. In these circumstances, the question of interest is *predictive* rather than attributive: will intervening on X cause future, yet unobserved, changes or realizations of Y in B, if it does so in A?

This distinction is related to a distinction made by Marcellesi (2015) between predictive and explanatory external validity inferences. Marcellesi's distinction focuses primarily on differences in epistemic aims: predictive external validity inferences are concerned with predicting the effects of an intervention in a novel context, whereas explanatory external validity inferences focus on explaining the occurrence of an observed effect by reference to a specific cause. In contrast to Marcellesi, my distinction is concerned, however, not (or not predominantly) with the epistemic aims of the extrapolative inference, but rather, and more importantly, with the kinds of causal information that are available in each scenario. In attributive extrapolation the effects of interest have already obtained in the target and were (or could have been) observed there, whereas in predictive extrapolation the effects of interest have not yet been experienced in the target and hence have not been and could not have been observed there.

To illustrate this key difference in more detail, let me modify Steel's example so that it matches the features that I consider distinctive of *predictive* extrapolative queries. An extrapolation concerned with predicting the causal relevance of AFB1 in humans would be one where we have never observed any human populations in which individuals were both exposed to AFB1 as well as exhibited HCC. This situation makes successful extrapolation using Steel's strategy recognizably more difficult. Put simply, if one has never observed humans exposed to AFB1, then it is significantly more demanding to establish the causal relevance of AFB1 for HCC in humans, as it requires more extensive knowledge of the causal mechanism governing HCC production from start to finish.

In order to better understand why this is the case, let me offer a framework that helps illustrate the differences between these two kinds of extrapolation with respect to the eventual aims of the inference; the evidence from the target that is available; the extent

of mechanistic evidence required to support these extrapolations; and the difficulties involved in obtaining such evidence. Specifically, I argue that differences with regard to the latter can help elucidate why Steel's CPT may be successful in evading the extrapolator's bind in cases of attributive extrapolation, such as Steel's AFB1 case, but not in cases of predictive extrapolation that are predominant in EBP.

5.3.3 From Counterfactuals to Mechanisms and Back

The framework I offer to help concretise the differences between attributive and predictive extrapolation is loosely based on the Neyman-Rubin potential outcomes framework for causal inference (Neyman et al. 1935; Rubin 1974; Holland 1986). It starts from the idea that important epistemological features of causal inference can be framed in terms of the process of constructing, observing, or mimicking appropriate *counterfactuals* that help determine whether a putative cause makes a difference for the outcome of interest and what the magnitude of this difference is.

Drawing on this framework, my aim is to show that attributive and predictive extrapolation are significantly different activities. Attributive extrapolation aims at reaching conclusions about *which causes and mechanisms* are responsible for a difference between two actual, *observed* states in the target. Predictive extrapolation, on the other hand, aims at inferring a *future counterfactual state* in the target. In short, one could say that attributive extrapolation is concerned with inferring the causes of suspected effects, whereas predictive extrapolation is concerned with inferring the future effects of interventions on suspected causes.

Let P and Q be state descriptions of a population A, i.e. factual and counterfactual characterizations of the states of certain variables in A. P and Q each comprise of measurements or predictions of a triple $< X, Y, K >$, where X is the suspected cause of interest (e.g. exposure to AFB1), Y is the outcome of interest (e.g. HCC), and K is a vector of additional measured or unmeasured variables that include other causes of Y (e.g. HBV infection) as well as moderating and mediating variables that figure in the mechanism governing the production of Y by X. A and B are subscripts for state descriptions P and Q that indicate the population to which the description applies, where A indicates the population from which one wishes to extrapolate and B indicates the target.

P and Q are supposed to figure as descriptions of actual and counterfactual states of affairs, either for one and the same individual or for one and the same population. Yet, the *fundamental problem of causal inference* is that P and Q are never jointly observed for the same units (Holland, 1986). Hence, an important question for researchers is often whether a state description P that describes an *actual* state for a set of individuals i (say, individuals exposed to an intervention) is a *valid counterfactual* for another, disjoint set of individuals j which are observed at their respective actual state Q (e.g. not exposed to an intervention), and vice versa. In standard causal inference settings, such as RCTs, randomization and additional methodological precautions are supposed to warrant using P as a valid counterfactual for Q and vice versa. To the extent that such precautions are successful, differences between outcomes Y of P and Q may be interpreted as measures of the *average causal effects* of the exogenously induced differences in X between P and Q (Rubin 1974). *Figure 2* illustrates the standard RCT case:[2]

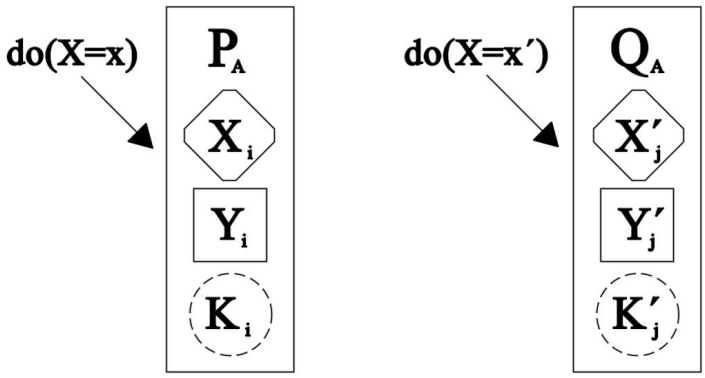

Figure 2: Constructed counterfactuals from an RCT on animals

Here, P_A and Q_A represent control units and treated units respectively. Again, Y is observed in both groups, which is indicated by squares, and X is exogenously controlled by the experimenter, which is indicated by $do(X = x)$ and $do(X = x')$, and the diamond-shaped symbol for X. The vector K of causes of Y other than X and moderators and mediators of the hypothesized $X \rightarrow Y$ effect are not observed, which is indicated by the circular dashed symbol. However, randomization of treatment ensures,

[2] The so-called *do-operator* $do(X = x)$ that figures in this diagram is borrowed from Pearl (1988). It is used as shorthand to indicate that an intervention is performed that sets X to a specific value $X = x$. Although Pearl's causal inference framework importantly differs from the potential outcomes approach, the do-operator provides a convenient (and hopefully not exceedingly offensive) way to represent interventions here.

in expectation, that the net effects of K on Y are the same for P_A and Q_A, so P_A and Q_A are still informative state descriptions for inferring causal effects over the full distribution of K.

The important thing to notice about this setting is that the aim of the inference is to answer the question: *what is the counterfactual for P_A with respect to a change in X*? The purpose of the experiment is to ensure that Q_A is a valid counterfactual for P_A by means of randomization and experimental control of X. The particular aspect of Q_A that is of interest is its value of Y'. If Y' is eventually observed, then this permits computation of the average causal effect of changes in X on changes in Y as the difference between the averages $\overline{Y_J'} - \overline{Y_\iota}$.

Let us suppose now that the above kind of experiment shows that AFB1 exposure (X) is positively causally relevant for HCC (Y) in a variety of animal species. How does the extrapolation proceed from these results to conclusions about the carcinogenicity of AFB1 in humans? For clarifying this, it is important to make the causal query explicit that is supposed to be answered by the extrapolation. At face value, this causal query seems to be *whether X is causally relevant for Y in humans.*

This would be the case, for instance, if we were to estimate that $\overline{Y_J'} - \overline{Y_\iota} > 0$ in an RCT on humans. Yet, since such experimental intervention would not only raise moral concerns but also fall prey to the extrapolator's bind, the aim of the extrapolation is to obtain a conclusion about the carcinogenic efficacy of AFB1 in humans based on evidence about similarities and differences in mechanisms between rats and humans, as well as other kinds of supplementary evidence.

Here is where the distinctive features of Steel's Aflatoxin case become most apparent: in Steel's example researchers had access to observations on humans that allowed constructing state descriptions P_B and Q_B that could figure as *candidate* counterfactuals for each other. Specifically, case-control and prospective cohort studies on human populations included observations of 1) individuals that were exposed and individuals that were not exposed to AFB1, 2) individuals that did and did not exhibit HCC, as well as 3) individuals that did and did not exhibit HBV. Thus, it was possible to construct two state descriptions P_B and Q_B, for which the relevant outcome (HCC), the suspected cause of interest (AFB1), and at least some other known causes of the outcome (HBV) were observed in their different realizations. *Figure 3* summarizes this

by representing that both P_B and Q_B can be constructed from observational data. This is indicated by square symbols for X and Y, and the solid square/dashed circle symbol for K, suggesting that at least some elements of K (HBV) are observed. Hence, P_B and Q_B can figure as *candidate counterfactuals* for each other.

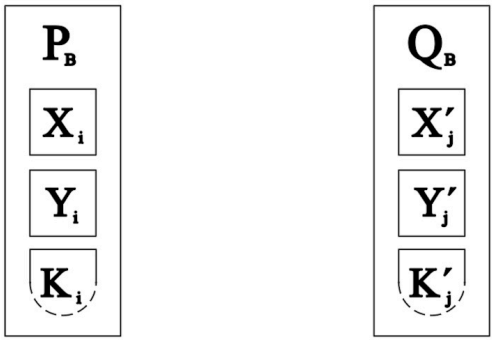

Figure 3: Candidate counterfactuals from observational data on humans

Against the background of these state descriptions, and the additional assumption that the observed difference between Y_j' and Y_i has *some* cause, it becomes clear that the aim of the extrapolation is not to answer the question of whether X is causally relevant for the production of Y in humans *simpliciter*. Instead, the aim is to answer the question whether *observed* differences in AFB1 exposure, i.e. differences between X_i and X_j', are in fact the causes of *observed* differences in HCC, i.e. Y_i and Y_j', or whether rather some other difference in K (e.g. HBV) is relevant for bringing about these observed differences. An attributive conclusion, then, would be to say that the observed differences in X caused the observed differences in Y.

On Steel's account, the case for this conclusion can be strengthened by providing evidence that the same or similar mechanism that is known to govern the production of Y by X in rats is also instantiated in humans. According to Steel, CPT helps reach this conclusion by comparing animal and human populations with respect to stages at which differences in mechanisms for the production of Y are known to be likely. Comparison of these stages proceeds by investigating whether known observable consequences of the presence of the hypothesized mechanism in rats are also present in humans. This is what Steel refers to as investigating *distinctive marks* or *symptoms* of the hypothesized mechanisms (cf. Salmon 1984 ch.4; Beach 2017), i.e. marks and symptoms that are expected to obtain, ideally, if and only if a suspected mechanism is operating. According to Steel, evidence of the presence of such marks and symptoms strengthens

the attributive conclusion that differences in X, rather than differences in some other variable in K (such as HBV), were indeed the cause of the observed differences in Y among the studied human populations.

5.3.4 Predictive Extrapolation in Terms of Counterfactuals

Let me expand on the distinctive features of *predictive* extrapolation. Suppose that we have RCT evidence from population A indicating that supplying microcredit to agents in A is efficacious in increasing household welfare indicators. Suppose that the hypothesized mechanism to explain this causal effect is that microcredit allows agents with inadequate access to capital markets to pursue entrepreneurial efforts, which in turn generate household income and wealth that are subsequently used for consumption (e.g. of more durable goods, healthier foods etc.), which increases several variables that figure in the welfare indicator of interest.

Against the background of this, our extrapolative query is whether the microcredit programme is also positively causally relevant in a target B. Again, at first glance, it might seem that the target of the extrapolation is the same general claim as above, i.e. *whether X is positively causally relevant for Y*. Yet, the circumstances under which such queries are answered in social science settings are often significantly different from those encountered in Steel's example. Specifically, the most important difference will often be that the putative cause and effect of interest have not yet been experienced and observed in the target.

This difference is crucial. While it may be possible to measure both household endowment X and welfare Y in the target before the intervention, as well as potentially suspected mediating variables of the envisioned effects, observing these baseline values will often not be informative for constructing counterfactuals. This is because observing naturally occurring levels of these variables is similar to only observing individuals that were not exposed to AFB1 in Steel's example. While we can measure whether they were exposed to AFB1, these measurements will return the same value for all individuals, i.e. zero. This means that essential observations for constructing counterfactuals are missing, i.e. observations that include *differences* in the suspected cause of interest as well as suspected mediating variables.

Our microfinance case differs slightly. Here, our intervention variable X, household endowment, will naturally assume values other than zero. Crucially, however, natural variation in X is not necessarily the same as variation in X induced by an intervention. The importance of this becomes clear when considering attempts to establish correlations among variables to guide inference about the presence of suspected causal relationships. Observing that high levels of household endowment are co-instantiated with high levels of welfare does not reliably indicate that *increasing* endowment would *increase* welfare – standard concerns about common causes are pertinent here, as are concerns about structure-altering interventions. Just because endowment and welfare naturally covary, this does not guarantee that an intervention on endowment will change welfare, since that very intervention might disrupt or otherwise meddle with the relevant structural relationships, might induce unanticipated counteracting effects, and so forth.

Similar concerns about limited informativeness apply to the putative mediators and ultimate effects of the suspected cause, e.g. changes in entrepreneurial activity, household consumption, and ultimately in welfare. In predictive extrapolation cases, I assume that the intervention of interest has not yet been implemented in the target and that its effects have not been experienced there yet. This means that the kind of variation in Y that is a candidate for being causally attributed to variation in X induced by our intervention has not been observed. While one may measure household welfare levels from survey data, this does not permit straightforward construction of a counterfactual, since what one observes is only natural variation in Y that is known to not have been caused by our envisioned intervention on X, for lack of it having been implemented. This makes the construction of a counterfactual Q_B significantly more difficult. *Figure 4* encodes this graphically by representing Q_B, y_j', and k_j' as unobserved (or at least difficult to observe), which is indicated by dashed symbols. The dashed diamond symbol for x_j' and the $do(X = x')$ arrow indicate that the counterfactual of interest is one where X would have been set by an intervention to $X = x'$.[3]

[3] Strictly speaking, we might say that in Steel's example no intervention on X has taken place in the past either, since it was not experimentally controlled exposure, but rather agents' natural behaviours that led some to be exposed to AFB1 and left others unexposed. This is noted, and the use of the do-operator here can be understood in a more relaxed way, as merely indicating that there are observations available on agents who have been exposed to AFB1, although perhaps not by an investigator-induced intervention.

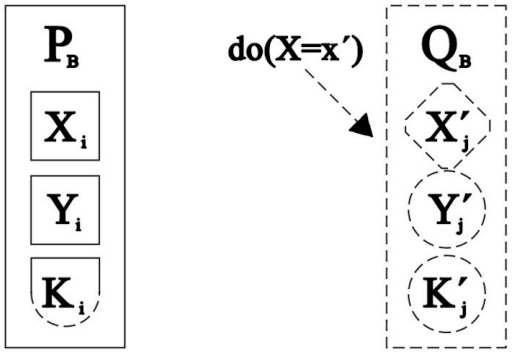

Figure 4: Q_B cannot be construed for lack of observations Xj', Yj', and Kj'

Figure 4 helps highlight the crucial differences between attributive and predictive extrapolation. The aim in the microfinance case is not to *attribute* the differences between y_j' and y_i to differences in x_j' and x_i, rather than differences in some other variable in K. Instead, the aim is to *predict* the counterfactual Q_B if X were to be set to $X = x_j'$ by an intervention. In other words, the question is: *what is the counterfactual Q_B, specifically with respect to y_j', if X* were to be set to $X = x_j'$. Obtaining such a prediction of y_j' helps to answer various causal queries. For instance, if y_j' is greater than y_i, we may conclude that an intervention that sets X to $X = x_j'$ will be positively causally relevant for Y on average. Moreover, computing the difference between y_j' and y_i yields the *magnitude* of the causal effect of the intervention on X.

This makes clear that there are differences between predictive and attributive extrapolation both in the causal claim to be extrapolated as well as in the evidence that is available to help extrapolate that claim. As the differences in the aims of the extrapolation are now clear, let me expand in more detail on the differences in the evidence available to facilitate attributive and predictive extrapolation.

Attributive extrapolation uses experimental evidence from a study population, observational evidence from the target, and evidence of similarities in mechanisms to infer that a specific *cause*, rather than some other cause, has causally contributed to bringing about certain characteristics of the observations in the target.

Predictive extrapolation, on the other hand, aims at predicting counterfactuals based on experimental evidence from the study population, observational evidence about the pre-intervention states of the target, and evidence of similarities in mechanisms. Its aim is hence to infer what *effect* will obtain in the target, if its suspected cause were intervened on in such-and-such way.

These differences have important ramifications for whether it is possible to evade the extrapolator's bind. In the case of attributive extrapolation, Steel demonstrates persuasively that it is possible to establish a strong case for extrapolating the attributive claim of causal relevance by means of evidence about similarities in mechanisms. But there is more evidence involved in this extrapolation than just evidence of similarities and differences in mechanisms. More specifically, the evidence involved in supporting the conclusion that AFB1 is causally relevant for the production of HCC in humans consists of at least the following ingredients:

1. Evidence from RCTs on animals that AFB1 is causally relevant for the production of HCC in some species
2. Evidence from in vivo and in vitro animal experiments characterizing the mechanism that governs the production of differences in HCC from differences in AFB1 exposure in some species
3. Evidence from animal experiments suggesting that there is between-species variation in both the quality as well as the magnitude of the effect
4. Evidence indicating that the between-species differences in carcinogenic efficacy in animals are related to differences at certain stages of metabolic mechanisms
5. Observational evidence on AFB1 exposure and HCC prevalence in human populations, such that the evidence covers all combinations of states that these variables can assume
6. Observational evidence concerning K and K', specifically concerning the prevalence of known alternative causes of HCC other than AFB1 in human populations (e.g. HBV infection)
7. Evidence of similarities and differences in mechanisms, i.e. observations of the presence or absence of characteristic marks and symptoms of the hypothesized mechanisms in humans, specifically concerning stages at which mechanisms are likely to exhibit between-species differences

The aim of Steel's account is to highlight the importance of 7), i.e. the evidence afforded by CPT. However, it is important to recognize that while CPT evidence makes an important difference to the overall strength of the extrapolation, it does so only in conjunction with the additional evidence that is available.

My main concern here is that in predictive extrapolation, analogues of 2)-6) are often not available. In these cases, it seems unlikely that CPT evidence would be similarly effective in providing a strong case for an extrapolative conclusion. What is more, without relatively complete knowledge of the mechanism in the experimental population already being available, it is doubtful whether CPT evidence is feasibly producible at all, since in order to *compare* mechanisms in study and target populations, one needs to know at least some characteristics of the mechanism in the study population, i.e. one needs to have evidence corresponding to 2)-4) in the above list. As I have argued in *Section 2*, such evidence is often not available in social science settings.

CPT evidence hence cannot do the same job as in Steel's example if little additional evidence is available to complement it, or the evidence that is needed for the *comparative* goals of process tracing to be realizable in the first place is absent.

This seems to suggest that even if CPT is feasible at all in cases where some or many of the above kinds of evidence are missing, the evidence of similarities in mechanisms may need to be significantly more extensive to reach the same level of support for extrapolating claims of causal relevance.

Finally, there is the issue of scope that I have emphasized above. Not only do typical social science and EBP cases of extrapolation differ significantly with respect to the availability of crucial kinds of evidence that are required for CPT, but also, the very aims of extrapolation are often different as well. Whenever the observational evidence from the target does not include observations where the suspected causes and their putative effects are realized, then the extrapolation is concerned with an altogether different kind of question: namely, to predict characteristics of the counterfactual that would be realized if the intervention were to be implemented in the target. As I have suggested, but not yet fully argued, such predictive extrapolation seems to be a significantly more epistemically burdensome activity than the attributive case that Steel considers. To help explain these epistemic challenges in more detail, let me construct a more detailed example.

5.4 CPT in Action: Predicting the Effectiveness of HIV Prevention Interventions

To build my example, I will draw broadly on some empirical literature regarding the effectiveness of HIV prevention interventions (e.g. Owczarzak et al. 2018; Sagherian et al. 2016; Covey et al. 2016), but will abstract away from concrete interventions and studies to focus my example on those aspects important for illustrating the general points I wish to make here.

Suppose we have a broad evidence base consisting of several RCTs in different populations, indicating that a specific sexual-behaviour change programme is highly effective in decreasing participants' risk of acquiring HIV. Let us assume that the intervention consists of reproductive health counselling sessions seeking to increase agents' understanding of HIV transmission and sexual assertiveness, as well as the distribution of free male condoms. Suppose further that the intervention has largely been tested in populations that, while heterogeneous in potentially important socio-demographic characteristics, have so far not included type-B individuals (where type-B may stand for any relevant subpopulation). Let us also assume that, for some reason, it is plausible to suspect that type-B individuals are both a highly vulnerable subpopulation (facing a high baseline probability of HIV infection) and may respond differently to the intervention of interest, e.g. the social contexts in which sexual behaviours of type-B individuals proceed might differ importantly from those inhabited by individuals represented in the studies conducted so far. Let us assume that our extrapolative query is whether the intervention of interest will be similarly effective in a population consisting of type-B individuals as in the populations studied so far. In accordance with my understanding of predictive extrapolation I will assume that the intervention of current interest has not yet been experienced by individuals in this target.

In addition to this outline of the intervention and populations studied, I will also need to make some assumptions concerning the intervention and outcome variables at issue and potential mediators of the effects of interest. At the most general level, let me assume that the sexual-behaviour change programme is represented by a binary exposure variable $X \in \{0,1\}$, i.e. whether an individual has participated ($X = 1$) in the programme or not ($X = 0$) (or has *intended* to participate, as in standard *intention to treat* designs, see e.g. Owzcarzak et al. 2018, 919). The outcome Y will be HIV infection status. This is notably different from typical outcomes measured in evaluations

of real-world sexual-behaviour interventions. Here, clinical endpoints such as HIV infection are usually not the main outcome of interest, but the focus is rather on plausible mediators, such as self-reported condom use, and aspects of sexual communication behaviours, such as sexual assertiveness (to be explained shortly). For the present theoretical arguments, however, it will be better to focus on the clinical endpoint, given my concerns about mechanism-based extrapolation using CPT focus on whether we can assert that suspected mediating variables of the effects of interest are involved in the right way in producing the outcome in the target (or preventing it, as in the case of HIV infection).

With this in mind, let me assume that, at the most general level of abstraction, the causal mechanism underlying the effect of interest is $X \rightarrow \boldsymbol{Z} \rightarrow Y$, where $\boldsymbol{Z}$ is a 'box of moderators and mediators', i.e. a set of variables $[Z_1, Z_2, Z_3 \ldots Z_n]$ that, sequentially, or in parallel, mediate and moderate the effects of interest. We can think of $\boldsymbol{Z}$ as comprising of all (or a broad range of) factors that are relevant to determining agents' sexual behaviours and ultimately (in conjunction with yet other background variables) their HIV outcomes Y. The envisioned effects of the behavioural intervention of interest would then supposedly be mediated and moderated by at least some variables in $\boldsymbol{Z}$. Let me offer some more detail on the supposed mechanism governing these effects, starting with a diagram of the mechanism:

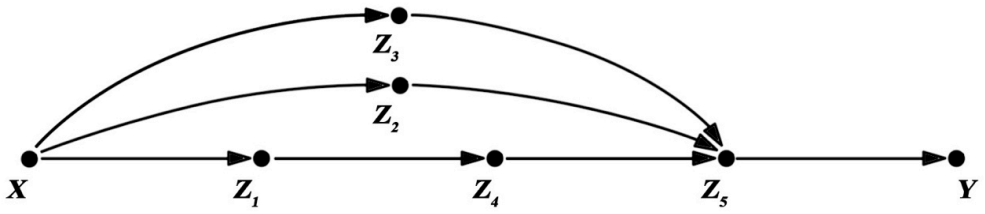

Figure 5: The suspected mechanism governing HIV infection prevention

Here, the intervention on X is supposed to achieve its ultimate aim of reducing the rate of HIV infection Y by inducing effects along three parallel pathways. First, it is supposed to increase agents' *understanding of HIV transmission* Z_1. Second, it should increase *condom availability* Z_2, by distributing free male condoms. Third, it should increase agents' *sexual assertiveness* Z_3, by which I mean agents' ability to promote the

enforcement of their desired use of condoms.[4] There are two additional mediators, Z_4 and Z_5, which are added for plausibility. First, Z_4 is agents' *intent to use condoms*. Increasing agents' understanding of HIV infection mechanisms and how condoms figure in these (Z_1) is supposed to help agents form suitable intentions to use condoms. Second, Z_5 is *condom use*, which is influenced by Z_1 (through Z_4), Z_2, and Z_3.

Favourably for Steel's strategy, condom use acts as a downstream bottleneck mediator of the effects of X on Y. For instance, and assuming interaction between the effects transmitted through Z_{1-4}, irrespective of whether agents are highly informed about HIV transmission, if their sexual assertiveness is low, or their partners do not want to use condoms, then understanding of HIV transmission and availability of condoms may fail to translate into condom use.[5] Likewise, intent to use condoms and high levels of sexual assertiveness alone may not be enough to translate into condom use if condoms are unavailable. Conversely, with condom use being a bottleneck, the diagram assumes that neither condom availability, nor intent to use condoms, nor sexual assertiveness can have effects on HIV infection other than through condom use.[6]

With this sketch of the mechanism in place, let me explain how CPT can experience difficulties in handling predictive extrapolation in this case. Recall that, to clarify whether the intervention of interest will be similarly effective in a novel target consisting of type-B individuals, we need to ensure that the causal mechanisms connecting the intervention and outcome variables are sufficiently similar between the experimental and target populations. For this, our main aim will be to ascertain that distinctive marks or symptoms of important parts of the mechanisms believed to govern the effects of interest in the experimental populations are also instantiated in the target.

---

[4] Two important things must be noted here: first, sexual assertiveness is not a success term, but rather an attitude. It can obtain even if attempts to enforce one's own desires about e.g. condom use are not always successful. Moreover, the characterization of sexual assertiveness used here is but one among potentially many others, including the ability to communicate one's desires to have sex, the ability to refuse unwanted sex, and the ability to communicate about sexual history and risk (see e.g. Loshek and Terrell 2015). I focus on one aspect merely for reasons of simplicity and not out of ignorance of, or disregard for, other important aspects of sexual assertiveness and their role in condom use negotiation.

[5] Specifically, Z_{2-4} can be understood both as mediators of the effects of X on Y as well as moderators of each other's effects. This would be the case, for instance, if Z_5 were determined in accordance with $Z_5 = Z_2 * Z_3 * Z_4$, which means that no change in Z_5 can obtain unless Z_{2-4} all take on values other than zero.

[6] This is an assumption of convenience, of course, since both understanding of HIV transmission and sexual assertiveness may still be highly effective in reducing agents' probability of HIV infection simply by inducing agents to refuse having unprotected sex, hence bypassing the downstream bottleneck. This is bracketed here only to ensure the applicability of Steel's CPT.

What could these marks or symptoms be in our example? Importantly, unlike in Steel's AFB1 example, we will be concerned here with the *arrows* to and from variables rather than the presence or absence of *variables* (or specific values of variables). In Steel's AFB1 example, and in many biomedical examples more generally, it is often the presence or absence of a variable, or a close concomitant of a variable, say e.g. DNA adducts (Steel 2008, 91), that is of interest for determining whether mechanisms are sufficiently similar. This is especially so when these variables are typically absent without the suspected mechanism being operational and their presence (or their assuming a particular value) in turn indicates the suspected mechanism being at work. In many social science settings, by contrast, variables such as household income, features of economic choice behaviours, health indicators, educational performance, etc. often exhibit natural variation, i.e. they naturally take on values other than zero irrespective of the details of the causal mechanisms in which they are involved. To stick with our example, condom use is a variable that will often take on values other than zero in a population B irrespective of whether the suspected mechanism from A is also at work there. Condoms can be used by people for a variety of reasons, even if the mechanism by which the intervention is supposed to be effective is not present or is unhelpfully disrupted. So, condom use as such is not a distinctive mark or symptom of the suspected mechanism being operational. It is hence not the presence or absence of such variables that is important to look at, but rather the presence or absence of causal relationships between them. In the language of mechanisms and diagrammatic representations of them, we are interested in the presence or absence of *causal arrows*, and distinctive marks or symptoms of their presence, absence, and quality.

Following Steel's strategy, let us now have a closer look at the downstream bottleneck Z_5 to help clarify how questions about mechanistic similarity and difference could be answered. What can we say about the arrows into and out of Z_5 in the target population? The arrow $Z_5 \rightarrow Y$ is plausibly believed to be invariant between populations. For the sake of argument, we may assume that if condoms were to be used, they would be used properly, so we can lay aside concerns about the nature of this arrow and potential differences in the causal effects of Z_5 on Y.

Importantly, however, even if it is uncontroversial to assume the presence of the $Z_5 \rightarrow Y$ arrow, or even if this were supported by observational evidence from the target or other, similar populations, this alone does not licence further inferences about

upstream similarities in mechanisms, which is how Steel proposes to evade the extrapolator's bind. Unlike in Steel's AFB1 example, where individuals have already been exposed to AFB1, our predictive extrapolation scenario assumes that target agents have not yet experienced the intervention of interest. So even if some agents in the target have been using condoms before, it will not be *because* they have been exposed to our intervention. Hence, the mere fact that some target agents are already using condoms is no indication of whether the mechanism from the experimental population is also instantiated in the target, and hence whether agents in the target, including specifically those who have not yet been using condoms, would be successfully induced to use condoms by our intervention. This inferential shortcut is not available in predictive extrapolation. I will say more on this important issue later. For now, let me proceed to discuss how we may clarify whether other causal relationships believed to be important for the effects of interest are suitably instantiated in the target.

What about the arrows from Z_2, Z_3, and Z_4 into Z_5? This is where things will get more difficult for CPT. Let me start with the $Z_2 \rightarrow Z_5$ arrow. Our aim here will be to find distinctive marks of the efficacy of increasing condom availability for increasing condom use. In fortunate cases, where condoms have been available in the target in the past, and at least some individuals in the target have been using condoms, we may use observational data to support that the $Z_2 \rightarrow Z_5$ arrow is in place. Things will be more difficult if condoms are so far mostly or entirely unavailable or unaffordable in the target, particularly in non-clinical outlets and at locations and times relevantly close to sex. If condoms are unaffordable or unavailable in the target, then it will be difficult to learn from observation alone whether, if condoms *were* to be made available, they *would* actually be used. For instance, the $Z_2 \rightarrow Z_5$ path might be disrupted in the target because agents hold religious or superstitious beliefs that induce them to reject using available condoms. This may be the case even for agents with sophisticated understanding of HIV transmission. Clearly, when condoms are not available nor used by agents, observations of condom availability (none) and use (none) are not elucidating, since potential disruptions of the suspected $Z_2 \rightarrow Z_5$ path, e.g. due to religious and superstitious beliefs, can only manifest themselves in observational data if condoms were already widely available. While we may sometimes be able to rule out specific concerns about such disruptions by drawing on (local) background knowledge about the target population, this may not always be enough to increase our confidence in the presence of the $Z_2 \rightarrow Z_5$ arrow beyond some desired threshold. What is missing is

that we can see relevant parts of the suspected mechanism 'in action'. As suggested, in our case this could comprise, for instance, of observations of individuals who already have access to condoms and actually use condoms. Yet, without condoms being available in the target so far, our beliefs in the presence of the suspected arrow $Z_2 \rightarrow Z_5$ will need to be supported by means other than observations of target agents' behaviours.

What could these alternatives be? We might consider asking agents whether they would use condoms if they were freely available. Yet, while this can sometimes be helpful, especially if agents are already familiar with condoms, it might be a poor guide to predicting condom use if agents have no previous experience in using condoms (as is plausible, for instance, in some development settings). Crucially, what may be required here to learn whether agents would use condoms, if they were available, is to introduce them to condoms, including their functional relation to the prevention of sexual risks such as HIV infection. Hence, while interviewing target agents about their predicted condom use behaviours can be informative for predicting the effects of interest in the target, one might worry that doing so effectively, i.e. in a way that promises to reliably extract the required information, may require exposing agents to the very intervention whose effectiveness we are interested in predicting, thus gradually undermining the *success* of our extrapolative inference.

More pressing concerns arise when considering the second causal pathway, mediated by $Z_4 \rightarrow Z_5$ and $Z_1 \rightarrow Z_4$, which is hoped to govern the envisioned increase of agents' understanding of HIV transmission and subsequent increases in their intent to use condoms. How can we clarify whether these relationships are suitably instantiated in the target? As suggested above, we might be lucky to encounter favourable conditions where natural variation in agents' understanding of HIV transmission already exists, so at least some agents exhibit levels of understanding similar to those sought to be induced by the intervention. In such cases, we might want to look for whether those agents who are, for extraneous and pre-existing reasons, better informed about HIV transmission are also more likely to (intend to) use condoms. This information could be elicited through interviews and questionnaires, for instance. If we were to find, say, that individuals who are better informed than others with respect to HIV transmission report significantly higher willingness to use condoms, this could strengthen the basis for the assumption that mechanisms are relevantly similar at $Z_1 \rightarrow Z_4$ in at least some agents in the target. This might also give us reasons to believe that if our intervention were

successful in increasing understanding of HIV transmission for those agents without previous understanding thereof, then this would yield similar differences in intended condom use as those promised by the observed differences obtained through questionnaires. Analogously, regarding $Z_4 \rightarrow Z_5$, we might want to look for whether agents who express higher willingness to use condoms also report higher condom use, which could increase our confidence that the $Z_4 \rightarrow Z_5$ relationship is also suitably instantiated in the target.

However, there may also be less favourable cases where there is little in the way of previous understanding of HIV transmission or intent to use condoms on the part of target agents, e.g. for lack of previous awareness of HIV, previous experience in using condoms, or understanding of their functioning in relation to HIV infection. Observing agents' behaviours in such cases will be a poor guide to telling whether mechanisms are sufficiently similar, since there are no agents with suitable levels of HIV transmission understanding and intent to use condoms that could serve as exemplars for predicting whether other agents would intend and behave similarly if their understanding of HIV transmission were to be intervened on. Without such an understanding already in place, and without observing any of the associated differences in intent and behaviour ultimately hoped for, we will need, again, to support our beliefs in mechanistic similarity at these stages by means other than observation.

Interviewing target agents might be unhelpful here, too. Asking agents, for instance, whether they believe that their intent to use condoms *would* increase if they *were* to be better informed about HIV transmission might not be illuminating if agents lack the very understanding of HIV transmission and condom functioning that the intervention is supposed to promote. Similarly, regarding $Z_4 \rightarrow Z_5$, asking agents whether they think that their *hypothetically* increased intent to use condoms would translate into actual use of condoms might be uninformative for largely the same reasons. Without already intending to use condoms (which is part of what the intervention is supposed to achieve), it will be difficult for agents to reliably predict whether they would be able to enforce such hypothetically formed intent in interaction with their sexual partners. This makes clear that neither observational evidence, nor more intimate forms of information, such as from interviews and self-reports, can be expected to figure as reliable guides for telling whether mechanisms are sufficiently similar between populations when agents lack prior experience with the sorts of changes that the

intervention is supposed to bring about. Again, the crucial concern here is that learning information that would be needed for supporting our extrapolation might require exposing agents to the very changes that our envisioned intervention is supposed to induce. Once again, doing so could be highly informative, but would also threaten to fall prey to the extrapolator's bind.

Similar concerns arise when considering $Z_3 \rightarrow Z_5$. Here, our aim is to tell whether *hypothetical* increases in agents' sexual assertiveness would yield subsequent increases in agents' condom use. Again, favourable circumstances might arise where observations and local knowledge about pre-intervention sexual assertiveness can help support assumptions about this relationship. For instance, conditional on condoms being available in the target in the past and other things being equal (e.g. agents' intent to use condoms) we may find that agents with higher sexual assertiveness (self-reported or otherwise measured) are significantly more likely to use condoms. This would be helpful for increasing our confidence in the $Z_3 \rightarrow Z_5$ relationship being present in the target.

Less favourable cases will again make this more difficult, however. Consider cases where target agents generally have low pre-intervention levels of sexual assertiveness, so low indeed that agents do not differ much in this respect from one another. Here, it would be difficult to tell whether higher levels of assertiveness sought to be induced by the intervention would translate into condom use, as there might be unanticipated ways in which the $Z_3 \rightarrow Z_5$ relationship could be disrupted in the target. Again, observing agents' sexual assertiveness and condom use will be a poor guide to telling whether increases in assertiveness would yield increases in condom use if both variables are realized at very low levels. Without elements of the intervention of interest already experienced by agents, we cannot observe parts of the suspected mechanism 'in action' to help us clarify issues of mechanistic similarity and difference.

Again, one alternative could be to interview target agents. Yet, for largely the same reasons as outlined above, this is unlikely to be helpful. If sexual assertiveness is poorly expressed in target agents, and indeed perhaps the very idea of being sexually assertive is unfamiliar to them, asking agents whether they *would be* more likely to use condoms if they *were* more sexually assertive is unlikely to provide useful information that bears on questions of mechanistic similarity and difference. This is no surprise. It is difficult for people to anticipate their behavioural response to things that they are mostly or

entirely unfamiliar with. Even if agents grasp the concept of sexual assertiveness, their effective ability to enforce condom use may nevertheless face unanticipated obstacles at the time of sex, e.g. when interacting with partners who are themselves highly assertive, have disproportionate relationship power, or, perhaps due to insufficient understanding of HIV transmission on their part, will successfully insist on not using condoms. Outside of previous familiarity with the kinds of changes sought to be induced by our intervention, it seems that, once again, we might need to introduce (parts of) the very intervention whose effectiveness we want to predict in order to tell whether mechanisms are sufficiently similar to licence extrapolation. And once again, this threatens the success of our extrapolation.

With these specific concerns in place, let me take some steps backwards from the concrete details of the example to highlight several general insights about the challenges involved in using CPT for predictive extrapolation.

Before all else, it is important to note that predictive extrapolation is not an in-principle insurmountable obstacle for CPT. As suggested above, there will be more favourable cases where, despite some additional epistemic burden placed on us, predictive extrapolation can nevertheless be achieved with the help of CPT, e.g. when good observational data from the target are available to indicate important features of the suspected mechanisms being present there, or where there is stronger overlap between experimental and target populations, or where background theory strongly motivates belief in essential mechanistic similarities, etc. However, there will also arguably be many other cases in which the distinctive features of predictive extrapolation will make successful uses of CPT extremely difficult or entirely infeasible.

One of the great promises of CPT, especially with a view towards evading the extrapolator's bind, is that it offers us an inferential shortcut. When comparing mechanisms and trying to avoid learning the full mechanism in both populations, Steel's suggestion is to focus on comparisons at downstream bottleneck stages Z. The shortcut provided by this is that as long as changes in X transmit down to the bottleneck Z, either mechanisms are similar upstream of Z or upstream differences do not matter. In many predictive extrapolation settings, however, this shortcut does not work. Observing, say, that agents with higher self-reported intent to use condoms are indeed more likely to use condoms does not reliably indicate that (hypothetical) increases in

HIV transmission understanding would translate into increased willingness to use condoms. In predictive extrapolation, distinctive marks or symptoms of individual relationships comprising a suspected mechanism are often not reliable indicators of the presence (or absence) of yet other relationships further upstream, or indeed an entire mechanism, unless these relationships or the whole mechanism were already (and distinctively) involved in producing the marks or symptoms at issue. This is precisely what makes predictive extrapolation so challenging for CPT. If whatever mechanism that will ultimately govern the effects of interest in a target setting has so far remained 'dormant' (e.g. due to lack of previous changes in relevant variables), then neither observations, nor agents' reports on details of such mechanisms, will usually be a reliable guide for answering questions about mechanistic similarity and difference. Here, we cannot infer upstream similarity or the irrelevance of upstream differences from downstream similarity, so Steel's shortcut fails. The information required by CPT, and that is available in attributive extrapolation, is afforded by the mechanism in the target being 'awake' and observed 'in action'. If this is not the case, other information is needed, i.e. information with bearing on how the mechanism *would* operate if it were to be intervened on, e.g. information obtained from interviews, local knowledge, background theory, and so forth. As suggested above, not all predictive extrapolation cases will be void of such information, and many might be more favourable, e.g. when at least parts of the targets' mechanisms have been in some relevant sense 'active' and allow both observation and other methods of eliciting information with bearing on issues of mechanistic similarity and difference to function properly.

Real-world cases, of course, can sit anywhere on this spectrum, and while I do not wish to take a stance on their distribution with respect to how challenging they are, it seems useful to briefly summarize several dimensions relevant to a general assessment of whether substantial difficulties are likely to be looming in predictive extrapolation.

First, little or no variation in intervention and suspected mediating variables of the mechanisms of interest will mean that observational data will be of little use in clarifying issues of mechanistic similarity and difference. Idle causes, we might say, make for idle inferences. This is because variation is often needed to tell, for instance, if there are distinctive correlations that would be expected to obtain if a certain causal relationship were present in the target. Without variation, there is no co-variation and hence no informative correlation, so other kinds of information will be needed to

underwrite our inferences. This point about the importance of covariance information will be more fully elaborated in the next chapter.

Second, interview- and questionnaire-based evidence concerning agents' predictions of their own behaviours under hypothetical interventions may sometimes be highly informative, but the reliability of such information is often questionable, particularly in cases where the interventions of interest have so far not been experienced by agents.

Third, in some cases, other, past interventions or exogenous changes in the target can be a useful guide towards addressing questions of mechanistic similarity and difference at higher levels of abstraction. If, say, other interventions seeking to inform or educate agents have been highly effective in the target in the past, this may increase our confidence that at least the educational aspects of our HIV prevention intervention might be similarly effective. This, of course, will need to involve thicker inferences, including the assumption that, say, past agriculture- and present sexual health-related information campaigns are similarly well-received by agents. More generally, inferences from the effectiveness of other, past interventions to those of current interest will need to involve assumptions about crucial similarities between them, and while not in-principle infeasible, supporting such assumptions can be extremely difficult, and, at times, require knowledge of parts of the very causal features that we need to support assumptions about. Generalizing from one intervention to another will often require a detailed understanding of how the interventions work, including identifying relevant parts of the mechanisms by which their effects are supposed to be transmitted as being relevantly similar, or alternatively (and glossing over mechanistic details) identifying the interventions as members of a common type and justifying a generalization across the type.

Fourth, there is a persistent and well-known concern about structure-altering interventions (see e.g. Steel 2008, 157-60), which is also pertinent to predictive extrapolation. Specifically, while fortunate cases with ample pre-existing variation in suspected mediating variables make it easier to clarify issues of mechanistic similarity and difference, there is nevertheless an important limitation to the use of such information, as we need to further assume that the envisioned intervention will yield similar effects as those induced by pre-existing variation in variables. To stick with the example, it is unclear whether naturally occurring differences in HIV transmission understanding have the same effects on intent to use condoms as the differences in HIV

transmission understanding that are envisioned to be induced by our intervention. For instance, agents who exhibit, for reasons unrelated to our intervention, highly developed previous understanding of HIV transmission might have 'self-selected' into becoming more knowledgeable about HIV as a function of some other important characteristic that is also relevant to shaping their sexual behaviours. Concerned that naturally acquired understanding of HIV transmission and the understanding facilitated by our envisioned intervention might be different in shaping agents' behaviours, we hence need to further assume that the effects of our envisioned intervention would manifest in the same ways in the target as the effects of naturally occurring differences in suspected mediating variables observed there. Such assumptions can be difficult to support.

Fifth, local knowledge from other, believed-to-be-similar populations can be extremely helpful. Yet, as is the case with using information about the effects of similar interventions in the target in the past, a substantive additional inference is required for establishing the relevance of such information to questions about the target. Such inference is possible, but also, again, requires further support, including, for instance, a strong inductive basis that licences inference to the target on the grounds of it being a member of some generalizable population-type with respect to mechanistic details.

Finally, background theory, too, might be extremely helpful for underwriting predictive extrapolation. Some causal relationships are relatively easy to generalize, e.g. that individuals' health outcomes are adversely affected by lethal doses of neurotoxins; more mundanely, that condoms are highly effective at preventing STI infection when used properly; or, more concretely, that condoms need to be available not just in general, but at times and locations relevantly close to sex. But background theory, like all other candidate sources of support, is not always available, nor does it always offer satisfactorily high levels of support for the assumptions needed. As with supporting inference where causal information is imputed from other populations or other interventions, background theory will need to be strong enough to reliably identify the target population as a member of a certain population type, which then licences imputing causal information from elsewhere to support assumptions about the suspected mechanisms of interest.

Predictive extrapolation, then, poses distinct challenges for Steel's CPT, particularly with a view towards its ability to evade the extrapolator's bind. If the causal mechanisms of interest in the target have so far remained 'dormant' and, for lack of

manifestation, their characteristics are difficult to identify from available observations or by means of interview- and questionnaire-based methods, other sources of support are required for CPT, including inference from other populations or past interventions; background theory; or any combination of these. As suggested, even if such resources are available, they will often provide weaker support and require more involved inferences, making it more likely that adequately justified predictive extrapolation will require additional first-hand causal information from the target that threatens to make the experimental results from which we wish to extrapolate less relevant to our envisioned conclusions.

5.5 Conclusions

I have argued that the scope of Steel's mechanism-based strategy for extrapolation is constrained to a specific class of problems of extrapolation. As Steel recognizes (2008, ch.8 – but not for the reasons provided here), this severely limits its usefulness for social science purposes.

First, the applicability of Steel's strategy is limited to cases where one has knowledge 1) of the stages at which differences between mechanisms are most likely to pose obstacles to extrapolation; 2) of which of these stages are downstream stages; 3) that mechanisms are consonant for all individuals in both populations; and 4) that downstream stages of mechanisms are bottlenecks. I have argued that these epistemic demands constitute a high lower bound on the mechanistic knowledge required for extrapolation.

In addition, and following some of Steel's concerns, I have argued that mechanisms encountered in social science often exhibit features that make it difficult to use CPT. There, mechanisms are often more complicated than the single-path mediated mechanism in Steel's examples; they might exhibit considerable heterogeneity between individuals; and they are often significantly more difficult to observe, directly and indirectly, thus limiting the extent to which mechanisms can be compared at all.

Adding to these general concerns, I have argued that the ability of Steel's strategy to evade the extrapolator's bind varies significantly between two kinds of extrapolation, attributive and predictive, which exhibit two important differences. The first concerns the aims of the extrapolation, i.e. to *attribute* observed effects to their suspected causes,

or rather to *predict* the future effects of (an intervention on) a suspected cause. The second, more important difference concerns the evidence typically available in these settings, i.e. evidence that comprises observations where both the suspected cause as well as its putative effects are realized in attributive extrapolation versus evidence where the suspected cause and the mechanisms governing its effects have so far remained 'dormant' in the target in predictive extrapolation. I have argued that predictive extrapolation imposes more severe epistemic demands on CPT as crucial information from the target remains unavailable if the target's mechanism has not been observed 'in action'. Here, one may often need to consider alternative kinds of support for crucial causal assumptions, such as background theory, information from past interventions, or from yet other populations. If these sources of support are unavailable, or are insufficiently strong to underwrite our assumptions, one may need to intervene on the suspected cause of interest in the target, thus falling prey to the extrapolator's bind.

In the next two chapters, I will argue that Steel's CPT is not alone in experiencing difficulties in handling predictive extrapolation. I will develop analogous arguments concerning strategies for extrapolation proposed in econometrics and computer science that seek to licence quantitative predictions of how causally relevant differences between populations will bear on differences in the magnitude of causal effects in a novel target. Perhaps unsurprisingly, this type of extrapolation is significantly more demanding than licencing merely qualitative conclusions, and the concerns developed here present even more pressing obstacles to such ambitions.

References

Banerjee, A., E. Breza, E. Duflo, and C. Kinnan. (2017). "Do credit constraints limit entrepreneurship? heterogeneity in the returns to microfinance". Buffett Institute Global Poverty Research Lab Working Paper No. 17-104.

Beach, D. (2017). "Process-Tracing Methods in Social Science". Oxford Research Encyclopedia of Politics. Retrieved 12 December 2017, from http://politics.oxfordre.com/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-176.

Beach, D., and R. B. Pedersen. (2016). *Causal Case Studies: Foundations and Guidelines for Comparing, Matching and Tracing*. Ann Arbor: University of Michigan Press.

Covey, J, H. E. Rosenthal-Stott, and S.J. Howell. (2016). "A synthesis of meta-analytic evidence of behavioural interventions to reduce HIV/STIs", *Journal of Behavioural Medicine*, 39(3): 371–85.

Davey, C., S. Hassan, C Bonell, N. Cartwright, M. Humphreys, A. Prost, and J. Hargreaves. (2017). "Gaps in Evaluation Methods for Addressing Challenging Contexts in Development", CEDIL PreInception Paper: London

Garcia, M. F., and L. Wantchekon. (2010). "Theory, External Validity, and Experimental Inference: Some Conjectures". *The ANNALS of the American Academy of Political and Social Science*, 628(1), 132–47.

Gold, L., N. Manley, and B. Ames. (1992). "Extrapolation of Carcinogenicity Between Species: Qualitative and Quantitative Factors", *Risk Analysis*. 12: 579-88.

Hanushek, E. A. (1979). "Conceptual and empirical issues in the estimation of educational production functions". *Journal of Human Resources*, 14(3): 351–388.

Heckman, J. (1992). "Randomization and Social Policy Evaluation". In: *Evaluating Welfare and Training Programs*, C. F. Manski and I. Garfinkel (eds.), pp. 201-230. Boston, MA: Harvard University Press.

Hengstler, J., B. van der Burg, P. Steinberg, and F. Oesch. (1999). "Interspecies Differences in Cancer Susceptibility and Toxicity", *Drug Metabolism Reviews*. 31: 917-70.

Holland, P. W. (1986). "Statistics and Causal Inference", *Journal of the American Statistical Association,* 81(396): 945–60.

International Agency for Research on Cancer (IARC). (1993). Some naturally occurring substances: Food items and constituents, heterocyclic aromatic amines and mycotoxins. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 56. Lyon.

Loshek, E, and H. K. Terrell. (2015). "The Development of the Sexual Assertiveness Questionnaire (SAQ): A Comprehensive Measure of Sexual Assertiveness for Women", *Journal of Sex Research*, 52(9): 1017-27.

Marcellesi, A. (2015). "External Validity: Is There Still a Problem?", *Philosophy of Science*, 82(5): 1308-17

Muller, S. M. (2013). "External validity, causal interaction and randomised trials: the case of economics", Unpublished manuscript.

Neyman, J., K. Iwaszkiewicz, and S. T. Kolodziejczyk. (1935). "Statistical Problems in Agricultural Experimentation", *Supplement to the Journal of the Royal Statistical Society,* 2(2), 107-80.

Owczarzak, J., M Broaddus, and S Tarima. (2018). "Effectiveness of an evidence-based HIV prevention intervention when implemented by frontline providers", *Translational Behavioural Medicine*, 8: 917-26.

Pawson, R., and N. Tilley. (1997). *Realistic Evaluation*, London: SAGE Publications.

— **(2001).** "Realistic Evaluation Bloodlines", *American Journal of Evaluation*, 22: 317-24.

Pawson, R. (2013). *The science of evaluation: a realist manifesto*. London: SAGE Publications.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligence Systems*. San Mateo, CA: Morgan Kaufmann.

Reiss, J. (2010). "Review: Across the boundaries: Extrapolation in biology and social science", *Economics and Philosophy*, 26: 382-390

Rubin, D. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology,* 66(5): 688–701.

Sagherian, M. J., T. B. Huedo-Medina, J.A. Pellowski, L.A. Eaton, and B. T. Johnson. (2016). "Single-session behavioural interventions for sexual risk reduction: a meta-analysis", *Annals of Behavioural Medicine*, 50(6): 920–34.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Schmitt, J., and D. Beach. (2015). "The contribution of process tracing to theory-based evaluations of complex aid instruments". *Evaluation*. 21(4): 429–47.

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science.* Oxford: Oxford University Press.

Todd, P. E., and K. I. Wolpin. (2003). "On the specification and estimation of the production function for cognitive achievement". *Economic Journal,* 113(485): 3–33.

Vivalt, E. (2019). "How Much Can We Generalize from Impact Evaluations?". Unpublished manuscript, ANU, Canberra. Retrieved from: http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf (retrieved, Feb. 28, 2019)

White, H. (2009). "Theory–Based Impact Evaluation: Principles and Practice", *The Journal of Development Effectiveness*. 1(3): 271-84

Wogan, G. (1966). "Chemical nature and biological effects of the aflatoxins", *Bacteriological Review*, 30 (2): 460-70.

— **(1992).** "Aflatoxin Carcinogenesis: Interspecies Potency Differences and Relevance for Human Risk Assessment". In R. D'Amato, T. Slaga, W. Farland, and C. Henry (eds.), *Relevance of Animal Studies to the Evaluation of Human Cancer Risk*. pp. 123-188. New York: Wiley.

— **(1999).** "Aflatoxin as a human carcinogen", *Hepatology*. 30(2): 573-575.

CHAPTER 6[1]

Interactive Covariate-Based Extrapolation

6.1 Introduction

Interactive covariate-based approaches have been proposed in the econometrics literature to help overcome problems of extrapolation (Hotz et al. 2005; Crump et al. 2008; Muller 2013 ms.; 2014; 2015). In a nutshell, these approaches aim to permit predictions of quantitative causal effects in a target, despite causally relevant differences. This proceeds by adjusting for differences in the distributions of *interactive covariates*, i.e. variables, such as moderating and mediating variables identified in *Chapter 2*, that can induce differences in causal effects between individuals and between populations. In virtue of being able to adjust for such differences, the approach extends significantly beyond the capabilities of other approaches, including using simpler construals of effectiveness arguments to tell whether causal effects will be the same in a target and using Steel's CPT, which can tell us when qualitative effects will be the same despite some causally relevant differences.

In what follows, I offer a critical discussion of the interactive covariate-based approach proposed by Hotz et al. (2005).[2] In *Section 2*, I offer a brief overview of how the approach works, including the assumptions it requires. In *Section 3*, following Muller (2014; 2015), I discuss existing concerns about the epistemic demands involved in underwriting these assumptions. In *Section 4*, I proceed to argue that underwriting these assumptions also raises important concerns about the extrapolator's bind. *Section 5* considers whether Steel's (2008) CPT can figure as a supplementary strategy to evade these concerns. While CPT can play such a role in principle, it faces distinctive challenges in *predictive* extrapolation cases and when using the quantitative observational evidence that is preferred by econometricians. In *Section 6* I argue that, to remedy this, econometricians need to consider other kinds of evidence, particularly *qualitative* evidence. This hence offers a novel argument for evidential pluralism and

[1] Parts of this chapter have been published in Khosrowi (2019).
[2] Some of the concerns developed here also apply to a recent contribution by van Eersel et al. (2019). This contribution, although interesting, will not be discussed here in more detail.

integration of different kinds of evidence in econometrics, EBP, and social science extrapolation more generally.

6.2 Interactive Covariate-Based Extrapolation

Interactive covariate-based strategies for extrapolation offered in the non-structural[3] econometrics literature (Hotz et al. 2005; Crump et al. 2008; Muller 2014; 2015) acknowledge that there are often likely to be important causally relevant differences between experimental and target populations. To overcome these obstacles, they propose a way in which we can still successfully extrapolate by taking such differences into account. In doing so, they focus on differences in the distributions of *interactive covariates*, i.e. variables W that can induce differences in causal effects between individuals and between populations, i.e. different kinds of moderating and mediating variables as discussed in *Chapter 2*.

To be sure, when following the outline of the strategy proposed by Hotz et al. (2005), it often seems that they are concerned with adjustment of causal effects by differences in mere *covariates* of the effect, i.e. variables that covary with different magnitudes of a causal effect. Yet, as suggested in *Chapter 2*, all kinds of things can covary with causal effects, including variables that have no capacity to induce differences in these effects. Descendants Z of the outcome Y are a good example, as they covary with causal effects but do not induce differences in effect magnitudes as they are downstream of the outcome. Similarly, when a moderator of an X-Y-effect is a common cause of a variable Z and the outcome Y (or some mediator), Z will also appear as a covariate of the X-Y-effect, but will otherwise have no bearing on the magnitude of that effect. While, if the relationships between Z and the mechanisms governing the effects on Y are invariant between populations (which requires corresponding assumptions), adjusting causal effects by distributions of Z can still permit accurate predictions, it will be crucial to single out *causally relevant* variables. This is particularly important if the aim is, for instance, to achieve a specific effect magnitude in the target. Doing so may often require co-interventions on moderators or mediators, so focusing on causally irrelevant variables would be detrimental here. In light of this, rather than focusing on mere

[3] Non-structural (micro-) econometrics, specifically the so-called *treatment-effects literature*, is concerned with estimating causal effects from experimental and quasi-experimental data. For detailed examinations of the differences between, and history of, structural and non-structural, reduced-form approaches see Boumans (2005); for a critical discussion see Keane (2010).

covariates, I take the aim of interactive covariate-based extrapolation to be to adjust for differences in *causally relevant* variables, specifically moderators and certain types of mediators. To capture this, and following existing emphasis on *causal*, rather than statistical interaction (e.g. Cartwright 1979; Muller 2013 ms.), I call these variables *interactive covariates*, i.e. variables that causally *interact* with the effects of an intervention and have causal bearing on the magnitude of these effects.

With this proviso in mind, the strategy of adjusting for differences in the distributions of interactive covariates between populations proposed by Hotz et al. (2005) proceeds in two steps. First, identify a causal effect in an experimental population as a *conditional average treatment effect* ($CATE$, see Muller 2014; 2015; see also Athey and Imbens 2016), i.e. an average treatment effect that is estimated conditionally upon the experimental populations' distribution of interactive covariates. This is supposed to capture how the average treatment effect in the experimental population hinges on that population's specific distribution of interactive covariates. Second, reweight the $CATE$ according to the observed distribution of interactive covariates in the target.

To give a stylized example, suppose that there is experimental evidence from an RCT that providing free textbooks to students in schools increases student performance on standardized tests. Let X be the number of textbooks that students have at their disposal and let Y be the outcome of interest, i.e. performance on tests. Suppose now that it is known that the causal pathway connecting the intervention variable X to the outcome Y is fully moderated by visual acuity W of individuals. Specifically, if students have extremely low visual acuity, they are unable to read the textbooks, and the effect of X on Y is completely suppressed. Suppose further that the distribution of W in the experimental population exhibits a high mean; students have 20/20 vision. Now, if one is interested in extrapolating the effect of distributing free textbooks to some target population, one will need to take into account the distribution of W in that population, since the effect may be suspected to vary over levels of W on the grounds that it does so in the experimental population. If, for instance, the mean of W is significantly lower in the target, then the effect of intervening on X intuitively expected in the target is likely lower as well. To tell us how exactly the effect in the target will differ, interactive covariate-based extrapolation proceeds by estimating the *CATE* of distributing textbooks in the experimental population, i.e. treatment effects that are estimated

conditionally on the distribution of W. When extrapolating, the prediction for the target is then computed by reweighting the experimental estimate across the observed distribution of W in the target population. The approach offered by Hotz et al. (2005) hence aims to permit quantitative extrapolation of causal effects in the presence of causally relevant differences by using suitably identified $CATEs$ from an experiment and quantitative observational data on the distributions of interactive covariates in the target.

Let me expand in more detail on the assumptions required for this strategy to proceed successfully. I adopt the notation of Hotz et al. (2005) and Muller (2014; 2015) that follows the standard Neyman-Rubin potential outcomes framework notation (Neyman et al. 1935; Rubin 1974).

Specifically, let $D_i \epsilon \{0,1\}$ denote the location of an individual i, where $D_i = 0$ indicates membership in the experimental population, and $D_i = 1$ in the target. Let $T_i \epsilon \{0,1\}$ be a treatment indicator, with $T_i = 0$ indicating no treatment and $T_i = 1$ indicating treatment respectively. Correspondingly, each individual has two potential outcomes, one without treatment $Y_i(0)$, and one with treatment $Y_i(1)$. Moreover, for each unit there is a vector of interactive covariates W_i.

The outcome for each individual is determined by:

$$Y_i \equiv Y_i(T_i) = T_i * Y_i(1) + (1 - T_i) * Y_i(0)$$

And the ATE for the experimental population is

$$ATE_{D=0} = E[Y_i(1) - Y_i(0) | D_i = 0]$$

In order to compute the ATE in the target population by adjusting for differences in the covariate vector W_i, three assumptions are required.

The first is that units' potential outcomes are probabilistically independent of treatment status (or assignment, in intention-to-treat cases), which is satisfied in ideal RCTs that achieve successful randomization, blinding of all participants, administrators of treatment, and evaluators, as well as absence of differential attrition between treated and untreated units as a function of interactive covariates. Formally, this is expressed as follows (Hotz et al. 2005, 246):

$$T_i \perp (Y_i(0), Y_i(1) | D_i = 0)$$

The second assumption*, unconfounded location*, is that all differences between locations $D_i = 0$ and $D_i = 1$ that are relevant for individuals' potential outcomes are differences in W_i (Hotz et al. 2005, 247; this mirrors one of the basic assumptions sketched out in *Chapter 3*; see also Angrist and Fernandez-Val 2013 for a similar assumption; see Stuart et al. 2018, 3 for a general discussion). In other words, while potential outcomes may vary between locations, they may not vary due to idiosyncratic features of the location that are not (or could not be) captured by some interactive covariate set W_i.[4] Conversely, this means that by conditioning on W_i, units' potential outcomes must be probabilistically independent of their location. Formally,

$$D_i \perp (Y_i(0), Y_i(1) | W_i)$$

Finally, the third assumption is *overlapping support* (Hotz et al. 2005, 247; see also Muller 2015, 5 who coined the term). It requires that there is overlap in the distributions of the interactive covariates that enter W_i, so that for all values $W_i = w$ that the interactive covariates in the experimental population assume, there is a non-zero probability that an individual in the target population will exhibit this value. Formally, for all $w \in W$ and for some $\delta > 0$

$$\delta < \Pr(D_i = 1 | W_i = w) < 1 - \delta$$

Minimally, this assumption implies that there are no macro-characteristics of the setting $D_i \epsilon \{0,1\}$ that causally interact with the treatment, that are different between populations, but homogenous within either population. For instance, consider again the stylized example concerning the effect of distributing free textbooks X on student performance Y as moderated by visual acuity W. If W only assumed one value $W = w$ for all individuals in $D_i = 0$, e.g. all students have 20/20 vision, then it is not feasible to estimate informative CATEs over this distribution in a way that would permit prediction of causal effects for any target population with a distribution other than $W = w$. In other words, it is not possible to predict how different levels of visual acuity will modify the ATE obtained in the study population, if variation in the outcomes that is attributable to differences in W has never been observed in the experimental population. In short, in cases where overlapping support of the interactive covariate distributions

[4] This largely mirrors a standard idea that has been put forward in the literature on probabilistic causation, where for each failure of a cause to increase the probability of its effect (here understood as the outcome Y), there must be a reason to explain this failure (see e.g. Cartwright 1979, 427), i.e. a reason that cites arrangements of causally relevant background circumstances as the culprit for suppressing the otherwise expected differences in probabilities of the effect.

fails, reweighting $CATE_{D=0}$ by the distribution of these variables in the target to obtain a prediction of $CATE_{D=1}$ is infeasible. I will discuss this assumption and some interesting consequences for experimental design in more detail in *Chapter 8*.

Jointly[5], the above assumptions permit identification of the causal effect in the target population as the expectation of the causal effect in the experimental population taken over the distribution of the interactive covariates in the target (see Hotz et al. 2005, 247 for the proof of this extrapolation theorem):

$$ATE_{D=1} = E[Y_i(1) - Y_i(0) | D_i = 1]$$

$$= E_W[E[Y_i|T = 1, D_i = 0, W_i] - E[Y_i|T = 0, D_i = 0, W_i] \,|D_i = 1]$$

This result allows performing quantitative extrapolation of causal effects, given suitably identified CATEs from an experimental population and observational data on the distributions of interactive covariates from both populations.

Interactive covariate-based extrapolation is hence similar in scope and capability to Bareinboim and Pearl's (2013) proposals (to be discussed in *Chapter 7*), which are intended to handle cases of extrapolation where experimental intervention in the target is not feasible, but observational data on moderating and mediating variables as well as confounders are available. In essence, the crucial commonality of these approaches is that both aim at expressing the expectation of a causal effect in the target in 'intervention-free terms' with respect to the target, i.e. as a function of the causal effect in the experimental population and the *observed* distribution of interactive covariates in the experimental and target population respectively. This is an important feature that is helpful for evading the extrapolator's bind. In short, while meeting the challenge posed by the extrapolator's bind does not preclude *all* experimental interventions in the target (nor all observations), any intervention conducted in the target to facilitate extrapolation may not be such that it allows learning the effect to be extrapolated. So at the very least, interventions on the eventual treatment variable in the target are precluded by the extrapolator's bind. Since the general setting that both Bareinboim and Pearl's approach and interactive covariate-based strategies assume is one in which no intervention is

[5] To be sure, the three assumptions outlined here, though central, are not the only assumptions required for covariate-based extrapolation to proceed successfully. Other, auxiliary assumptions are needed as well. For instance, it is necessary that the stable unit treatment value assumption (SUTVA, cf. Rubin 1980) holds, which means that units' potential outcomes are probabilistically independent of each other. However, while such assumptions are somewhat contentious, they are not the focus of the criticisms that I discuss below, which is why I bracket them here.

performed in the target, this intuitively seems to help evade the extrapolator's bind. With this in mind, let me turn to examine the epistemic requirements involved in interactive covariate-based extrapolation following the exposition by Muller (2014; 2015), before I turn to offer some criticisms that call into question whether interactive covariate-based extrapolation can truly evade the extrapolator's bind.

6.3 Epistemic Requirements

For interactive covariate-based extrapolation to be feasible, several epistemic requirements need to be met. Following the comprehensive exposition offered by Muller (2014), interactive covariate-based extrapolation imposes at least three important epistemic requirements.

The first is that all interactive covariates that are involved in producing an effect must be known[6], observable, and observed in both populations (Muller 2014, 41; 2015, 5). This requirement poses several distinct epistemic challenges for interactive covariate-based extrapolation.

For instance, as anticipated above, some variables that are detectable as (unconditional) covariates of the treatment effect will covary with the effect not because they are moderators or mediators of that effect (adjustment for which might be necessary for obtaining a correct expectation of the causal effect in the target) but, for instance, because they are correlated with such moderators or mediators, e.g. when they have a common causal parent with a moderator or are themselves a common effect of a moderator and the outcome. Conversely, not all moderating and mediating variables are discernible covariates of the treatment effect. For instance, two moderators whose interactive effects cancel each other out on average will not be detectable as covariates of the treatment effect unless a suitable conditioning set is chosen (e.g. conditioning on a specific value of one moderator to recognize that the other indeed covaries with the effect of interest).

In addition, not all interactive covariates are observable. This poses significant obstacles to interactive covariate-based extrapolation, particularly in settings where relevant moderators and mediators are latent characteristics of individuals (such as

[6] It is unclear what exactly this requirement amounts to, but it seems that proponents of interactive covariate-based strategies assume that it is sufficient to know that these variables are causally relevant in some way, although not necessarily in which way.

psychological features, e.g. agents' skill levels or features governing their susceptibility to behavioural biases) that are not easily captured by observable proxy variables without strong assumptions on their relation to proxies.

Finally, parameters that govern (or capture) how interactive covariates W bear on the effect magnitudes of interest, such as γ in the structural equation $Y = \gamma * W * X + v$, do not always lend themselves to unbiased estimation unless there is exogenous variation in the respective variables X whose influence they govern (see Imai et al. 2013). They are also difficult to compare across populations unless such exogenous variation can be exploited in both populations, which is an important point that I will discuss in more detail later.

The second important requirement for interactive covariate-based extrapolation to proceed successfully is that measurements of interactive covariates must be *comparable* across populations (Muller 2014, 41; 2015, 5). This requirement seeks to rule out several conceptually distinct types of cases where one and the same interactive covariate measurement may not be comparable between individuals for the purpose of reweighting causal effects. Muller (2014) focuses on cases where an interactive covariate is measured in different ways across populations or between individuals, such as when different individuals self-report variables in systematically different ways that are probabilistically dependent on location (see e.g. Fumagalli 2013 for an overview of such concerns in the context of measuring economic agents' well-being). This echoes standard concerns about consistent operationalization of measurement concepts and constructs (see e.g. Reiss 2008), in requiring that measurement protocols should be the same and that measurement concepts and constructs should have the same empirical referents across contexts.

Finally, the third important requirement for interactive covariate-based extrapolation is that the size of the experimental population must be large enough relative to the size of the covariate vector W to allow researchers to obtain not only unbiased but also precise estimates of the $CATE$ (Muller 2014, 41). If this condition fails, the standard errors on the estimates of interest for reweighting purposes may be too large to permit even qualitatively unambiguous prediction of causal effects in the target. While it is difficult to offer general rules on what combinations of sample size, dimensionality of the covariate vector W, and the variance of the covariates in W allow sufficiently precise parameter estimates, since this hinges on specific features of the estimation

problem of interest, it is possible to perform generic power analyses to investigate the implications of different arrangements of these factors for precision and statistical power.

With this overview of the principled aims, strategies, identification assumptions, and epistemic requirements involved in interactive covariate-based extrapolation in place, let me proceed to discuss a more basic, and I believe more important, concern about the assumptions that the approach involves.

6.4 Hopes, Assumptions, and the Extrapolator's Bind

A second, more fundamental set of concerns, takes issue with yet other, heretofore unspoken assumptions required for interactive covariate-based strategies to be successfully applied. To explain these concerns, let me invoke the distinction between different levels at which causally relevant differences between populations can obtain provided in *Chapter 2*: populations can differ in 1) the distributions of variables (including interactive covariates), 2) the functional form and parameters involved in the structural equations that best represent how causal effects are transmitted and outcomes are determined, and 3) in the basic structure of causal mechanisms, e.g. whether or not a variable W is involved in the mechanism that governs the production of Y at all. In a nutshell, my concern here will be that while it is an important achievement of interactive covariate-based approaches to help accommodate and adjust for differences at the first level, i.e. differences in interactive covariate distributions, accounting for these differences can only be successful if there are no relevant differences between populations at other levels (this is suggested, but not elaborated in Deaton and Cartwright 2016, 39; see also Horton 2000 and Rothwell 2005 for similar points about extrapolation in evidence-based medicine).

To illustrate with a familiar example, supplementary teaching S might be positively relevant for the effect of schooling X on student performance Y in an experimental population; it helps students review material discussed in class. However, it might be negatively relevant in a target where students are stigmatized by their peers for being in need of supplementary teaching, making them less confident in their abilities and

decreasing their performance on tests.[7] Clearly, adjusting for differences in the distribution of *S* between populations is only useful if we are confident that populations do not differ at any of the lower levels, specifically not with respect to *whether* and *how* *S* is involved in producing the effect of interest.[8] Moreover, merely *assuming* that populations are similar at lower levels would amount to extrapolation based on hope. Populations frequently differ in their structural makeup, e.g. because institutions, norms, individuals' psychological characteristics, and other features differ between them. So we need to support empirically (or otherwise) the claim that populations are sufficiently similar to warrant extrapolation. That is not only difficult, but even if feasible, can raise important concerns about the extrapolator's bind.

Let me start at the most basic level to illustrate. To support the assumption that experimental and target populations are similar with respect to the basic structure of causal mechanisms we need to learn something about the mechanisms in both populations. Say, for instance, the mechanism in the experimental population is understood to be $X \rightarrow Z \rightarrow Y$. Then, in order to ensure that the mechanism in the target is similar, we might need to learn whether all causal relationships comprising this mechanism are present there as well, i.e. we need to learn that $X \rightarrow Z$ and that $Z \rightarrow Y$. But learning this in a straightforward fashion by means of observations or interventions in the target can make the information obtained from the experimental population redundant to answering whether X is causally relevant for Y in the target.

Similar concerns apply when learning about similarities in *parameters* associated with variables that figure in the mechanisms as well as the *functional form association* of these variables. While we might, given suitable study designs, be able to learn how a variable W induces differences in a causal effect in an experimental population, this only gives us half the information we need. To see this, let us generously assume that we are fortunate enough to know the structural equations best representing how individuals' outcomes in the experimental population are determined. Let Y be the outcome, X the treatment variable, β the parameter that captures the baseline causal

[7] One might argue that rather than the parameter associated with *W* being positive in *A* but negative in *B*, what happens here is rather that *W* is associated with the outcome in *B* via an additional causal pathway that is mediated by stigmatization, which is negatively relevant for performance. I am open to alternative characterizations, as not much hinges on this. The point is merely that we need to ensure that there are no differences in whether (causal structure) and how (parameters/functional form) *W* is involved in producing *Y*. Otherwise, adjusting for differences in the value/distribution of *W* will not help us predict the correct causal quantity in *B*.

[8] Others seem to ignore this as well, e.g. Crump et al. (2008), Stuart et al. (2018).

effect of X, W an interactive covariate of the X-Y-effect, γ the parameter that captures the interactive effect of W and X on Y, and v an idiosyncratic error capturing the effects of other variables on Y. For ease of illustration, let the outcome equations for individuals in population A and B be of an additively separable form:[9]

$$Y_A \Leftarrow c + \beta * X_A + \gamma_A * W_A * X_A + v_A$$

$$Y_B \Leftarrow c + \beta * X_B + \gamma_B * W_B * X_B + v_B$$

It is easy to see that the marginal effect on Y induced by a given change in X depends on β, γ, and W, since $\Delta Y/\Delta X = \beta + \gamma * W$. Now, even if β is the same between populations, adjusting for differences in W can only be successful if $\gamma_B = \gamma_A$, i.e. if the *way in which* W induces differences in the effect is the same in both populations. So even if we are fortunate enough to have learned γ_A, we still need to learn γ_B in order to validate that $\gamma_B = \gamma_A$ (or is otherwise sufficiently similar). Likewise, concerning *functional form*, and granting that $\gamma_B = \gamma_A$, we need to ensure that the way in which W is functionally associated with the other variables figuring in the structural equations is the same in both populations.

Just like validating that mechanisms are similar between populations, establishing that they are similar in population-level parameters and functional form is generally difficult. For one, observational data for estimating γ_B might not be available. Even if they are, *unbiased* estimation of γ_B requires substantive assumptions, e.g. that there are no common causes of W and differences in treatment effects that could induce significant, but ultimately spurious interactions between X and W. To avoid such assumptions, γ_B can be identified by performing so-called *factorial experiments*, i.e. experiments where both X and W are exogenously varied (see e.g. Imai et al. 2013; Pearl 2014). However, doing so may not only be difficult – think of variables such as age that cannot be meaningfully intervened on – but also, factorial experiments in the target will often involve intervention on X and hence trivially fall prey to the extrapolator's bind as we can learn the causal effect of interest by doing so (unless, of

[9] The equations simply encode the causal assumptions elaborated here and represent that *Y* is causally determined in accordance with the equations. Indices are suppressed for simplicity, which means that we assume individuals within *A* and *B* respectively to be perfectly alike. Importantly, the particular functional form used here, where the term involving *W* is additively separable, is only assumed for ease of illustration. The problem highlighted persists in the more general case where $Y \Leftarrow f(\gamma, W, X, v)$. As long as $f(\cdot)$ involves some causal interaction between *X* and *W*, where the marginal effect of *W* is not separable from *X*, we need to ensure that the functional form of $f(\cdot)$ and the value of γ are the same between populations.

course, we look the other way – a possibility which I do not consider a genuine solution for evading the extrapolator's bind).

Proponents of interactive covariate-based extrapolation may object at this point that causally relevant differences at the levels of parameters, functional form, and basic structure of mechanisms pose no special problem to their approach and can be handled if there are observable proxy variables that correlate with these differences, as we can simply adjust for differences in the distributions of such proxy variables. For instance, even though the signs or values of important causal parameters may often be difficult to observe (or estimate) in practice, this does not mean that there could not be more readily observable features that covary with differences in parameters and functional form, and that offer themselves as targets for reweighting. To use a standard example from life sciences, the causal effects of medical interventions frequently differ by age or sex, but age or sex are often not doing the causal work of inducing differences in effects. Instead, finer-grained physiological details with which age and sex are correlated are responsible for inducing the observed differences. Yet, higher-level variables such as age and sex can still be used for interactive covariate-based adjustment if there is a reliable enough relation between those underlying details that do the causal work and the observed variables by which the adjustment proceeds.

Likewise, there could be close concomitants of differences in the basic causal structure of mechanisms. For instance, the presence or absence of a general social norm in a population, parts of which may be involved in the transmission and/or modification of causal effects, may be relatively straightforward to observe. This is despite the fact that the concrete details of what causal features related to (and co-instantiated with) the presence or absence of this norm are involved in transmitting or modifying the effects of interest may remain entirely unknown to us. Hence, in different ways, we might still be able to use interactive covariate-based extrapolation to adjust for causally relevant differences at the levels of parameters, functional form, and basic causal structure, if there are observable variables that reliably correlate with these more basic and less accessible causal features.

There are three reasons to be sceptical about this possibility, however. First, it is generally unclear whether differences in parameters, functional form, and the basic structure of causal mechanisms frequently have readily observable correlates that are amenable to this solution. Put simply, agents and populations may sometimes, but do

not always, wear mechanism-types, functional forms, or parameter values on their sleeves (cf. Weinberger 2014 who uses the same expression to flag this concern; see Little 1993 for related concerns; see Strevens 2007 for a more optimistic view).

Second, features of causal mechanisms (and differences in such features) that are important for extrapolation often do not readily manifest themselves in ways other than in agents' behavioural response to the specific intervention of interest – think, for instance, about agents' latent psychological characteristics such as those relevant in the HIV-prevention example from the previous chapter. Here, particularly in virtue of the fact that agents have not yet experienced an intervention of interest, including any associated changes in mediating variables, it is unlikely that distinctive higher-level features of agents or whole populations will be reliably indicative of what mechanism-type will be at work in the population. So while there could be proxy variables that correlate with important parameters and features of mechanisms, measurements of such variables would often only be useful after the intervention of interest was already experienced by agents, potentially raising concerns about the extrapolator's bind.

Third, even if many causally relevant differences in parameters, functional form, and the basic structure of mechanisms had readily observable proxies to permit interactive covariate-based reweighting, measuring and accommodating differences in such variables would still require extensive causal knowledge about both populations, including details concerning how the proxy variables of interest are associated with the underlying differences in parameters, functional form, and basic causal structure and whether this association is the same in both populations. The latter is especially likely to be difficult to support in practice. This suggests that the concerns developed here are not easily remedied by appealing to the principled possibility of adjusting for differences in proxy variables, at least not without making substantive assumptions and raising yet further concerns about the extrapolator's bind.

The upshot is this: interactive covariate-based extrapolation presently proceeds on the assumptions that experimental and target populations are relevantly similar at the level of the basic structure of causal mechanisms as well as the parameters and functional form associations of the interactive covariates by which adjustment proceeds. Supporting these assumptions, however, will not only often be difficult, but can also raise important concerns about the extrapolator's bind.

This suggests the need for a supplementary strategy to underwrite interactive covariate-based extrapolation, one that steers clear of the extrapolator's bind. One candidate for this could be Steel's CPT. Let me consider whether supplementing interactive covariate-based strategies with CPT could help evade some of the problems sketched out above.

6.5 Can Steel's CPT Save Interactive Covariate-Based Extrapolation?

At least under some conditions, Steel's CPT can help us extrapolate causal relevance claims, e.g. whether X is causally relevant for Y in B if it is causally relevant in A, while evading the extrapolator's bind. In virtue of this, it could be useful for underwriting interactive covariate-based extrapolation as it seems to offer a way to evade the extrapolator's bind, at least when it comes to supporting that experimental and target populations are similar at the level of the basic structure of causal mechanisms. In what follows, however, I offer reasons to think that CPT will not be useful for this purpose in the kinds of cases that econometricians and EBP researchers typically encounter, especially given their preference for using quantitative observational evidence from the target to facilitate extrapolation. To help explain why, I will draw again on the distinction between *attributive* and *predictive* extrapolation developed in the previous chapter.

As I have argued there, the kind of extrapolation Steel discusses in developing his strategy is special. This *attributive extrapolation* aims to attribute an observed effect causally to its suspected causes. The kind of extrapolation typically encountered in EBP and econometrics is, however, importantly different. This *predictive extrapolation* aims to predict the future effects of (interventions on) suspected causes and proceeds under conditions where neither the intervention of interest nor its suspected effects have yet been observed or experienced in the target.

In what follows, I argue that despite the promises of CPT, problems of predictive extrapolation are unlikely to be overcome by supporting interactive covariate-based strategies using CPT without falling prey to the extrapolator's bind, at least not without substantial changes to what evidence is considered relevant in supporting extrapolation.

As I have argued, the important difference between attributive and predictive extrapolation concerns the evidence that is available to support extrapolation. In

discussing the attributive extrapolation of the effects of AFB1 from animals to humans, Steel's aim is to highlight the importance of process tracing evidence, i.e. observations of the presence or absence of distinctive marks of a hypothesized mechanism. Yet, Steel's case also proceeded against the background of several other kinds of supplementary evidence that established a basis for comparing mechanisms, including evidence that 1) helped to characterize the mechanism of interest in animals, 2) indicated between-species variation in the effect of interest that could present obstacles to extrapolation from animals to humans, and 3) clarified between-species differences in mechanisms that induced these between-species differences in effects.

As I have argued, and as Steel recognizes, an immediate problem in many social science contexts is that such evidence is difficult to produce (2008, Ch. 8). Even if such evidence could be readily obtained from experimental populations, to *compare* mechanisms we still need some evidence characterizing the mechanism in the target. This is crucial. In contrast to the AFB1 example where experiments on components of the putative mechanism in the target could be performed (e.g. on human cell cultures), in many contexts of interest in EBP and econometrics similar means for observing mechanisms in the target are unavailable (Steel 2008, 166). Instead, the most salient way in which EBP researchers and econometricians could compare mechanisms in line with standard methodological tenets and evidential preferences in these fields is by using quantitative observational data from the target. Such data could help determine whether distinctive marks of the suspected mechanism, e.g. in the form of distinctive covariance and probabilistic dependence/independence signatures between variables, are realized there.[10]

In Steel's example, such evidence is supplied by observational studies on humans. These studies offered covariance information suggesting that relative risk of HCC is significantly higher in humans that have been exposed to AFB1 and that this association remains stable even when conditioning on HBV infection (IARC 1993).

Such evidence is not available in predictive extrapolation cases. The crucial difference here is that when neither the intervention of interest nor its suspected effects

[10] This seems coherent with how extrapolation is supposed to proceed according to Hotz et al. (2005). On their approach, the evidence from the target used to facilitate extrapolation is quantitative observational evidence concerning the distributions of interactive covariates. Exemplary applications such as Dehejia et al. (2015) and Gechter (2016) suggest that econometricians would also be inclined to rely on quantitative observational data from the *target* to underwrite extrapolation.

have yet been observed and experienced in the target, quantitative observational evidence cannot speak to questions about mechanistic similarity and difference.

An example helps illustrate this. Suppose we learn that distributing free insecticide-treated bed nets helps decrease malaria infection rates in population A. Suppose further that bed nets must be properly installed to curtail malaria infection, and that whether nets are properly installed can differ significantly between populations, e.g. agents in some populations might use them as fishing nets instead (McLean et al. 2014). Let me represent this by the simplistic mechanism $X \rightarrow Z \rightarrow Y$, where Z, the number of properly installed nets, is a mediating variable on the path from distributed nets X to malaria infection Y (where Z is negatively relevant for Y).

How can we make sure that this mechanism is sufficiently similar between an experimental population A and a novel target B where bed nets have not yet been distributed? Quantitative observational evidence that could help indicate that the mechanism in B is similar to that in A would be that Z is higher conditional on X than unconditionally, indicating that distributed nets are properly installed, and that Y is lower conditional on Z than unconditionally, suggesting that properly installed nets in fact reduce malaria infection.[11]

The crucial problem is that such information cannot be usefully obtained from the target if no bed nets have ever been distributed there. If that is the case then X and Z exhibit no variation, since $X = 0$ and $Z = 0$ for all individuals, and Y will only assume its natural value that is induced through relevant malaria infection pathways.[12] This means that Z conditional on X and Z unconditionally will be equal, and that Y conditional on Z, and Y unconditionally will be equal as well. So if no bed nets have ever been distributed in the target, there will be no (co-)variation in the outcomes of interest or the intermediate stages of the mechanism that could help us tell whether mechanisms are sufficiently similar.

More generally, in cases where the intervention of interest has not yet been experienced and observed in the target, quantitative observational data on variables that figure in the suspected mechanism are not informative about mechanistic similarities

[11] This is called *pattern evidence* on Beach and Pedersen's (2016) typology of process tracing evidence.
[12] In the macroeconometrics literature, this is known as the problem of *non-excitation* (cf. Salmon and Wallis 1982; Engle et al. 1983).

and differences between populations.[13] This means that if, as econometricians and EBP researchers do, we primarily consider quantitative observational data from the target as relevant for underwriting extrapolation, we cannot tell whether the target exhibits characteristic signatures of the suspected mechanism being operational. As Steel anticipates, "[…] the operation of a program can be examined only where it is implemented […]" (2008, 166). So while this problem can be remedied by intervening on *X* in the target, doing so would trivially fall prey to the extrapolator's bind.

To be sure, as anticipated in *Chapter 5*, one could argue that quantitative observational evidence from the target can still have *indirect* bearing on questions of mechanistic similarity and difference if there have been similar, and well-understood interventions (or exogenous changes) in the target in the past.[14] This seems possible, but would also seem to require substantive assumptions concerning how such past interventions (or exogenous changes) relate to those of current interest, e.g. whether their effects are governed by the same mechanisms and in the same way, as well as whether the intervention of current interest is structure-altering or not. Such assumptions are similarly difficult to support as, and not recognizably weaker than, those at issue here.

The concerns outlined above are not surprising. Quantitative observational data only have bearing on questions concerning features of causal mechanisms if there is sufficient variation in at least some of the putatively causally relevant variables. More generally, we might say that quantitative observational evidence can only be informative about the causal mechanisms governing observable phenomena if these mechanisms have been 'active', and consequently had the opportunity to 'write', as it were, distinctive marks, symptoms, signatures etc. into the data that we can obtain. Without such opportunity, when mechanisms have remained 'dormant' so far, quantitative observational evidence from the target remains a poor guide to clarifying issues of mechanistic similarity and difference.

---

[13] Interestingly, econometricians recognize important differences between cases where the effects of interest have not yet been experienced in a population and those where they have been (e.g. Heckman 2005; Imbens and Wooldridge 2009; Athey and Imbens 2017; Braithwaite and Walker 2018). However, they do not seem to recognize the importance of this distinction for extrapolation.

[14] Steel anticipates this general intuition when discussing concerns surrounding structure-altering interventions (2008, 157-60).

6.6 Predictive Extrapolation: Where Next?

The previous discussion suggests three things. First, interactive covariate-based strategies for extrapolation involve wide-ranging but ultimately unsubstantiated assumptions that populations are sufficiently similar at the level of parameters, functional form, and basic structural-mechanistic features relevant to the effects of interest. Second, supporting these assumptions raises concerns about the extrapolator's bind. Third, even Steel's CPT is not immune to these concerns: at least in predictive extrapolation cases, quantitative observational data from the target are of little help in clarifying issues of mechanistic similarity and difference. So even if econometricians and EBP researchers were to use CPT to underwrite extrapolation, they could not rely on quantitative observational data from the target for this purpose.

This is not to suggest that interactive covariate-based strategies are fundamentally flawed. It is practically difficult, but perhaps not insurmountably so, to support that populations are relevantly similar at the level of parameters, functional form, and mechanisms. Similarly, as emphasized in *Chapter 5*, my aim is not to suggest that CPT is an inadequate strategy for extrapolation in general. I consider CPT to be a promising strategy; but my concern is that predictive extrapolation poses distinct challenges for CPT, specifically that quantitative observational data from the target are not useful for CPT in such cases. So if preferences for such evidence are maintained, it seems that interactive covariate-based extrapolation cannot be underwritten by CPT, and if applied at all, would need to proceed on hope that populations are sufficiently similar rather than evidence that this is so.

As this seems highly unsatisfactory, I now want to consider some suggestions for what might be done about this. Since the primary aims of this chapter are critical in nature, I will not attempt to develop an alternative strategy for underwriting extrapolation here. However, it seems useful to consider at least some tentative ways to respond to the challenges put forward, specifically proposals offered in the EBP-related literature that is not married to preferences for quantitative data.

There is a rich literature on supporting process tracing, including for purposes of extrapolation, by *qualitative* evidence, e.g. sociological, anthropological, and ethnographic evidence obtained from sources such as interview studies, participatory observation, and expert judgment (see e.g. Blatter and Blume 2008; Kay and Baker 2015; Schmitt and Beach 2015; Fairfield and Charman 2017 for a Bayesian approach to

integrating different kinds of qualitative evidence). Econometricians and EBP researchers have so far been reluctant to consider such evidence (see Kern et al. 2016). However, the arguments developed here suggest that this reluctance is misguided and that producing, using, and integrating other kinds of evidence, including qualitative evidence, may be useful, and in some cases perhaps even necessary, for underwriting extrapolation beyond the level of mere hope that crucial assumptions are satisfied.

Let me draw on the bed net example again to illustrate how this could proceed. For instance, analogously to in-vitro studies on human cell cultures that helped pin down specific features of parts of the suspected mechanism in humans in Steel's AFB1 example, it might be possible to investigate whether the insecticidal effects of the bed nets to be distributed in the target do in fact obtain on mosquitos sampled from the target, at least qualitatively. So for some parts of the suspected mechanisms in the target, it may be relatively straightforward to isolate and study parts of it (or well-understood analogues of it), without 1) having to rely on quantitative observational data from the target (such as observing that infection rates are lower conditionally on properly installed nets than unconditionally) or 2) introducing bed nets in the target and triggering the extrapolator's bind.[15]

For other causal relationships, this may be more challenging. For instance, it seems that investigating whether agents in the target will properly install bed nets can again raise concerns about the extrapolator's bind, e.g. when we introduce bed nets provisionally in (at least part of) the target. Proponents of qualitative approaches to EBP might suggest that this could be avoided by considering sources of evidence such as participatory observation, agents' self-reports, or expert judgment. For instance, they might point out that agents can sometimes reliably report on counterfactual states of affairs that have bearing on questions of mechanistic similarity and difference, e.g. when these counterfactual states are importantly determined by agents' own decision-making (see e.g. Kay and Baker 2015; Fairfield and Charman 2017 for such suggestions). If, for instance, agents do in fact have alternative uses for bed nets, such as using them as fishing nets, then it seems that they could, under some conditions, reliably report that, if they were given free bed nets, they would not use them as bed nets but rather as fishing nets. Similarly, agents might also be able to report on counterfactual states of affairs that hinge on other agents' decision making, or on

[15] I thank an anonymous referee of the *Journal of Economic Methodology* for suggesting this example.

existing social norms, where agents, as well as experts with local knowledge of the target, might be able to anticipate, at least qualitatively, how these norms may interact with the intervention of interest (see Cartwright and Hardie 2012 for related suggestions).[16]

Of course, considering qualitative evidence also raises a host of new challenges. Agents might be overconfident about their propensity to adhere to implementation protocols; they may experience substantial difficulties in anticipating the effects of some interventions, such as predicting the effect of deworming on student achievement; they may be incentivized to strategically misreport their expected future behaviours; and so on.

More generally, agents' self-reports are often plausibly suspected to be unreliable, and various important precautions need to be undertaken to support the reliability of such evidence, e.g. triangulating qualitative conclusions by considering multiple sources and using different elicitation methods; having agents report on others' behaviours instead of their own; ensuring that agents are not improperly incentivized to strategically misreport; etc (see Schmitt and Beach 2015 for practical examples concerning the importance of such precautions).

What is more, it is not to be expected that qualitative evidence of the kind outlined above will be sufficiently informative by itself to tell us how much exactly, for instance, a particular effect will be amplified or diminished by local causal features of the target. However, it is important to recognize that this does not preclude quantitative predictions of causal effects in the target. If qualitative evidence increases our confidence that crucial features of causal mechanisms are qualitatively similar between populations, e.g. that a moderating variable W is likely to interact with an intervention in the same qualitative way in both populations, this can offer important support (although perhaps not full-fledged warrant) for the assumptions that are necessary for interactive covariate-based extrapolation to proceed. If this is successful, interactive covariate-

[16] This is not dramatically different from what Steel's CPT recommends. The arguments presented here add an important nuance, however, which is that in predictive extrapolation, for lack of observational evidence indicating that there is *some* causal pathway from the intervention variable to the outcome of interest, CPT may need to be supported with *more* evidence that has bearing on questions of mechanistic similarity. When available evidence from the target is insufficient to clarify these questions, this may, again, make it likely that we fall prey to the extrapolator's bind. At least in these cases, the attributive/predictive extrapolation distinction has an important ramification for CPT: it can only evade the extrapolator's bind in a constrained class of cases.

based strategies may justifiably be used to obtain *quantitative* predictions of causal effects, just as envisioned by their proponents.

So qualitative evidence is not a silver bullet to address the shortcomings of interactive covariate-based strategies. However, when quantitative data from the target are unlikely to help clarify whether populations are sufficiently similar to licence extrapolation at all, considering qualitative evidence with bearing on these issues, despite its potential shortcomings and the additional methodological burden placed on us, may still be recognizably superior to proceeding on mere hope that populations are sufficiently similar.

Most importantly, considering (certain forms of) qualitative evidence promises to help us evade the extrapolator's bind. Here, the intervention is not introduced in the target *in fact*, but only *hypothetically*, in the minds of agents who may possess relevant expertise to report on features of causal mechanisms and processes that they are part of, and that have important bearing on the effects of interest. This would steer clear of the extrapolator's bind because while agents' self-reports may help us rule out important causally relevant differences between populations, the quantitative causal effects of interest could probably not be reliably inferred by asking them any number of questions. So qualitative evidence can be useful for clarifying issues of similarity and difference between populations, but is not a sufficient means to predict quantitative causal effects in the target.

So what is the main suggestion for how interactive covariate-based extrapolation in EBP should proceed in light of the arguments provided here? I will expand in more detail on positive proposals in *Chapter 8*. For now, it seems clear that predictive extrapolation (which is predominant in EBP) poses distinct challenges for interactive covariate-based strategies (as well as for CPT). While there is no obvious remedy, it seems that considering qualitative evidence with bearing on questions of similarity and difference between populations is an option that should be explored in more detail, as such evidence might be able to give us at least some purchase on whether the assumptions required for interactive covariate-based extrapolation are satisfied.

In light of this, it seems reasonable to suggest that the ability of qualitative evidence to speak to these issues should be investigated further, and that the production and use of such evidence should be encouraged in widely circulated methodological guidelines such as those issued by the Campbell Collaboration, What Works Clearinghouse,

CONSORT, GRADE, J-PAL, and others. In addition, more attempts should be made to develop strategies for *integrating* quantitative and qualitative evidence, including in domain-general theories (sometimes called programme theories, theories of change, or logic frames) that aim to offer comprehensive accounts of *how* the interventions of interest achieve their intended effects, and under what conditions they might fail to do so (see e.g. Davey at al. 2017 for similar suggestions).[17] EBP institutions such as J-PAL, 3ie, and others already make attempts along these lines (White 2009). These are only early steps, however, and a persuasive, general methodology for underwriting extrapolation by means of integrating qualitative and quantitative evidence is still missing.

As the arguments provided here suggest, there is much promise in developing such a methodology. It seems that there can be cases where qualitative and quantitative evidence, when considered in tandem, can help us extrapolate causal effects in a way that is superior to doing so based on either type of evidence alone. Qualitative evidence by itself can at best clarify issues of qualitative causal relevance. Quantitative evidence by itself, on the other hand, can help us make extrapolative predictions of causal effect magnitudes in the target, but these predictions are only credible if crucial assumptions about similarities between populations are adequately supported. At least in predictive extrapolation, it is clear that this role cannot be played by further quantitative observational evidence from the target. But as the arguments provided here suggest, qualitative and quantitative evidence can play complementary roles: one helps clarify whether populations are similar at the level of basic causal structure (and potentially structural parameters), the other helps investigate causal effect magnitudes of interventions implemented in one setting and with adjusting for differences in the distributions of variables that can modify these effects. Considered together, both types of evidence can hence underwrite extrapolative conclusions that would not be accessible from either type of evidence alone.

Providing the details of a method for integrating qualitative and quantitative evidence is beyond the scope of this thesis, but I hope that the arguments presented here will reinforce similar suggestions made by other philosophers (e.g. Cartwright 2013; Cartwright and Hardie 2012; Grüne-Yanoff 2016) by providing reasons to think that extrapolation may not only be greatly facilitated by considering qualitative evidence

[17] see Clarke et al. 2013; 2014 for similar suggestions concerning evidence-based medicine.

pertaining to the causal mechanisms governing the effects to be extrapolated, but that a wide range of real-world extrapolations may be exceedingly difficult to underwrite without doing so. This, I hope, will help motivate further contributions that encourage econometricians and EBP researchers to add previously neglected kinds of evidence to their arsenals in the pursuit of underwriting extrapolation by more than hope alone.

References

Angrist, J., and I. Fernandez-Val. (2013). "ExtrapoLATE - ing: External Validity and Overidentification in the LATE Framework", in D. Acemoglu, M. Arellano, and E. Dekel, eds., *Advances in Economics and Econometrics*, Cambridge: Cambridge University Press.

Athey, S., and G. W. Imbens. (2016). "Recursive Partitioning for Heterogeneous Causal Effects". *PNAS* . 113(27): 7353-60.

— **(2017).** "The state of applied econometrics: causality and policy evaluation". *Journal of Economic Perspectives* 31(2): 3–32.

Bareinboim, E. and J. Pearl. (2013). "A general algorithm for deciding transportability of experimental results", *Journal of Causal Inference*, 1: 107-134.

Beach, D., and R. B. Pedersen. (2016). *Causal Case Studies: Foundations and Guidelines for Comparing, Matching and Tracing*. Ann Arbor: University of Michigan Press.

Blatter, J., and T. Blume. (2008). "In Search of Co-variance, Causal Mechanisms or Congruence? Towards a Plural Understanding of Case Studies". *Swiss Political Science Review*. 14(2): 315-56.

Boumans, M. J. (2005). How Economists Model the World into Numbers. London and New York: Routledge.

Braithwaite, T. and J. L. Walker. (2018), "Causal inference in travel demand modeling (and the lack thereof)", *Journal of Choice Modelling*, 26: 1-18.

Cartwright, N. D. (1979). "Causal laws and effective strategies", *Nous*, 13: 419–37.

— **(2013).** "Evidence, Argument and Prediction". In: V. Karakostas, and D. Dieks (eds.), EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings. Cham, Switzerland: Springer International Publishing Switzerland.

Cartwright, N. D., and J. Hardie. (2012). *Evidence-Based Policy: A Practical Guide to Doing it Better*, Oxford: Oxford University Press.

Clarke, B., D. Gillies, P. Illari, F. Russo, and J. Williamson. (2013). "The Evidence that Evidence-Based Medicine Omits", *Preventive Medicine*, 57: 745–47.

— **(2014).** "Mechanisms and the Evidence Hierarchy", *Topoi*, 33: 339–60.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik, (2008). "Nonparametric tests for treatment effect heterogeneity". *The Review of Economics and Statistics*, 90(3): 389–405.

Davey, C., S. Hassan, C. Bonell, N. Cartwright, M. Humphreys, A. Prost, and J. Hargreaves. (2017). "Gaps in Evaluation Methods for Addressing Challenging Contexts in Development". CEDIL PreInception Paper: London

Deaton, A., and N. D. Cartwright. (2016). "Understanding and misunderstanding randomized controlled trials". Technical report. National Bureau of Economic Research.

Dehejia, R., C. Pop-Eleches, and C. Samii. (2015). "Local to Global: External Validity in a Fertility Natural Experiment". SSRN Electronic Journal. 10.2139/ssrn.2647649.

Engle, R. F., D. F. Hendry, and J.-F. Richard. (1983). "Exogeneity", *Econometrica*, 51(2): 277-304.

Fairfield, T., and A. Charman. (2017). "Explicit Bayesian analysis for process tracing: guidelines, opportunities, and caveats", *Political Analysis*, 25(3): 363-80.

Fumagalli, R. (2013). "The Futile Seach for True Utility", *Economics and Philosophy*, 29 (3): 325-47.

Gechter, M. (2016). "Generalizing the Results from Social Experiments: Theory and Evidence from Mexico and India". Unpublished manuscript, Pennsylvania State University.

Grüne-Yanoff, T. (2016). "Why Behavioural Policy needs Mechanistic Evidence", *Economics and Philosophy*, 32(3), 463-83.

Heckman, J. (2005), "The scientific model of causality", *Sociological methodology*, 35: 1–97.

Horton, R. (2000). "Common sense and figures: the rhetoric of validity in medicine: Bradford Hill memorial lecture 1999", *Statistics in medicine*, 19: 3149–64.

Hotz, J. v., G. W. Imbens, and J. H. Mortimer. (2005). "Predicting the efficacy of future training programs using past experiences at other locations", *Journal of Econometrics*, 125: 241–270.

International Agency for Research on Cancer (IARC). (1993). Some naturally occurring substances: Food items and constituents, heterocyclic aromatic amines and mycotoxins. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Vol. 56. Lyon.

Imbens, G.W., and J.M. Wooldridge. (2009). "Recent developments in the econometrics of program evaluation", *Journal of Economic Literature*, 47: 5–86.

Imai, K., D. Tingley, and T. Yamamoto. (2013). "Experimental designs for identifying causal mechanisms", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176: 5–51.

Kay, A., and P. Baker. (2015). "What Can Causal Process Tracing Offer to Policy Studies? A Review of the Literature", *The Policy Studies Journal*, 43(1): 1-20.

Keane, M. P. (2010). "Structural vs. atheoretic approaches to econometrics", *Journal of Econometrics*, 156: 3-20.

Kern, H. L., E. A. Stuart, J. Hill, and D. P. Green. (2016). "Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations", *Journal of Research on Educational Effectiveness*, 9(1): 103-127.

Khosrowi, D. (2019). "Extrapolation of Causal Effects – Hopes, Assumptions, and the Extrapolator's Circle", *Journal of Economic Methodology*, 26(1): 45-58.

Little, D. (1993). "On the Scope and Limits of Generalizations in the Social Sciences", *Synthese*, 97(2): 183-207.

McLean, K. A., A. Byanaku, A. Kubikonse,V. Tshowe, S. Katensi and A. G Lehman. 2014. "Fishing with bed nets on Lake Tanganyika: a randomized survey" Malaria Journal 13:395.

Muller, S. M. (2013). "External validity, causal interaction and randomised trials: the case of economics", Unpublished manuscript.

— **(2014).** "Randomised trials for policy: a review of the external validity of treatment effects". Southern Africa Labour and Development Research Unit Working Paper 127, University of Cape Town.

— **(2015).** "Interaction and external validity: obstacles to the policy relevance of randomized evaluations", *World Bank Economic Review*, 29(1): 217-25.

Neyman, J., Iwaszkiewicz, K., & S. T. Kolodziejczyk. (1935). "Statistical Problems in Agricultural Experimentation". *Supplement to the Journal of the Royal Statistical Society,* 2(2), 107-180.

Pearl, J. (2014). "Reply to Commentary by Imai, Keele, Tingley and Yamamoto Concerning Causal Mediation Analysis." Psychological Methods. 19(4): 488-92.

Reiss, J. (2008). *Error in Economics: The Methodology of Evidence-Based Economics*. London: Routledge.

Rothwell, P. M. (2005). "External validity of randomized controlled trials: 'to whom do the results of the trial apply'", *Lancet*, 365: 82–93.

Rubin, D. B. (1974). "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies", *Journal of Educational Psychology,* 66(5): 688–701.

— **(1980).** “Comment on: ‘Randomization analysis of experimental data in the fisher randomization test’ by D. Basu”, *Journal of the American Statistical Association*, 75: 591–93.

Salmon, M. and K. F. Wallis. (1982). "Model Validation and Forecast Comparisons: Theoretical and Practical Considerations", Ch. 12 in G. C. Chow and P. Corsi (eds.) *Evaluating the Reliability of Macroeconomic Models*. New York: John Wiley.

Schmitt, J., and D. Beach. (2015). “The contribution of process tracing to theory-based evaluations of complex aid instruments”, *Evaluation*. 21(4): 429–47.

Steel, D. (2008). *Across the boundaries: Extrapolation in biology and social science*. Oxford University Press.

Strevens, M. (2007). “Why Represent Causal Relations?” In A. Gopnik and L. Schulz (eds.), *Causal Learning: Psychology, Philosophy, Computation*, New York: Oxford University Press.

Stuart, E. A., B. Ackerman, and D. Westreich. (2018). "Generalizability of Randomized Trial Results to Target Populations: Design and Analysis Possibilities", *Research on Social Work Practice*, 28(5): 532–37.

van Eersel, G. G., G. V. Koppenol-Gonzalez, and J. Reiss. (2019). "Extrapolation of Experimental Results through Analogical Reasoning from Latent Classes". *Philosophy of Science*, 86(2): 219-35.

Weinberger, N. (2014). “Review: Evidence-Based Policy: A Practical Guide to Doing it Better”, *Economics and Philosophy*, 30: 113-20

White, H. (2009). “Theory–Based Impact Evaluation: Principles and Practice”, *The Journal of Development Effectiveness*. 1(3): 271-84.

CHAPTER 7

Graph-Based Extrapolation

7.1 Introduction

In this chapter, I critically discuss approaches to extrapolation developed in the computer science literature by Elias Bareinboim and Judea Pearl (Pearl and Bareinboim 2011; 2014; Bareinboim and Pearl 2012; 2013; 2014; 2016; Pearl 2014; 2015). These contributions build on and unify methods developed in causal graph theory and structural causal modelling that have been used for a variety of purposes, including most prominently the identification of causal relationships and causal effects from observational data (e.g. Spirtes et al. 2000; Tian and Pearl 2003; Shpitser and Pearl 2006; Huang and Valtorta 2006). My focus here is on uses of these methods for purposes of extrapolation. This *graph-based approach*, as I will call it from now, involves graphical causal models, called *directed acyclic graphs* (DAGs), that encode qualitative causal assumptions and are accompanied by corresponding structural causal models (SCMs), as well as a formal calculus by which expressions that identify causal quantities in a target population can be derived.

The approach offered by Bareinboim and Pearl (henceforth B&P) has several distinct advantages over other extrapolation strategies. First, in addition to permitting extrapolative inference from experimental data, it also does so from observational data, provided that these data permit unbiased identification of the effect to be extrapolated.

Second, the graph-based approach can address a wide range of causal queries, including quantitative causal queries about outcome distributions and effect magnitudes in cases where experimental and target populations differ in known causally relevant respects. In virtue of this, its capabilities accommodate and extend (potentially significantly) beyond those of other approaches discussed in previous chapters.

Third, the graph-based approach offers both an expressive graphical and formal framework to represent problems of extrapolation as well as a powerful formal machinery to help overcome them. The former helps encode causal background knowledge and makes such knowledge, as well as the absence of such knowledge, explicit. This promotes the transparency and tractability of extrapolative inference, as,

unlike on other approaches where important causal assumptions often remain implicit, the graph-based approach permits 'reading off' substantive causal assumptions from the graphical causal models themselves. Moreover, the formal calculus and algorithms developed by B&P help decide whether causal effects can be extrapolated at all, and if so, yield so-called *transport formulae*, i.e. expressions by which the target causal quantity of interest can, in principle, be estimated. Notably, for a broad class of problems, the calculus offered by B&P has been proven to be complete (B&P 2012), meaning that whenever a causal query can be answered, and the required background conditions are satisfied, then answers to causal queries follow deductively from background knowledge, assumptions, observations, and the rules of the calculus underlying the approach. This deductive rigour is taken to be desirable as many commentators note that there is often too little rigour involved when issues of extrapolation are discussed and addressed (see e.g. Cartwright 2013, 16; Pearl 2014).

Fourth, the graph-based approach can proceed from experimental or quasi-experimental observational data from a study population and observational data from a target. In virtue of not requiring experimental data from the target it hence seems to be a promising candidate for evading the extrapolator's bind. What is more, several transport formulae derived by B&P suggest that their approach can substantially reduce measurement cost in both populations, e.g. by requiring only few quantities to be measured in the target, thus intuitively decreasing the likelihood of falling prey to the extrapolator's bind. This suggests that the graph-based approach may enable genuinely *ampliative* conclusions that extend potentially significantly beyond the supplementary knowledge and evidence required to infer them.

The above achievements lead B&P to conclude that the problem of extrapolation has been solved by their approach (B&P 2016, 7352). In this chapter, I provide reasons to resist this conclusion. I do so by arguing that the graph-based approach encounters important problems related to representing causally relevant differences of various kinds as well as empirically supporting unspoken causal assumptions required for its successful application. These problems have, so far, remained unmentioned by critics of the graph-based approach and unaddressed by its proponents, but are significant enough to make it likely that the approach will fail to achieve successful extrapolation in a wide range of real-world extrapolation scenarios.

The chapter is organized as follows: *Section 2* provides an in-depth overview of the graph-based approach as a basis for my subsequent criticisms. In *Section 3* I take issue with important limitations of causal graphs in representing causally relevant differences between populations. While some of these limitations might be overcome, doing so comes at the expense of rendering many extrapolation problems insurmountably difficult to handle. In *Section 4* I discuss what causal assumptions are required to licence graph-based extrapolation in several pertinent examples used by B&P and how supporting these assumptions raises important epistemic concerns. First, for several related reasons, they can render the approach epistemically over-demanding (see also Steel 2013, 197). Second, much like the strategies discussed in previous chapters, these epistemic demands also raise important concerns about the extrapolator's bind. Building on my criticisms developed in previous chapters, I discuss how the distinction between attributive and predictive extrapolation further aggravates these concerns. I conclude that, without any explicit discussion of these challenges, the graph-based approach obscures, rather than clarifies, important epistemic obstacles to successful extrapolation.

Let me begin with an overview of the basic ingredients of the approach, which will help articulate the criticisms to be developed later.

7.2 Causal Graphs: The Basics

The graph-based approach has four main elements: 1) *Directed acyclic graphs* (DAGs) graphically encode the structural causal models and qualitative causal assumptions pertaining to the populations involved in an extrapolation. 2) *Do-calculus* is the formal calculus by which derivation of expressions used to estimate target causal quantities proceeds. 3) *D-separation* is a graphical criterion that helps 'read off' probabilistic independence features from graphs and identify causal effects by 'shielding' them from causally relevant differences. 4) *Selection diagrams* are an extension of DAGs that help encode knowledge and assumptions about causally relevant similarities and differences between populations and, together with the rules of do-calculus, form the basis for deriving expressions that help compute the target causal quantities to be inferred by an extrapolation. Let me provide an overview of these elements, following the expositions offered by Scheines (1997), Greenland and Brumback (2002), Pearl (2009), Steel (2010), and Elwert (2013).

7.2.1 Directed Acyclic Graphs and Structural Causal Models

At the heart of the graph-based approach are two ingredients: causal graphs and accompanying structural causal models. Causal graphs encode qualitative causal knowledge and assumptions pertaining to (the structure of) causal relationships that obtain among a set of variables. Structural causal models capture and concretise these relationships formally: they capture the causal relationships stipulated by a causal graph by specifying which variables figure as arguments in determining the values of other variables, and, in parametric cases, provide additional details about the functional form of the causal relationships between variables, properties of distributions, and parameter values. These latter, *parametric* details extend beyond the information encoded by the graph itself and are not required for the construction of a causal graph.

The causal graphs used in the graph-based approach are called *directed acyclic graphs* (DAGs). DAGs consist of a set of *nodes* that represent variables, and a set of *edges* that represent relationships between the variables. An edge between X and Y is *directed* when it points from one variable to another with an arrowhead, such as in $X \rightarrow Y$. The direction of the arrow indicates the direction of the causal relationship that is assumed (or suspected) to hold between the two variables. If there is an arrow between two variables, such as in $X \rightarrow Y$, this means that there is a *direct causal effect* of X on Y. The absence of an arrow between two variables implies the substantive assumption (or claim) that there is no causal effect of either variable on the other. An edge (or arc) between two variables X and Y is bi-directed when it has two arrowheads. This indicates that there is a (potentially unmeasured) common cause (or set of such causes) of X and Y that induces a probabilistic dependence between X and Y and that is not explicitly modelled in the graph, or captured by the structural causal model.

Relationships between variables are described with kinship terminology. If there is a directed edge $X \rightarrow Y$, then X is a *parent* of Y, and Y is a *child* of X. If Y can be reached, directly or indirectly, from X by following an uninterrupted sequence of directed edges, such as in $X \rightarrow Z \rightarrow Y$, then Y is a *descendant* of X. An uninterrupted sequence of directed edges connecting X and Y is called a *path*. The arrows along a path may point in any direction. A *causal path*, which is often defined relative to an intervention variable X and an outcome Y (see e.g. Elwert 2013, 12), is a path where all arrows point away from X to Y. A graph consisting of nodes and edges is *directed* when every edge

has an arrow at (at least) one end. A graph is *acyclic* when it does not contain any path that starts and ends with the same node, such as $X \rightarrow Y \rightarrow Z \rightarrow X$.

DAGs are assumed to be *complete* in the sense that it is assumed that a DAG captures everything that matters for a causal effect, causal mechanism, or process of interest. Specifically, causal DAGs are assumed to include all common measured and unmeasured causes of any pair of variables included in the DAG. They are not assumed to be complete, however, in the sense of including (explicitly) all exogenous causes of the variables that figure in the graph. Variables that are not explicitly represented in the DAG, but that bear on the values of variables in the graph, are captured by error variables such as U_X in the graph $U_X \rightarrow X \rightarrow Y$ (Scheines 1997, 189).

DAGs are non-parametric, abstract objects and while they encode qualitative causal assumptions, such as whether there is a causal relationship between a pair of variables or not, they make no assumptions or claims about the distribution of variables (e.g. normal, Poisson, etc.), the functional form association between variables (e.g. linear, nonlinear), or the magnitude of (marginal) causal effects of variables on other variables, i.e. *parameters* (Elwert 2013, 12). For this, an accompanying structural causal model (SCM) is needed.

A structural causal model M consists of four elements (sets will be denoted in bold face): 1) a set of exogenous variables $\boldsymbol{U}$ representing factors outside of the model that affect variables and relationships inside the model, 2) a set of endogenous variables $\boldsymbol{V}$ that are dependent on some subset of the exogenous variables in $\boldsymbol{U}$, 3) a set of functions $\boldsymbol{F}$ that represent the causal relationships by which values are assigned to the endogenous variables $\boldsymbol{V}$, and 4) a joint probability distribution $P(u)$ over $\boldsymbol{U}$.

With these four elements in place, a causal model M specifies that the values of the variables in $\boldsymbol{V}$ are set in accordance with the functions $\boldsymbol{F}$ and values of exogenous variables $\boldsymbol{U}$, where the probability distribution $P(u)$ over $\boldsymbol{U}$ induces a probability distribution $P(v)$ over $\boldsymbol{V}$ according to the functions $\boldsymbol{F}$.

There are two general types of models that can accompany a DAG: parametric and non-parametric. Non-parametric models specify a set of functions governing the production of the variables in $\boldsymbol{V}$ but do not expand on the functional form details (including parameter values) or the nature of the joint distribution $P(u)$ (aside from implying a set of probabilistic independencies that obtain among the members of $\boldsymbol{V}$).

For instance, a non-parametric model involving $\boldsymbol{V}$ (X, Y, Z) and $\boldsymbol{U}$ (U_X, U_Y, U_Z) could look as follows:

$$X = f_x(z, u_x)$$

$$Y = f_y(x, u_y)$$

$$Z = f_z(u_z)$$

This would correspond to no more than the simple graph $Z \rightarrow X \rightarrow Y$ (errors omitted).

Parametric models, on the other hand, offer more details on the functional form association of variables, as well as parameters figuring in the functions $\boldsymbol{F}$ in M. Here, for instance, a model could look as follows:

$$X = \alpha z^2 + u_x$$

$$Y = \log(x) + \beta z + u_y$$

$$Z = u_z$$

With a model M specified, a graph G is a causal graph of M if it captures all and only the relationships among the variables in $\boldsymbol{U}$ and $\boldsymbol{V}$ that are encoded in the functions $\boldsymbol{F}$. (B&P 2016, 7346). This relation may be a one-to-one correspondence in the case of a fully specified parametric model, or a looser connection in the case of a partially specified, semi-parametric or non-parametric model. Here, there typically exists a set $\boldsymbol{M}$ of models that are compatible with one and the same graph G, i.e. models that preserve the same qualitative relationships stipulated by the graph, and all (and perhaps only) the probabilistic independence features implied by it, but may differ (potentially radically) in their parametric details.

DAGs and their accompanying SCMs can represent interventions by means of the so-called *do-operator* (Pearl 1988). The do-operator is supposed to capture an ideal intervention on a variable X that sets X to a specific value $X = x$, thus eliminating the equations from M that describe how X is otherwise set, and leaving all other equations and hence the values of all other variables in $\boldsymbol{U}$ and $\boldsymbol{V}$ untouched, except those that are effects of X. The do-operator has graphical and formal expressions. Formally, an intervention that sets $X = x$ is written as $do(x)$, where $do(x)$ changes or replaces the function governing the value of X from $X = f_x(pa, u_x)$ (where $\boldsymbol{pa}$ is the subset of $\boldsymbol{V}$ that denotes the parents of X) to $X = x$. Graphically, the do-operator removes all arrows

pointing into X, including double-headed arrows, from a graph G, which yields the x-manipulated graph $G_{\overline{X}}$.

7.2.2 D-separation

So far, I have focused on the relationship between causal graphs and their corresponding structural causal models. However, as both graphs and their accompanying models are supposed to help us learn about causal phenomena, there must be some connection to these phenomena over and above the causal assumptions that go into building graphs and models. Specifically, the connection between graphs, models, and *data* needs to be clarified.

When aiming to answer questions about the world, graphs and models must face data first in order to help us tell whether they offer an adequate representation of the causal phenomena of interest. This typically proceeds by examining which (class) of causal graphs/models would be consistent with some observed data. Such questions are decided by investigating whether graphs and their accompanying models imply, and are hence compatible with, observed probabilistic independencies in the data as well as any antecedent information about causal relationships that is available.

On B&P's account, the connection between graphs and data is established by the graphical concept of *d-separation*. D-separation is derived from/identical to the Causal Markov Condition[1], a condition that establishes a connection between probabilistic dependence/independence and causality (Pearl 1988; 2009). In the context of causal graphs, *d-separation* is useful because it allows one to 'read off' all the probabilistic independence features entailed by the Causal Markov Condition for a given DAG.

D-separation is defined as follows (Pearl 1988): A path between two variables X and Y is said to be *d-separated* (or blocked) if:

1) The path contains a non-collider that has been conditioned on. For instance, in the causal paths $X \rightarrow Z \rightarrow Y$ and $X \leftarrow Z \rightarrow Y$, Z is a non-collider that blocks (stops the flow of information along) the path from X to Y when conditioned on.

[1] The Causal Markov Condition (CMC) states that for any two variables X and Y in a variable set $\boldsymbol{V}$, conditional on its parents, X is independent of all variables Y in $\boldsymbol{V}$ except its descendants (Hausman and Woodward 1999, 523).

2) The path contains a collider that has not been conditioned on, e.g. when $X \rightarrow Z \leftarrow Y$ and neither Z nor any of Z's descendants has been conditioned on.

A set Z is said to *d-separate* X from Y just in case Z blocks every path from a node in X to a node in Y (Pearl 2009, 17).

For instance, in the causal chains $X \rightarrow Z \rightarrow Y$ and $X \leftarrow Z \rightarrow Y$, X and Y are probabilistically dependent, but if we condition on Z they become independent. Conversely, if $X \rightarrow Z \leftarrow Y$ then X and Y are independent unconditionally, but if we condition on Z, they become dependent.

A theorem proved by Verma and Pearl (1988; see also Pearl 2009, 18) shows that if two (sets) of variables X and Y are d-separated by conditioning on a (potentially empty) set of variables Z, then X is independent of Y conditional on Z, or $X \perp\!\!\!\perp Y|Z$. This has important probabilistic implications (Pearl 2009, 18): If X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every probability distribution compatible with G. This means that the conditional independence implications derived from a graph G can be used to tell whether the graph is compatible with an observed probability distribution $P(v)$. D-separation is then used to list the set of all conditional independencies that the graph implies. If all of these independencies are realized in the data, then the graph is compatible with the observed distribution. If not, then the graph may need to be changed. This establishes a connection between the causal assumptions encoded in a graph (specifically in the form of missing arrows that indicate probabilistic independence) and observed data.[2]

7.2.3 Do-calculus

Do-calculus (Pearl 1995) is a set of syntactic rules by which expressions involving causal queries and causal quantities, such as $P(y|do(x))$, and observations such as probability distributions $P(y)$ can be transformed into and expressed in terms of each other. The rules of do-calculus are outlined in Pearl (2009) and B&P (2014) and will not be repeated here. For the present purposes, it suffices to note that such rules exist and that they are the main formal instruments by which extrapolation on B&P's approach proceeds.

[2] This assumes, of course, that the non-trivial task of inferring probability distributions from observed finite frequencies has somehow been achieved.

Specifically, do-calculus is used to derive so-called *transport formulae* for the post-intervention distributions of outcomes in a target population $P^*(y|do(x))$, i.e. the conditional probability of Y given an intervention $do(x)$ in the target (indicated by the star superscript). Here, the aim is to determine whether, using the rules of do-calculus, it is possible to express the target causal quantity $P^*(y|do(x))$ in terms of a right-hand side expression that does not contain a do-operator applied to a variable measured in the target. This is important because one of the main aims in successful extrapolation is to avoid interventions in the target, which could raise concerns about the extrapolator's bind and obviate the need to extrapolate causal effects from an experimental population in the first place. If a transport formula for a causal effect contains no expressions concerning the target to which a do-operator is applied, then no interventions in and only observational data from the target are needed to identify the effect of interest in the target.

Let me proceed to discuss the main graphical instrument that the graph-based approach employs for addressing problems of extrapolation.

7.2.4 Selection Diagrams

An extension of DAGs, called *selection diagrams*, are a graphical instrument used to represent causally relevant differences between populations (or 'disparities' as B&P call them) and enable the graph-based approach to identify causal effects in a target despite such differences. This proceeds by adding so-called *selection nodes* to variables where causally relevant differences are suspected to obtain between populations.

Building a selection diagram begins by assuming that an experimental population Π and a target population Π^* share a causal graph G', which constitutes an 'overlapping' of the causal diagrams G and G^* of both populations (B&P 2016, 7351; footnote). Two models M and M^* underlying this shared causal graph G' then induce a selection diagram D if:

1) Every edge in G' is also an edge in D.
2) D contains an extra edge $S_i \rightarrow V_i$ whenever there might exist a discrepancy in an underlying function $f_i \neq f_i^*$or background factors $P(u_i) \neq P^*(u_i)$ between M and M^*.

Selection diagrams are hence DAGs that are augmented with a set $\boldsymbol{S}$ of selection variables, i.e. variables indicating that there are exogenous causally relevant differences in the mechanisms that assign values to the endogenous variables in $\boldsymbol{V}$. According to B&P, an S-variable connected by a directed edge to an endogenous variable Z indicates differences in the 'mechanisms' that assign values to Z. Conversely, the absence of a selection node pointing to a variable Z implies that the mechanism responsible for assigning values to Z is the same in both populations.

S will hence assume different values to represent differences in these mechanisms, and switching between populations is represented by conditioning on different values of S-variables. For instance, if the target quantity of interest is the interventional distribution of Y given an intervention on X in the target, i.e. $Q = P^*(Y|\, do(x))$, then this is just equal to the interventional distribution of Y in the experiment conditional on the target's value of S, i.e. $S = s^*$. Formally, $P^*(Y|\, do(x)) = P(Y|\, do(x), s^*)$.

By encoding causally relevant differences between populations in selection diagrams, the graph-based approach involves the substantive assumption, outlined in *Chapter 3*, that differences in causal effects between populations can, in principle, be attributed to, explained in terms of, and predicted by accounting for causally relevant differences between populations; these differences are what S-variables capture. If accounted for in the right way, conditioning on S-variables as well as d-separating S-variables from the outcome by conditioning on other (sets of) variables, allows us to correctly predict causal effects in a target despite causally relevant differences. This is hence similar to the *unconfounded location* assumption involved in interactive covariate-based strategies (*Chapter 6*), where it is assumed that units' potential outcomes are probabilistically independent of population membership (or location) conditional on a vector W of interactive covariates that are believed to induce (or at least reliably correlate with) differences in causal effects between populations (see Pearl 2015 for a discussion of differences between the graph-based and interactive covariate-based approaches).

7.2.5 Transportability

B&P cover issues of extrapolation under the general heading of *transportability* (B&P 2014, 7350). Generally, transportability is understood as a licence to transport causal relations from one population to another.

More specifically, a causal relation R is transportable from Π to Π^* if it can be identified given some combination of experimental data and observational data from the target (B&P 2014, 588). B&P offer three theorems to clarify how it can be decided whether a relation R, such as $R = P^*(y|do(x), z, s)$, is transportable.

The first theorem aims at clarifying the general conditions under which R is transportable. R is transportable if it can be reduced, using do-calculus, to an expression in which the set of selection variables S that capture causally relevant differences between populations only appears as a conditioning variable in terms that do not contain do-operators (B&P 2014, 588). This ensures that the information required from the target (i.e. the terms that are conditioned on $S = s^*$) can be observational, so no interventions in the target are needed to answer the query of interest, thus satisfying an important desideratum for successful extrapolation.

The second theorem says that strata-specific causal effects, such as $P^*(y|do(x), z)$, are transportable if Z d-separates Y from S in the X-manipulated version of D, $D_{\bar{X}}$, i.e. the selection diagram in which, due to the intervention $do(x)$, all arrows pointing to X have been removed. Formally, this is represented by the following constraint:

$$(Y \perp\!\!\!\perp S|Z, X)_{D_{\bar{X}}}$$

Whenever there exists a set $\boldsymbol{Z}$ that satisfies this constraint, $\boldsymbol{Z}$ is called *s-admissible* (B&P 2014, 589). S-admissibility therefore helps identify conditioning sets that render the outcome independent of selection variables S and hence independent of *differences* in selection variables between populations, thus facilitating the identification of effects that are transportable despite causally relevant differences between populations.

More generally, d-separation and s-admissibility help us identify conditioning sets that can be used to 'shield' the causal effects of interest from causally relevant differences between populations in the form of selection variables. If and when feasible, d-separation will help us establish a conditioning 'blanket' around the effect we are interested in, such that causally relevant differences that lie outside of this blanket do not make a difference to the effect of interest one way or another. In virtue of this,

transport formulae involving expressions licenced by the d-separation or s-admissibility property can substantially reduce measurement cost; not all causally relevant differences are relevant obstacles to extrapolation and need to be accounted for.

Finally, B&P's third theorem generalizes transportability to broader classes of cases, including cases where an X-Y-effect is mediated by a variable Z and where the way in which Z mediates the X-Y-effect is different between populations, indicated by a selection node pointing into Z (e.g. Figure 4c in B&P 2014, 587), as well as more complicated cases where the X-Y-effect is mediated by yet other variables W on the path from X to Z (e.g. Figure 6a in B&P 2014, 591).

B&P invoke two subgoals to help handle such cases: The first subgoal is *trivial transportability*. According to B&P "[…] a causal relation R is said to be trivially transportable from Π to Π*, if $R(\Pi^*)$ is identifiable from (G*, P*)" (B&P 2014, 589). They further remark that "[t]his criterion amounts to an ordinary test of identifiability of causal relations using graphs […]. It permits us to estimate $R(\Pi^*)$ directly from observational studies on Π*, unaided by causal information from Π" (B&P 2014, 589). This is the case, for instance, when the selection diagram D is as follows:

$$X \rightarrow Y \leftarrow S$$

Since this graph is assumed to be complete, implying that the X-Y-effect is unconfounded, then the X-Y-relationship $P^*(y|do(x))$ can be identified from observational data from Π*, i.e. $P^*\big(y|do(x)\big) = P^*(y|x)$.

So saying that a causal relationship or effect is trivially transportable amounts to no more than saying that it can be straightforwardly identified from observational data from the target. In this sense, transportability falls outside the scope of successful extrapolation as discussed in *Chapter 3*, where (significant) changes in the evidence E obtained from an experimental population would make at least some difference to our extrapolative conclusion C. This is not the case when C is obtained only on the basis of data from the target. Hence, determining that a causal relation is trivially transportable is not a form of successful extrapolation. It only amounts to identification of causal effects from observational data in the target. However, trivial transportability can still be useful for successful extrapolation when the final transport formula derived does not only contain trivially transportable effects, and information from the interventional

distribution from a study population still figures relevantly in the final transport formula in a way that satisfies the constraints on relevance outlined in *Chapter 3*.

The second transportability subgoal invoked by B&P is called *direct transportability*: "A causal relation R is said to be directly transportable from Π to Π^*, if $R(\Pi^*) = R(\Pi)$" (B&P 2014, 589). This means that the causal relation of interest is the same in both populations, and no adjustment is needed to infer the target quantity. According to B&P, a condition for the direct transportability of a relation $R = P^*(y|do(x), z)$, i.e. the z-specific effect of X on Y, is as follows:

$$(S \perp\!\!\!\perp Y|X, Z)_{G_{\overline{X}}}$$

R is directly transportable when, in the X-manipulated version of G, $G_{\overline{X}}$, X blocks all paths from S to Y once we condition on Z. Put simply, conditioning on X and Z allows rendering Y independent of the differences induced by S, so these differences do not matter for our prediction of the Z-specific effect of X on Y in the target. This effect is (assumed to be) the same in both populations, so learning it in the experimental population allows transporting it to the target.

By itself, direct transportability is a form of successful extrapolation, albeit not a very impressive one. It applies just in case causally relevant differences between populations do not matter for the causal effect to be extrapolated (at least after conditioning on an s-admissible set $\boldsymbol{Z}$), and the causal effect is indeed known to be (or justifiably assumed to be) the same between both populations. Nevertheless, at face value, some form of successful extrapolation seems to be achieved here, at least as long as the knowledge about the target required to decide whether an effect is directly transportable is not too extensive. I will discuss this important issue later.

With these two subgoals in place, B&P's strategy is to show that more involved cases of extrapolation, where, say, X-Y-effects are not immediately directly transportable, can nevertheless be successfully addressed. This proceeds by an iterative divide-and-conquer procedure, where one decomposes an extrapolation problem into smaller sub-problems, demonstrates that these sub-problems can be solved by trivial or direct transportability, and then uses these results to establish the transportability of the relation of ultimate interest. So at least in some more involved cases where the effect of interest is not immediately transportable, but where the transportability of this causal effect depends on that of others, an iterative procedure can still be used to decide the

transportability of the effect of ultimate interest.

B&P's third theorem expresses this as follows (B&P 2014, 590): The average causal effect $P^*(y|do(x))$ is directly transportable if either of the following conditions holds:

1) $P^*(y|do(x)$ is trivially transportable.
2) There exists a set of covariates Z (possibly affected by X) that is S-admissible, and for which $P^*(z|do(x))$ is (directly) transportable.
3) There exists a set of covariates W that satisfies $(X \perp\!\!\!\perp Y|W)_{D_{\overline{X(W)}}}$ and for which $P^*(z|do(x))$ is (directly) transportable.

Condition 1) can be ignored as establishing trivial transportability of an X-Y-effect does not constitute successful extrapolation. Conditions 2) and 3) are more interesting. A combination of these conditions can be used to iteratively decide the transportability of primary relations of interest, e.g. $P^*(y|do(x))$, in more complicated cases. Here, for instance, we might need to first decide the trivial or direct transportability of another relation, such as $P^*(z|do(x))$ or $P^*(w|do(x))$, and, based on these results, we may find $P^*\big(y|do(x)\big)$ to be transportable, either directly, or with adjustment for observational distributions of other variables Z, W that bear on the X-Y-effect (B&P 2014, 590-91; Examples 8 and 9). To the extent that our derived transport formula in such cases still uses information from the experimental population in the sense outlined in *Chapter 3*, this seems to evade the extrapolator's bind.

This completes the overview of B&P's extrapolation theorems. According to B&P, these theorems show that "[…] despite *glaring differences* between […] two populations, it might still be possible to infer causal effects at the target population by borrowing experimental knowledge from the source populations" (B&P 2016, 7350; emphasis added). At face value, this satisfies at least some of the desiderata highlighted in *Chapter 3*: experimental knowledge from a source population is used to infer a causal conclusion about a distinct target, despite some causally relevant differences between the populations.

With this overview in place, let me turn to my criticisms of B&P's extrapolation strategy, which call into question whether their approach indeed manages to achieve successful extrapolation in a broad range of cases.

Graph-based approaches have been criticised on various fronts. For instance, Aalen

et al. (2016) argue that DAGs cannot adequately capture data produced by time-continuous causal processes. Others suggest that DAGs cannot adequately represent effect modification by moderating variables (Weinberg 2007, but see Elwert 2013, 255), nor reversals in the direction of causal arrows (see Hausman et al. 2014). Moreover, Deaton and Cartwright (2018), argue that DAGs 1) cannot capture simultaneous causality (as is common in economic analyses of equilibration), 2) involve substantive assumptions for extrapolation that are as extensive as those that RCTs are supposed to help us avoid, and 3) require extensive knowledge not just of differences between populations but similarities as well.

In what follows I develop two further criticisms: first, the graph-based approach is limited in what causally relevant differences it can capture, and hence in the range of causal queries it allows us to address. Second, it rests on unspoken assumptions about causally relevant similarities, and empirically supporting these assumptions is likely to fall prey to the extrapolator's bind. Let me expand on these concerns in turn.

7.3 Limitations of Selection Diagrams

The first concern is with limitations of the graph-based approach. Specifically, while it is clear how selection diagrams can represent differences in the distributions of endogenous variables, it is not clear whether they can also represent other causally relevant differences between populations. I will argue that the approach is indeed limited in some of these respects, and while some of these limitations are not principled in nature, no attempts have been made so far to discuss them or to develop ways to overcome them. Doing so is important, however, to delineate more clearly what the graph-based approach is useful for and when its limitations preclude successful application, suggesting that other strategies are needed to address the cases in question.

The general question to be addressed in this section is just what kinds of causally relevant differences between populations the graph-based approach can represent and potentially help overcome. To structure this, and recalling the different kinds of causally relevant differences delineated in *Chapter 2*, we may ask whether selection diagrams can represent causally relevant differences in 1) variable distributions, 2) functional form associations between variables and parameters, and 3) the basic structure of causal mechanisms, or indeed any combination of the above.

The answer to the first question seems clear: selection variables S "[…] represent exogenous conditions that determine the values of the variables to which they point." (P&B 2011, 160) Following this, *differences* in selection variables represent differences in the exogenous features of causal mechanisms that determine the values of the variables into which the selection variables point. Importantly, however, we do not need to know *what* exogenous differences bring about these differences in variable distributions; we only need to know *that* such differences are induced for some reason. While this makes clear that selection diagrams can capture differences in the values and distributions of variables, it remains unclear whether they can also capture differences in structural parameters, functional form association, or indeed in the basic structure of causal mechanisms.

B&P remain vague on these important issues. In some places, it seems that selection nodes S pointing into variables Z in $\boldsymbol{V}$ are only supposed to indicate that populations are selection biased with respect to the distribution of Z. Indeed, the terminology of 'selection' suggests that selection variables are supposed to capture selection mechanisms that induce different distributions of variables in different populations (see for instance B&P 2016, 7345) as a result of underlying differences in individuals' propensity to self-select (or be selected) into the respective population, i.e. differences in selection mechanisms.

In yet other places B&P seem to suggest that, particularly in virtue of the non-parametric nature of DAGs, all causally relevant differences can be captured by selection diagrams, potentially including differences in causal structure, functional form associations, and parameters. According to B&P, their approach allows „[...] extrapolating experimental findings across domains […] that differ both in their distributions and in their *inherent* causal characteristics" (B&P 2016, 7350, emphasis added). Similarly, they claim that selection variables "S may represent *all* factors by which populations may differ or that may 'threaten' the transport of conclusions between populations" (P&B 2014, 587, emphasis added). There is one way in which this is true, but another in which it is less clear whether selection diagrams can capture such differences. Let me expand.

It seems true that selection diagrams can capture all kinds of causally relevant differences between populations, in the sense that all differences in the details of the *exogenous* mechanisms by which variables Z in $\boldsymbol{V}$ are assigned their values are captured

by differences in the distributions of these variables Z.[3] As the story goes, the underlying reasons for these differences do not matter – they may be differences in the distributions of exogenous variables or in the basic structure of the (exogenous) mechanisms, functional form of structural equations, or structural parameters involved in assigning values to Z. All that matters is that there are differences in the distributions $P(V)$ assigned by such mechanisms to variables Z in $\boldsymbol{V}$. So with respect to *exogenous* causally relevant differences, those that obtain outside of the model and induce differences in $P(V)$ inside the model, the graph-based approach seems to be able to capture all kinds of causally relevant differences.

This still leaves unclear, however, whether and how selection diagrams can capture *endogenous* causally relevant differences, i.e. differences in the structure of the causal graphs G and G^* that underlie a selection diagram, and differences in the functional form association and parameters that figure in the structural equations $\boldsymbol{F}$ relating the variables in $\boldsymbol{U}$ and $\boldsymbol{V}$. Let me turn to these two issues now.

7.3.1 Differences in Basic Causal Structure

One intuitive place to look for more details on these issues is to consider again how selection diagrams are supposed to be constructed. Recall that, according to B&P, this proceeds as follows (2016, 7351):

Two models M and M^* underlying a shared causal graph G' induce a selection diagram D if:

1) Every edge in G' is also an edge in D.
2) D contains an extra edge $S_i \rightarrow V_i$ whenever there might exist a discrepancy in an underlying function $f_i \neq f_i^*$or background factors $P(u_i) \neq P^*(u_i)$ between M and M^*.

However, this definition of selection diagrams still leaves unclear how the respective graphs G and G^* of the experimental and target populations that underlie the shared causal graph G' need to relate to one another. B&P speak of populations "sharing" a

[3] This brackets cases where substantively different mechanisms nevertheless assign the same values to Z, potentially over broad variations in $P(U)$. Yet, further differences in exogenous mechanisms may still induce relevant differences in causal effects of interest outside of such stability conditions and may hence matter for extrapolation even if they are not captured by differences in selection variables.

causal diagram (B&P 2014, 588) and of the shared causal diagram G' underlying the selection diagram D being the result of an "overlapping" of the causal diagrams G and G^* of the two populations (B&P 2016, 7351, footnote).

But there are different ways in which a shared causal diagram may be constructed. For instance, G' could consist only of edges that are identical between G and G^* (an intersection, if you will), including perhaps the constraint that edges must have the same direction in order to count as shared. Or it could constitute the union of all edges in G and G^*, including edges that are only part of G but not G^*, and vice versa (this possibility seems to be favoured by Mooij et al. [2018, 13] as well as Huitfeldt et al. [2016, 7]).

However, this important issue of how shared graphs G' underlying selection diagrams are constructed is left unmentioned in B&P's papers with the exception of a footnote, where they suggest that "[i]n extreme cases in which the two domains differ in causal directionality [reference suppressed], acyclicity cannot be maintained. This complication as well as one created when G is a edge-super set of G^* require a more elaborated graphical representation and lie beyond the scope of this paper" (B&P 2014, 587; footnote 18).

This suggests that cases where the direction of the edges between a pair of variables differs in G and G^* cannot be handled by the selection diagram approach (at least not without undisclosed changes to the approach). Moreover, at face value, this may suggest that cases where there are additional edges in G that are not in G^* (or vice versa), cannot be represented by selection diagrams either. Selection diagrams seem limited, then, in that they cannot represent endogenous causally relevant differences in the structure of the mechanisms governing the outcomes of interest.

This is contravened in other places, however, adding further confusion to the issue. For instance, in an earlier paper, B&P comment that selection diagrams "[…] can also represent *structural* differences between […] two domains." (P&B 2011, 249; emphasis added) For instance, when there is an arrow from X to Y in Π but not in Π^*, they suggest that a selection node S should be added to Y to 'disable' the arrow in Π^* when $S = s^*$ and enable it in Π, when $S = s$. Based on this suggestion, they claim that their "[…] analysis will apply therefore to *all factors* by which domains may differ or that may 'threaten' the transport of conclusions between domains, studies, populations, locations or environments" (P&B 2011, 249; emphasis added). This suggests, at least, that

selection diagrams may be able to capture some differences in the basic structure of causal mechanisms concerning the presence or absence of arrows that are contained in either population.

It is unclear, however, whether this extends to cases other than the most simplistic. How, for instance, can we represent cases where an X-Y-effect is mediated by different variables in two populations, say e.g. because the effects of a social policy intervention on X on an outcome Y are governed by formal institutions Z in one population and informal social norms W in another? We can model such a situation with two different mediated paths $X \rightarrow Z \rightarrow Y$ in Π and $X \rightarrow W \rightarrow Y$ in Π^*, where Z and W may stand for some relevant mediating aspects of the formal institution and the informal social norm respectively. Here, the issue is once again unclear, as there are no mentions of such cases in any of B&P's papers on transportability. One way to handle such cases could be to add a total of three selection nodes to the selection diagram: a selection node S into Y, just as in the simpler case above, and two further selection nodes S' and S'' into Z and W respectively. *Figure 1* illustrates how the graphs G and G' could be combined into a selection diagram D:

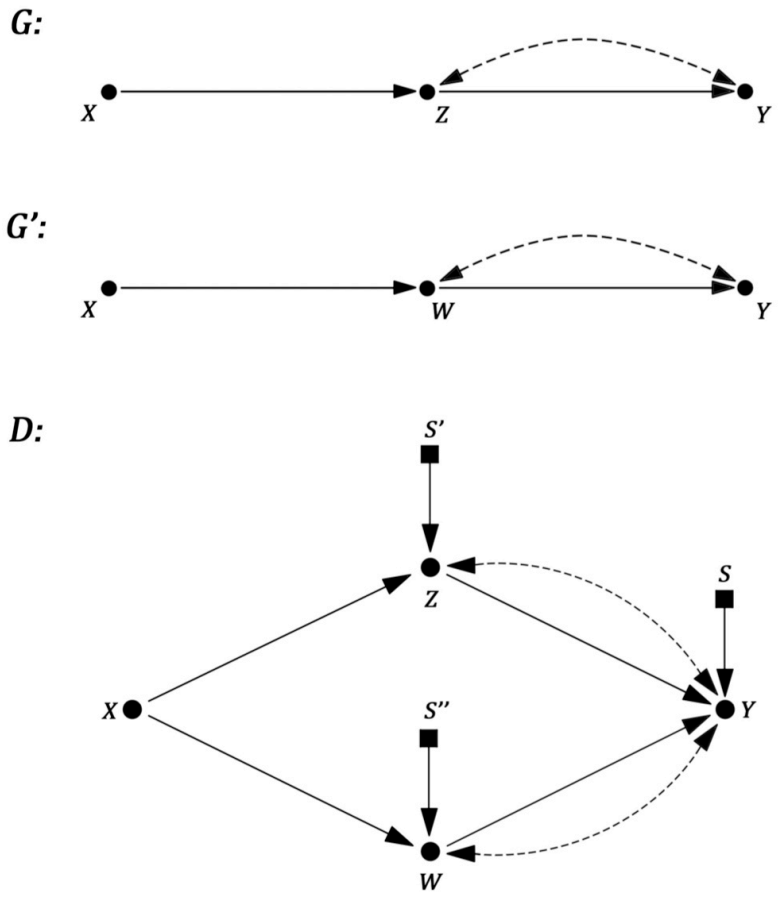

Figure 1: Graphs G and G' combined into a selection diagram D

However, it is unclear whether this would successfully represent the case. For instance, it seems that the selection node S pointing into Y would need to simultaneously disable the path $W \rightarrow Y$, enable the path $Z \rightarrow Y$ in Π, and do the opposite in Π^*. S would hence need to meddle with the *arguments* that figure in the structural equations for Y in Π and Π^* respectively, e.g. by deleting W from f_y in Π and Z from f_y in Π^*.

More involved cases can easily be envisioned, and it remains unclear whether selection diagrams could represent them. At the very least, it seems that such operations (where selection nodes add or delete arguments such as Z and W from the structural equations determining Y in different locations) would be at odds with yet other statements of B&P on how selection diagrams capture properties of two populations simultaneously. They state, for instance, that "[t]his is possible if we assume that the structural equations share the *same set of arguments*, though the functional forms of the equations may vary arbitrarily" (B&P 2013, 112). Moreover, it seems that in order to approximate the requirement that the structural equations determining Y share the same arguments, one would at the very least need to meddle with the functional form of f_y and make some auxiliary parametric assumptions, such as saying that $Y = f_y(w, z, u_y) = 0 * w + g(z, u_y)$ to indicate that although W is an argument in the equation determining Y in Π, Y is invariant to changes in W, other things being equal. While this would contravene the causal knowledge encoded in the graph G representing Π, where there is *no* arrow $W \rightarrow Y$, and W is *not* an argument in f_y, it would be the closest approximation to representing the absence of an arrow $W \rightarrow Y$ in Π. It is unclear whether this is how B&P envision such cases to be handled, or indeed whether doing so would yield yet other problems.

At this point, it would seem desirable to explore in more detail the implications of how different ways of representing this case and other, similar cases by constructing different selection diagrams would bear on the transport formulae that could be derived using do-calculus. Here, it would seem interesting to explore whether any of the inferences so licenced would be misleading, as well as whether there are other ways of reaching correct conclusions in these cases without employing the graph-based approach. Due to the substantial ambiguities left by B&P's remarks about selection diagram construction, such formal explorations are beyond the scope of this chapter, and will need to be pursued in future work.

Aside from several important vaguenesses, then, the informal considerations above at least suggest that even relatively simple arrangements of causally relevant differences between populations at the level of the basic structure of causal mechanisms might pose problems for the construction of selection diagrams, although some simplistic cases might still be captured. At the very least, it seems clear that more extensive commentary and more fully developed theory is needed to support anyone interested in building and using such diagrams for purposes of real-world extrapolation.

At worst, the above concerns suggest that selection diagrams might be ill-equipped to handle a variety of ways in which populations may differ at the level of the basic structure of causal mechanisms. This could be an important limitation of the approach, however, when considering that such cases might be commonly instantiated when extrapolating across cultural and institutional boundaries, where similar effects might be mediated by different variables that play similar functional roles. In such cases, the addition of selection nodes to a DAG will generally make it less likely that causal effects can be transported. Indeed, even the relatively simple case of two mediating paths differing between populations discussed above, may already undermine transportability. Whenever there are different pathways governing an *X*-*Y*-effect, this will need to be represented with a selection node pointing into the outcome variable. This, in turn, may often trigger variants of the so-called *s-bow arc* case, illustrated in *Figure 2* below.

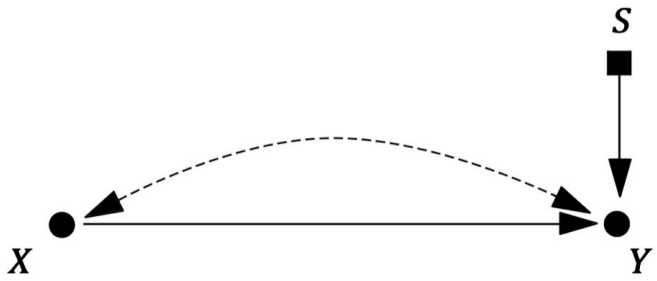

Figure 2: The s-bow arc structure

This is a case where a selection node *S* points into the outcome variable *Y*, and the path from *X* (or any other parent of *Y*) is confounded by a bi-directed arc. It marks the simplest possible case where a causal effect is non-transportable (B&P 2012, 701), and non-transportability will extend to all cases where the s-bow arc structure obtains as a subgraph. So whenever there is a causally relevant difference in how *Y* is determined, and there is a confounding arc between *Y* and its closest ancestor on the *X*-*Y*-path, this

means that transportability is precluded. Since part of this graph structure is constituted by a selection variable pointing into Y, and this will need to be the case in any selection diagram where an effect is mediated or moderated by different variables in both populations, this makes the occurrence of the s-bow arc structure likely in such cases. All that is needed here to trigger the s-bow arc structure is an additional confounding arc, which in real-world cases, especially in social science contexts where selection effects are ubiquitous, is often likely to obtain.

Importantly, however, the insight that transportability is likely to fail in the above scenario does not imply that successful extrapolation is precluded. Some kinds of informal extrapolation might still be feasible. For instance, when the mediating variables involved in these cases play functionally similar or identical causal roles in both populations, and this is supported by background theory and understanding of the mediating variables at issue, this can still justifiably increase our confidence in the similarity of causal effects between populations, and serve as a basis for some informal forms of successful extrapolation (although potentially restricted to answering qualitative queries). So despite the concern that transportability may frequently fail in such cases, successful extrapolation is by no means precluded.

In light of this, and the above arguments concerning the difficulties of representing such differences in selection diagrams, it seems that B&P should aim to make further progress on the issue of which causally relevant differences can be represented and how. Without this clarification, even relatively simple problems would seem to substantially constrain the ability of their strategy to help us achieve successful real-world extrapolation.

7.3.2 Differences in Functional Form and Parameters

Let me turn to the second question, which is whether selection diagrams can represent differences in functional form and parameters. The concerns here will be somewhat different to the above: although B&P do not comment at all on whether selection diagrams can capture endogenous differences in functional form association between variables or in structural parameters, it seems that selection diagrams can represent these cases, but it is also clear that the graph-based approach is unlikely to enable any inferences in cases where such differences obtain.

Causally relevant differences in functional form association between variables can obtain in various ways. It is easy to offer an example where such differences are important because they yield differences in post-intervention outcome distributions, which the graph-based strategy might then fail to recognize.

To provide a simple example of differences in functional form associations, assume two populations Π and Π* with models M and M^*. Let the non-parametric version of M and M^* be:

$$Y = f_y(z, x, u_y)$$

$$X = f_x(u_x)$$

$$Z = f_z(u_z, s)$$

Now let the true parametric form of f_y differ between M and M^* as follows:

$$Y = z + x + u_y$$

$$Y^* = z * x + u_y$$

So Z is an additively separable cause of Y in Π and a fully interactive moderating variable in Π*. It is clear that even at one and the same pre-intervention distribution of Z, the outcome distributions $P(y)$ and $P^*(y)$ will differ (except in cases where $z * x = z + x$). In this case, computing the outcome distribution of interest according to B&P's transport formula $Q = P^*(Y|\, do(x))$=$\sum_z P(Y|\, do(x), z)P^*(z)$ will not recover the correct outcome distribution in the target (except, again, in cases where $z * x = z + x$).

Similarly, when structural parameters differ between populations, we might face a situation such as the following. Let the true parametric form of f_y differ between M and M^* as follows:

$$Y = \alpha_{ZY} z + \alpha_{XY} x + u_y$$

$$Y^* = \beta_{ZY} z + \beta_{XY} x + u_y$$

where

$$\beta_{ZY} \neq \alpha_{ZY} \neq \beta_{XY} \neq \alpha_{XY}$$

Again, even at one and the same pre-intervention distribution of Z and X, the outcome distributions $P(y)$ and $P^*(y)$ will differ and computing the outcome distribution of interest according to B&P's transport formula,

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

will not recover the correct outcome distribution in the target.

Although B&P do not discuss such cases, it seems that causally relevant differences in functional form and parameters can be represented in the same way in a selection diagram. In both cases, a selection node needs to be added to Y (or any other variable whose structural equation is the object of such differences) in order to capture such differences.

However, while it is important to recognize that selection diagrams can capture such differences, it is an entirely different issue of whether the graph-based approach can enable any inferences in cases where they obtain.

Causally relevant differences in functional form association and especially in parameters are likely in many real-world problems of extrapolation. Think, for instance, about the ubiquitous and significant real-world variation in agents' response to evidence-based development policy interventions, as documented in meta-analyses such as Vivalt's (2019). The insight that each such difference may require the addition of selection variables to a selection diagram creates important practical challenges for the graph-based account. As B&P recognize, in the limit, when selection nodes need to be added to all variables in a selection diagram, transportability is entirely precluded. Even significantly milder cases, such as occurrences of the s-bow arc structure, preclude transportability already. How, then, can we avoid this? The answer is simple: we can only avoid this problem in cases where few causally relevant differences exist in a selection diagram. At the same time, this means that our efforts to extrapolate will be burdened with making a potentially large number of substantive assumptions. Whenever there is no selection node pointing into a variable, this expresses the substantive assumption that populations are identical at all three levels discussed here, i.e. there are no differences in the basic structure of the causal mechanisms, functional form association and parameters, or variable distributions. Clearly, these cannot be assumptions of convenience, but must be supported. In the second line of criticism that I will now turn to, I focus on these assumptions and argue that underwriting these

assumptions is likely to raise concerns about the extrapolator's bind, even in cases where there are few causally relevant differences to begin with.

7.4 Causal Assumptions and the Extrapolator's Bind

So far, I have raised concerns about limitations of the graph-based approach with respect to encoding certain kinds of causally relevant differences. Even if such limitations did not obtain, however, and even in those cases where they do not apply, there remain further important obstacles to successfully using the graph-based strategy for extrapolation. These are of an epistemic nature and concern the issue of what is required to underwrite graph-based extrapolation.

The main question to be addressed in this section is: how can we support the crucial assumptions about similarity and difference between populations involved in constructing selection diagrams? I will argue that several important epistemic challenges in underwriting such assumptions have been unhelpfully glossed over by B&P, but need to be addressed in order to assess whether the graph-based strategy is a promising candidate for overcoming real-world problems of extrapolation. As I will argue, the graph-based approach encounters important problems: the knowledge about the target required to underwrite assumptions about similarity and difference between populations is likely, once again, to be so extensive as to trigger concerns about the extrapolator's bind.

At this point, one might wonder whether this line of criticism is misguided from the beginning. One may ask, for instance, whether supplying the details of empirical strategies to underwrite the assumptions required by the graph-based approach should even be part of this approach, or whether this would be too much to ask. B&P might insist that their approach does not aim to provide a manual for extrapolation from start to finish. For instance, they might argue that their approach does not aim to clarify how the estimation of the quantities demanded by transport formulae to compute causal effects should proceed; this would be left to statistical approaches that can complement their graph-based strategy, if and when available. What is more, other approaches, too, require extensive causal and probabilistic information from both populations to licence

extrapolation.[4] B&P might also point out that even without providing further details on these issues, the graph-based approach nevertheless provides results that are interesting and practically relevant. On this narrative, their approach contributes to overcoming problems of extrapolation by offering general algorithms to derive transport formulae, and in doing so, abstracts away (and perhaps needs to abstract away) from a variety of substantial and non-trivial empirical challenges, such as measuring probability distributions and other quantities needed to compute causal effects with the help of transport formulae.

This is not an uncontroversial stance, as it may seem unclear whether and how the graph-based approach substantially contributes to addressing real-world problems of extrapolation if it were to face insurmountable obstacles at the estimation stage that would preclude us from ever obtaining informative answers to our extrapolative queries. B&P might reply, however, that it is not a shortcoming of the approach, but indeed a virtue that, at least on reflection, helps bring such issues to the fore. Even so, our conclusions about the real-world capabilities of the approach should then perhaps be phrased somewhat more cautiously, at least adding that researchers may expect substantial downstream complications when engaging in graph-based extrapolation.

I will not take further issue with this debate.[5] The concerns I develop here are different, and focus on what seems to me a more basic desideratum for successful uses of the graph-based strategy: validating the empirical causal assumptions that it requires. This, I maintain, is an issue that cannot be abstracted away from as easily as, for instance, the finite-sample statistical complications involved in measurement and estimation of probability distributions and parametric details. Moreover, this concern is epistemically prior to such issues, in that it obtains before any selection diagram capturing any real-world problem of extrapolation can be constructed.

My concern is the following: if we cannot adequately support the empirical assumptions needed to construct selection diagrams, then how could the graph-based approach be useful for real-world extrapolation? Without adequate support, very little can be hoped to be learnt from using the graph-based approach, as one might think,

[4] While it is true that other approaches can be similarly epistemically demanding, B&P's approach does seem to involve a thicker layer of statistical inference to function than, for instance, interactive covariate-based extrapolation, which can proceed from frequency data alone (e.g. on covariate distributions), and mechanism-based extrapolation following Steel, which makes few quantitative assumptions to begin with.

[5] Interested readers may follow the discussions on Andrew Gelman's blog between several statisticians and both Elias Bareinboim and Judea Pearl: https://statmodeling.stat.columbia.edu/2015/12/05/28262/

following Cartwright, that an extrapolation is only ever as strong as its weakest link (Cartwright, 2013, 16).[6] If the derivation of transport formulae hinges on substantive assumptions about similarities between populations that are difficult to substantiate beyond the level of mere assumptions, then this would fall radically short of reasonable demands for how extrapolation should be justified and we might as well extrapolate based on hope alone, without the substantial complications involved in drawing graphs that cannot be empirically supported and deriving transport formulae that may lack any empirical bite. Put differently, and putting a spin on Cartwright's famous dictum "no causes in, no causes out" (Cartwright, 1989), we might say "no justified causal assumptions in, no justified causal conclusions out". It is clear that successful graph-based extrapolation requires justification for the underlying assumptions that are necessary to perform it. If there is no such justification, then any conclusion obtained from this method will also lack justification, and hence not be useful for addressing real-world problems of extrapolation involving actual stakes, e.g. when policy action could be taken based on potentially radically mistaken conclusions about the effectiveness of a policy in novel contexts.

So for the graph-based strategy to be a promising candidate for addressing real-world problems of extrapolation, we need some ideas about how to support the substantive assumptions that it involves. My demand here is not, of course, that these ideas need to be provided by B&P. However, providing some supplementary account of how to support such assumptions must be possible, and in a way that does not raise concerns about the extrapolator's bind. In what follows, I argue that providing such an account will be extremely difficult, and that these difficulties call the applicability and usefulness of the graph-based strategy into question. While this does not touch upon the validity of the results that can be derived by using it, or the generality of these results underwritten by the completeness of do-calculus, my arguments suggest that the graph-based strategy, by itself, is ill-equipped to handle a wide variety of real-world extrapolation problems. As a result, B&P's claim that the problem of extrapolation has been solved by their strategy (B&P 2016, 7352) is unwarranted. This claim unhelpfully glosses over basic epistemic problems that need to be addressed first, i.e. before any selection diagram can be constructed and any practically relevant licences to transport causal effects can be established.

[6] To be sure, I do not wish to endorse this view here, as there are cases where it seems less plausible. I expand on this issue in *Chapter 8*.

To show what assumptions the graph-based approach involves and how validating them raises concerns about the extrapolator's bind, let me revisit two examples repeatedly used by B&P to illustrate the capabilities of their approach.

7.4.1 Selection Diagrams and the Extrapolator's Bind

As outlined above, the main aim in graph-based extrapolation is to identify a causal effect in a target using 1) a selection diagram D that represents causally relevant differences between experimental and target populations, and 2) using do-calculus to determine whether the causal quantity of interest in the target can be expressed in terms of experimental data from a study population and observational data from the target. To evade the extrapolator's bind, the aim is to show that the target quantity can be expressed in terms of expressions about the target that do not contain a do-operator.[7] If this is successful, it seems that the extrapolator's bind is indeed evaded, as no intervention in the target is needed to reach a conclusion about the causal quantity of interest.

Let me revisit a simple example used by B&P and elaborate on the underlying assumptions involved, how these assumptions might be supported empirically, and how doing so raises concerns about the extrapolator's bind. In several of their papers, B&P offer the example of a causal effect of X on Y that is estimated in an experiment in LA, and the causal query of interest concerns the effect of X on Y, i.e. $P^*(Y|\, do(x))$, in a distinct target population in NYC. As before, call these populations Π and Π* respectively. The selection diagram D representing the two populations is as follows:

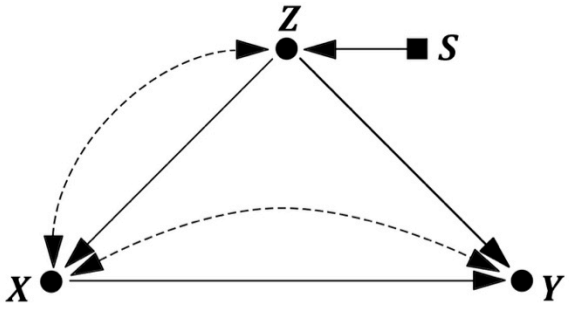

Figure 3: selection diagram reproduced after B&P (2016, 7350; fig. 5a))

[7] This is necessary but not sufficient to fully evade the extrapolator's bind (at least in cases where the do-operator at issue would be applied to X in the target). We might still be in possession of observational rather than interventional data, which may nevertheless permit identification of the effects of interest from these data alone.

As usual, X is the intervention variable and Y the outcome. Z denotes individuals' age and is a common cause of X and Y. In addition, there are two confounding arcs between X and Z and between X and Y respectively, indicating that there are further, unmeasured common causes of these pairs of variables. Finally, the S-node pointing into Z denotes the existence of a selection variable S, which induces differences in the distribution of age Z between Π and Π*.

The extrapolation proceeds as follows: The query to be answered is $Q = P(y|do(x), S = s^*)$. So we are interested in the post-intervention distribution of Y given $do(x)$ and given that the selection variable S assumes the value representing the target, namely s^*. Since the difference in S marks the only difference between populations, this is equal to the target's post-intervention distribution $P^*(y|do(x))$, i.e. the quantity that we want to learn. The aim is now to derive a transport formula that combines experimental data from Π and observational data from Π*, i.e. $P(y|do(x))$ and $P^*(x, y, z)$ respectively, to obtain an interventional distribution $P^*(y|do(x))$ in Π*. I will not expand on the details of the derivation here, which can be found in B&P (2014). For the present purposes it suffices to note that by using the rules of do-calculus the quantity of interest $Q = P^*(y|do(x))$ can be rewritten as follows (B&P 2014, 585):

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

In plain English, the post-intervention distribution of Y in the target is equal to the sum over all z of the product of the z-specific interventional distribution of Y in the experimental population and the observed distribution of Z in the target.

So what happens here is merely a traditional re-weighting of causal effects (understood in B&P's sense as outcome distributions). Similar to interactive covariate-based extrapolation discussed in *Chapter 6*, the effects are estimated conditionally upon a variable Z known to influence these effects (i.e. distributions), before reweighting these effects (i.e. distributions) by conditioning on the observed distribution of Z in the target.

The extrapolation discussed by B&P involves several substantive assumptions. Some of these are encoded in the selection diagram, while others remain implicit. Specifically, it is assumed that (in no specific order):

1) The joint probability distributions $P(x, y, z)$ and $P^*(x, y, z)$ are measured in

both populations (where $P(x, y, z)$ is often an interventional distribution called $I(x, y, z)$). For most real-world problems of extrapolation, this assumes that these distributions have somehow been inferred from finite frequency data.

2) For any variable Z that modifies the causal effect of interest, the z-specific causal effects of X on Y have been measured in the experimental population over all values of Z.

3) Sufficient knowledge and data about Π and Π* are available to permit construction of the two models M and M', the two DAGs G and G^*, the shared causal graph G', and the selection diagram D that is induced on G'.

4) The two models M and M', and the two DAGs G and G^* underlying the selection diagram D, are consistent with the measured joint distributions $P(x, y, z)$ and $P^*(x, y, z)$ respectively.

5) G and G^* are complete, i.e. there are no unmeasured common causes of any variable in $\boldsymbol{V}$ that are not encoded in the graphs.

6) The two populations only differ in their age distribution $P(z)$ or, as B&P put it, in the underlying *mechanisms* that determine the age distribution, such as differences in a selection mechanism by which agents at different age levels are differentially drawn to Π and Π*. This implies that all functions in $\boldsymbol{F}$ and structural parameters figuring in those functions that bear on the causal effect of interest are the same in both populations, including that z-specific effects of X on Y are *invariant* between populations.

These are strong assumptions. To name just a few standard concerns: 1) is highly controversial among econometricians and statisticians dealing with real-world cases that pose the non-trivial problem of inferring probability distributions from finite frequency data.[8] 2) is highly controversial, too, as even large-sample RCTs do not straightforwardly permit inference of strata-specific causal effects. To obtain unbiased estimates of such effects, multiple trials, potentially along many finely partitioned strata, might need to be conducted. This concern is multiplied by the number of potential moderating and non-linearly mediating variables involved in governing a causal effect of interest, and is even further aggravated if such variables interact with one another. 5), too, poses a formidable epistemic challenge that is not recognizably

[8] For an overview, consider the discussion at https://statmodeling.stat.columbia.edu/2015/12/05/28262/

easier to overcome than the challenges raised by standard unconfoundedness and ignorability assumptions involved in statistical estimation of causal effects from observational data (see Deaton and Cartwright 2018).

In what follows, I do not take further issue with these concerns. I will focus instead on assumptions 3) and 6). These assumptions raise issues that are epistemically prior to the challenges posed by 1), 2), and 5), and are indeed prior to the construction of any selection diagram, which is needed before B&P's strategy can be applied to any real-world problem of extrapolation.

Assumptions 3) and 6) can be discussed in one fell swoop. The only causally relevant difference between populations permitted, and explicitly encoded in the selection diagram, is in the distribution of age Z. Conversely, this entails that the two populations must be identical at the levels of causal structure, functional form association, and structural parameters.[9] Let me discuss these assumptions in turn.

First, even if we assume that the graph G for the experimental population is correctly specified, to construct the above selection diagram we must still ascertain that G^* does not differ from G at the structural level. To do so, we must learn something about G^* from information pertaining to the target. B&P suggest that G^* can be learnt by means of causal discovery methods (B&P 2012, 700, footnote 7; for details on causal discovery methods see Spirtes et al. 2000; Pearl and Verma 2001; see Hyttinen et al. 2015 for an attempt to integrate causal discovery and do-calculus-based identification methods; see Mooji et al. 2018 for a similar proposal). However, particularly in predictive extrapolation scenarios considered in *Chapters 5* and *6*, it seems unlikely that this can be successful. When the intervention of interest has not yet been experienced in the target, and important parts of the mechanisms of interest have remained 'dormant', there will often be insufficient variation in variables of interest in the target setting, and observational data from the target will hence not contain suitable probabilistic independence information required for orienting the edges of a graph by causal discovery methods. So without suitable variation, particularly in the intervention variable and mediating variables on the pathway from the intervention variable to the

[9] To be sure, some differences are still permitted, but only if they apply to nodes that can be screened off from the effect of interest by some s-admissible conditioning set. However, this should not be understood as a licence to be sloppy about potential causally relevant differences that might turn out to be irrelevant. In the pursuit of transparency and rigour, it still seems preferable to settle issues of (potential) causally relevant differences first, and *then* demonstrate that they are immaterial to an extrapolation.

outcome, causal discovery methods cannot learn causal graphs in the target as a basis for comparing them with the study population graph.

Similarly, in cases where the intervention of interest has been experienced in the target in the past, or some closely related and well understood exogenous changes in the variables on the path from intervention to outcome variable have taken place, there are still reasons to think that this could raise concerns about the extrapolator's bind. A definitive proof for this concern lies beyond the scope of this chapter, but the key idea is simple. It seems that, especially for complicated graphs, observational data from the target that would be sufficient to learn the target graph G^* as a basis for building a selection diagram D would already be extensive and informative enough to allow straightforward identification of the effect of interest in the target (see Hyttinen et al. 2015 for related concerns).

This means that, if extrapolation is to be supported at all, information with bearing on questions of similarity and difference in basic causal structure must be imputed from elsewhere. As discussed in the previous chapter, this will often mean that we have to consult sources of evidence that are typically deemed less reliable than quantitative observational data. This can have important ramifications for what degree of empirical support graph-based extrapolation can enjoy. With quantitative data we can, under somewhat milder assumptions, trust that the mechanisms of interest have faithfully written important identifying information into the data that we can obtain, e.g. in the form of distinctive probabilistic independence signatures.[10] With other sources, this might not be the case, and significantly stronger assumptions might be necessary. For instance, we can often not readily trust agents' reports about what causes what, and in what way, as there might not be any sufficiently reliable connection between such testimony and relevant features of the causal mechanism of interest. This concern is aggravated when it is plausible to suspect that individuals reporting on such issues have never experienced the mechanism in question being operational, which is particularly acute in many areas of EBP where novel interventions are considered for deployment in a target.

An easy way out of this problem would be to intervene in the target to learn crucial parts of the mechanism there but this, of course, threatens to fall prey to the extrapolator's bind. So B&P's suggestion (2012, 700; footnote 7) to use causal

[10] At least when mechanisms have been appropriately ‚active'.

discovery methods for learning the target graph underlying a selection diagram (or the study population graph, for that matter) is likely to be either unsuccessful or trigger the extrapolator's bind.

My second concern is that the selection diagram used by B&P implies that populations must be identical at the level of functional form of the structural equations and the parameters figuring in these equations. This may seem surprising at first, as the graph-based approach is decidedly non-parametric, i.e. results derived with its help should hold irrespective of the parametric details about functional form and parameters involved in the (potentially unknown) structural equations that best represent how the outcomes of interest are produced. So the non-parametric nature of the graph-based strategy may suggest that questions about identity in functional form and parameters do not matter; they concern parametric details, and any results derived by using the graph-based strategy should remain untouched by arbitrary variations in such details.

However, it is important to specify more clearly what exactly the non-parametric nature of the approach entails for the assumptions that must be made about experimental and target populations. It is true that using graph-based extrapolation theorems does not require full-fledged knowledge of all parametric details of both populations. We do not need to know, nor explicitly model, the functional form of the causal relationships that hold between any two variables, nor do we need to assume or estimate structural parameters or make distributional assumptions, in order to derive transport formulae. A wide range of such formulae will be valid non-parametrically, i.e. irrespective of the particular parametric details that govern the outcomes of interest in the two populations. This is an important advantage of the graph-based strategy, since it is potentially considerably less epistemically demanding than other, statistical approaches, which may require full-fledged parametric specifications of outcome models (Pearl 2015).

However, while it is true that transport formulae are valid non-parametrically, this does not mean that no parametric commitments whatsoever are required. For many interesting problems of extrapolation, in particular those involving moderating variables Z of an X-Y-effect or non-linear mediating variables such as those discussed in *Chapter 2*, we must assume that z-specific effects are *invariant* between populations. This is also true in B&P's simple example. While we can remain non-committal on the exact functional form of the functions $\boldsymbol{F}$ involved in M or M^*, and the exact values of the structural parameters involved in these functions, we need to entertain some parametric

assumptions (which currently remain unspoken). These do not pertain to either model M or M^* in isolation, but they apply to the *relation* between these models.[11] Assuming that z-specific effects are invariant between Π and Π* will typically mean that we must assume that the functional form of structural equations involving Z and the values of structural parameters figuring in these functions pertaining to Z are *identical* between populations, whatever they might ultimately turn out to be.[12] If this were not the case, extrapolative conclusions sanctioned by the approach could be radically off the mark.

To illustrate, assume two populations Π and Π* with models M and M^*. Let the non-parametric version of M and M^* be:

$$Y = f_y(z, x, u_y)$$

$$X = f_x(u_x)$$

$$Z = f_z(u_z, s)$$

Two populations can be identical at this level, but differ at the level of parametric details, e.g. when the form of the functions $\boldsymbol{F}$ involved in M or M^* or structural parameters figuring in these functions differ between them. For instance, f_y may differ between M and M^* as follows:

$$Y = \alpha_{XZY} * z * x + u_y$$

$$Y^* = \beta_{XZY} * z * x + u_y$$

where

$$\beta_{XZY} \neq \alpha_{XZY}$$

In such a case, z-specific effects would not be identical, as the structural parameters β_{XZY} and α_{XZY} relevant to governing the X-Y-effect differ. So even if x, z, and u_y are identical, marginal effects of $do(x)$ on Y will differ between populations. Wrongly assuming identity between β_{XZY} and α_{XZY}, in turn, would likely bias our predictions of $P^*(y|do(x))$. For instance, β_{XZY} could be positive and α_{XZY} negative, so adjusting

[11] This is, essentially, a *representation* of the relation $\boldsymbol{R}$ that obtains between experimental and target populations.

[12] I say 'typically' because it is in principle possible that differences in functional form do not induce differences in causal effects for some region of two functions. For instance, two non-monotonic functions may differ along some range of their arguments but are identical along another range that is more typically encountered in practice. Moreover, we might also envision cases where differences in parameters between populations cancel each other out. I do not consider such fortuitous arrangements to be practically relevant.

$P^*(y|do(x))$ for the difference in the distribution of z induced by s (note that Z depends on s in the structural model) would likely lead to upwards-biased predictions (for a positive change in X induced by $do(x)$).

More generally, whenever structural parameters or other parametric details differ between populations, this requires the addition of selection variables at those nodes into which the arrows point whose underlying parametric details differ. The absence of such nodes, in turn, indicates the substantive, but unspoken, assumption that populations are *parametrically identical*, although they may still vary arbitrarily *as long as* they are parametrically identical. All edges in a selection diagram without a selection node into the child express this assumption. This is especially important for outcome nodes Y into which two edges Z and W point. Without further parametric assumptions, Z and W may always interact, e.g. because Z is a moderator of the W-Y-effect or vice versa. In all such cases, the particular form of interaction may vary arbitrarily, but it must be identical between the two populations if there is no selection variable into Y.

How, then, can such assumptions about parametric identity be empirically supported? The easiest way, of course, is to estimate functional form relationships and parameters in both populations and compare them. This, however, is likely to raise concerns about the extrapolator's bind.

Those cases where causal effects are transportable at all will need to involve at least some nodes that are free from selection variables, and will hence involve arrows that require assumptions about parametric identity, most importantly those on directed paths from an intervention to an outcome variable. For these relations, we would either need to estimate functional form and structural parameters from observational data or obtain them from experimental data.[13] Both strategies raise concerns about the extrapolator's bind. Let me return to the selection diagram of B&P's example to concretise:

[13] I am assuming realistic settings where the functional form of data generating processes is not unambiguously handed to us by background theory but must, at least in part, be learnt and disambiguated from data, such as when multiple specifications (e.g. linear, quadratic, cubic, etc.) are tested for fit and the best fitting specification is chosen as a basis for further investigations.

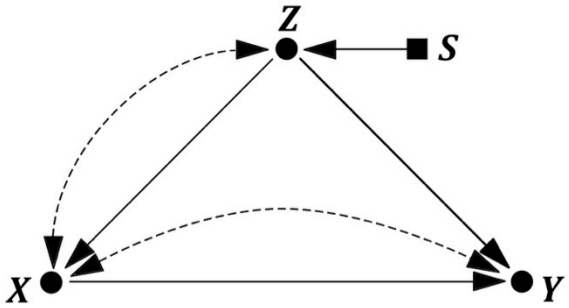

Figure 3: Selection diagram reproduced after B&P (2016, 7350; fig. 5a))

This diagram encodes the assumptions that Π and Π* are parametrically identical in the $X \rightarrow Y$, $Z \rightarrow Y$, and $Z \rightarrow X$ arrows. Since we are extrapolating from an experiment on Π in which $do(x)$ was performed, we can ignore the $Z \rightarrow X$ arrow, since this arrow (and the corresponding dashed arc) will be deleted in both populations in the mutilated graph $G_{\bar{X}}$ where X is intervened on:

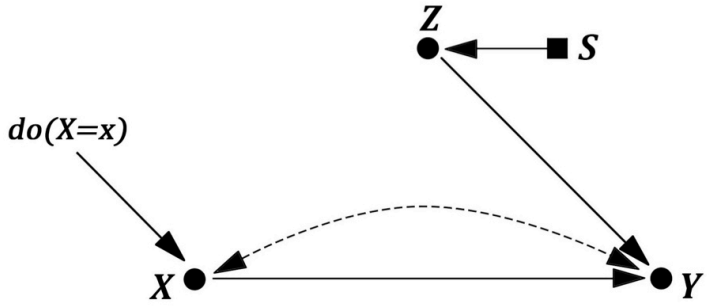

Figure 4: Mutilated graph $G_{\bar{X}}$

This still leaves us with two arrows, however: $X \rightarrow Y$ and $Z \rightarrow Y$. The $X \rightarrow Y$ effect in Π is known. B&P's transport formula for this example makes use of direct transportability. Recall that a causal relation R is said to be directly transportable from Π to Π* if $R(\Pi^*) = R(\Pi)$. The relation at issue here is $R = P(y|do(x), z)$, i.e. the *z*-specific causal effect of X on Y. So we are asserting that this conditional causal effect is the same in both populations. However, in B&P's extrapolation this is only true by stipulation. Even though the transport formula makes it seem as if the direct transportability of $P(y|do(x), z)$ is the *result* of the extrapolation, it is just identical to one of the *assumptions* underlying the selection diagram D, where it was assumed, by stipulating the absence of further selection nodes, that age-specific effects are the same. So not much is learned by direct transportability over and above what is already assumed when the selection diagram is construed. The crucial question however, which is bracketed in B&P's discussion, is how we can support such assumptions.

One way to support that the z-conditional $X \rightarrow Y$ effect is parametrically identical in the target is to identify and measure it in the target as well. The same is true for $Z \rightarrow Y$. If this is possible from observational data from Π^*, this falls prey to the extrapolator's bind, because the z-conditional $X \rightarrow Y$ effect together with the distribution of z already answers our query about the target. The same, of course, is true if we were to learn the z-conditional $X \rightarrow Y$ effect in the target by intervening on X (and potentially Z) in Π^*. In either case, there would be no need to extrapolate from Π, as the information acquired from the target is sufficient to answer our causal query.

So, if the parametric identity assumptions involved in this extrapolation are supposed to be supported by means of observational or experimental data from the target, no ampliative inference takes place. The knowledge about the target required for constructing the selection diagram and supporting the assumptions it encodes already contains the answer to our query and extrapolation subsequently fails, even though a transport formula that seems to help evade the extrapolator's bind has been successfully derived.

The important thing to recognize, then, is that whether or not the extrapolator's bind is evaded by B&P's approach cannot be decided only on grounds of what measurements from the target are required by a transport formula. The construction of the selection diagram, too, must be considered when assessing whether the epistemic demands involved in using graphs to extrapolate raise concerns about the extrapolator's bind.

This also makes clear again that, to evade the extrapolator's bind, we may need to consider sources of support for the required assumptions other than data from the target. For instance, strong background theory attesting that populations are invariant in features pertinent to the effects of interest, or other, well-established licencing facts about causal invariance that cover both populations, could help us provide such support. The important feature that these sources of support have in common is that they increase our confidence about parametric identities between populations without requiring a detailed look at the causal or probabilistic makeup of the target and establishing these identities manually, as it were. While they will perhaps not be strong enough to fully warrant assumptions about parametric identity, they may nevertheless provide enough support to think that the risks of error in extrapolation are adequately addressed, and that our extrapolation is supported by more than hope alone.

At the same time, this is where things will often get difficult in practice. Strong enough background theory and licencing facts are often unavailable in social science settings (in contrast to epidemiology, for instance), and we are soon back to square one where it remains unclear whether the rigour and transparency purportedly offered by the graph-based approach are of any help when we experience difficulties in underwriting the assumptions that it involves. Again, proponents of the graph-based approach might be quick to respond that it is a virtue, not a shortcoming, of their strategy that it informs us about just how challenging extrapolation can be and what knowledge and how much support is required for it to be successful and sufficiently credible. Indeed, it seems true that B&P's approach makes important progress here as it helps us pinpoint more precisely what assumptions are in need of support, and what assumptions we may need to bet on if support is hard to come by. In B&P's words:

> Our analysis is based on the assumption that the learner is in possession of sufficient knowledge to determine, at least qualitatively, where two domains may differ. In practice, such knowledge may only be partially available and, as is the case in every mathematical exercise, the benefit of the analysis lies primarily in understanding what knowledge is needed for the task to succeed and how sensitive conclusions are to knowledge that we do not possess (P&B 2011, 253).

But this is precisely the point of contention. While similar remarks are made in several of their papers (although this is by far the most explicit example), for the most part B&P gloss over epistemic obstacles that are substantial enough to render unclear whether their approach can provide us with an epistemically and practically feasible strategy for addressing any real-world problems of extrapolation, and whether it can outperform other, perhaps more informal, modes of extrapolative reasoning once real epistemic constraints are taken into consideration. This is aggravated by the arguments provided above, as they suggest that not only is graph-based extrapolation epistemically demanding in general, but also, even if we could obtain the necessary "knowledge that we do not [yet] possess", this may often render the very act of extrapolation from a study population redundant to our purposes.

There is hence a substantial gap between B&P's suggestions that the problem of extrapolation has been solved by their approach (B&P 2016, 7352), if and when transport formulae can be derived, and realizing that even in cases where a transport formula can be derived, a large number of real-world problems of extrapolation cannot

be successfully addressed by their approach since it falls prey to the extrapolator's bind and a supplementary strategy to help evade this problem remains unavailable.

7.4.2 Transport Formulae and the Extrapolator's Bind

Let me turn to a second, more involved example used by B&P which helps illustrate how concerns about the extrapolator's bind are further aggravated by the distinction between attributive and predictive extrapolation developed in *Chapter 5*.

In several papers, B&P offer a more complicated example which is supposed to illustrate how their third, more general extrapolation theorem works, i.e. the divide and conquer-type strategy. Here, we begin from a more involved selection diagram, where the relation of interest is not immediately transportable. This problem is then divided into sub-problems of finding transport formulae for relations other than that of ultimate interest. If such formulae exist, they can be combined to yield a transport formula for the effect of ultimate interest. The selection diagram used by B&P is as follows:

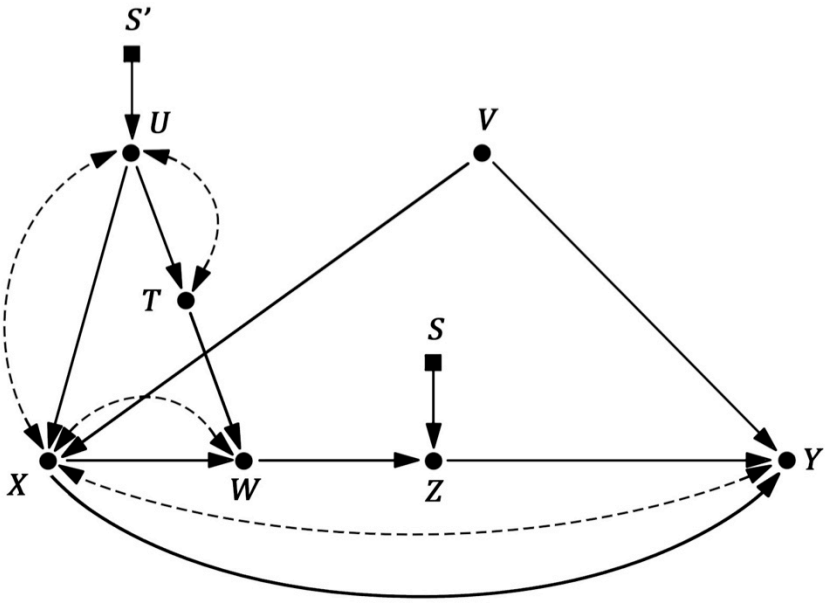

Figure 5: More complicated selection diagram after B&P (2013, 110; fig.2)

As usual, the aim in this example is to answer $Q = P^*(Y|\, do(x))$. There are two causally relevant differences that present potential obstacles to successful extrapolation, S pointing into Z and S' pointing into U.

The derivation of a transport formula for $P^*(Y|\, do(x))$ in this case is somewhat more involved, and requires application of B&P's third extrapolation theorem (B&P

2014, 590). The final transport formula reached after iterated application of this theorem is as follows (see P&B 2011, 254 for the derivation):

$$\sum_{z} P(y|do(x), z) \sum_{w} P^*(z|w) \sum_{t} P(w|do(x), t)\, P^*(t)$$

Intuitively, this transport formula seems to evade the extrapolator's bind. As B&P remark, "[t]he main power of this formula is to guide the learning agent in deciding what measurements need be taken in each domain. It asserts, for example, that variables U and V need not be measured, that the W-specific causal effects need not be learned in the experiment and only the conditional probabilities $P^*(z|w)$ and $P^*(t)$ need be learned in the target domain." (P&B 2011, 252)

Just like the previous case, this case raises potential problems with supporting structural and parametric identity assumptions implied by the selection diagram. Such problems are raised, for instance, by the requirement that T-specific effects of X on W and the W-specific effects of X on Y are invariant between populations, since otherwise, reweighting by $P^*(t)$ would not recover the correct quantity. I will bracket such concerns here, however, and instead focus on issues surrounding the quantities that the transport formula instructs us to estimate in the target, i.e. $P^*(z|w)$ and $P^*(t)$.

My concern is that measuring these quantities in a way that is informative for the extrapolation can, again, raise concerns about the extrapolator's bind. This is best illustrated by considering a concrete causal narrative for the selection diagram. As B&P do not offer a concretisation of their diagram in terms of a real-world extrapolation, I will impute a narrative onto the diagram to make my concerns more vivid.

Drawing on the *bed net example* used in *Chapter 6*, let us assume that *Figure 5* is a graph representing the suspected causal mechanism governing the effectiveness of an intervention seeking to decrease malaria infection in a population. Due to the slightly more complex structure of the graph, I will assume that the intervention of interest on X is a composite intervention consisting of distributing insecticide-treated bed nets and antimalarial drugs. There are two directed paths from X to Y. I will assume that the unmediated arrow directly from X to Y governs the effects of antimalarial drugs and that the second, mediated path from X to Y governs the effect of bed nets. The latter path is mediated by W, which I understand as properly installed bednets, and Z, which I understand as mosquitos in the proximity of an agent. I will assume that X is positively

relevant for W, i.e. distributing nets contributes to properly installed nets, and that W is negatively relevant for Z, as properly installed nets hinder mosquitos from biting humans. Z, in turn, is positively relevant for Y, malaria infection. Further, I will understand U as the level of education an agent has, V as their understanding of malaria transmission mechanisms, and T as agents' wealth.

Let me explain the arrows obtaining between these variables, starting with directed arrows and moving on to bi-directed arcs. The arrow between U and X indicates that education plays a role in agents' decisions to obtain bed nets. The arrow from U to T indicates that education also plays a role in determining agents' wealth position. The arrow from T to W indicates that wealth plays a role, perhaps a moderating one, in agents' propensity to properly install bed nets. For instance, higher wealth could mean that agents have weaker incentives to use bed nets for alternative purposes, such as for fishing. V, i.e. agents' understanding of malaria transmission mechanisms, is a common cause of X and Y. Here, the path from V to X could capture that agents with more sophisticated understanding of malaria transmission mechanisms have a higher propensity to acquire bed nets. Relatedly, such understanding may also bear on their general ability to avoid exposure to mosquitos, thus decreasing their probability of infection Y.

For completeness, let me expand the narrative to capture the four remaining bi-directed confounding arcs. The arc between U and T could, for instance, obtain because the wealth level of agents' parents influences both agents' education level and their wealth, e.g. through support and inheritance. The second arc between U and X may capture the effect of parents' education level on their children's education level, as well as on children's propensity to obtain bed nets (for instance because more educated parents tend to perform better at teaching their children about the effectiveness of bed nets). Third, the arc between bed nets X and properly installed nets W may obtain because there is a common cause, agents' understanding of how bed nets work, that both increases their propensity to obtain nets as well as their propensity to properly install them once obtained. Finally, the confounding arc between X and Y may indicate agents' mosquito detection abilities. This ability may both increase their awareness of mosquito-induced health threats at a given level of exposure and hence increase their propensity to obtain nets, as well as decrease their propensity for infection, e.g. by inducing them to proactively kill mosquitos.

In this causal narrative, the selection nodes S and S' indicate that populations differ in the distribution of education and the distribution of mosquitos in the proximity of agents. For instance, agents in the target may be less well educated and live in a setting with naturally higher background concentration of mosquitos, thus increasing the prevalence of mosquitos in agents' proximity.

Both differences can be relevant obstacles to extrapolation. Nets may be effective in dealing with a certain amount of mosquitos, but since malaria infection is an all-or-nothing affair, a substantially higher background rate of mosquitos may make it disproportionally more likely that otherwise inconsequential faults of nets, such as imperfect installation or holes created through frequent use, become relevant threats to infection. Similarly, significant differences in education, wealth, and their associated effects on agents' propensity to properly install bed nets may significantly modify the effect of nets on malaria infection outcomes.

With this in mind, let us have a closer look at the quantities that B&P's transport formula instructs us to estimate in the target, i.e. $P^*(z|w)$ and $P^*(t)$. The latter, $P^*(t)$, will not pose problems. It is plausible to assume that wealth exhibits variation between agents and has some well-behaved distribution induced by background factors. $P^*(z|w)$, however, is more problematic. Recall that Z is mosquitos in the proximity of agents. We may assume that this quantity can be readily measured in some way, and there is no obstacle to this as Z, like T, will likely assume some well-behaved distribution induced by background factors that contribute to the presence of mosquitos. However, the transport formula demands that Z be measured *conditionally* on W, and this is where the problems begin. Recall that in predictive extrapolation cases, which are common in EBP applications, we assume that the intervention of interest has not yet been experienced in the target. In our example, this means that there are no bed nets in Π^*. So W will have a distribution with mean and variance zero, and hence $P^*(z|w) = P^*(z)$. This means that we are taking a measurement in a population where there are no installed bed nets (because there have never been any bednets to install) and hence, for lack of being installed, bed nets are ineffective at decreasing the number of mosquitos in the proximity of agents. This is, hopefully, different for the interventional distribution in Π^*, where $I_X^*(z|w) \neq P^*(z|w) = P^*(z)$, so at least some nets would be effective at preventing mosquitos from being in the proximity of agents if they were installed. But this interventional distribution can of course not be used, since it requires, by definition,

an intervention on X in the target, which would trivially fall prey to the extrapolator's bind. So we are stuck with an observational measure $P^*(z|w)$ that is supposed to capture how effective installed bed nets are at decreasing Z in the target, but we are measuring it in a population where there have never been any bed nets. This provides a, hopefully, wrong answer to our extrapolative query by telling us that the nets are entirely ineffective (although the co-intervention of antimalarial drugs may still be effective). With this measure figuring in our transport formula, the final quantity computed will hence, hopefully wrongly, indicate that the effect of distributing nets in the target will be zero, as the pre-post intervention distributions of Y in the target will only differ by the effect induced by antimalarial drugs.

There is a sense in which this still gives us a correct answer to an extrapolative query, but it is an answer that applies to the wrong target population. We are extrapolating to a target where installed nets are known to be ineffective at decreasing Z, indicated by $P^*(z|w)$. What is wrong about this extrapolation, however, is that it does not apply to the target we are interested in because it misses essential information about how effective bed nets would be in the target of actual interest, if there were any bed nets there. This information is needed, in the form of an informative observational measure of $P^*(z|w)$, to compute the correct post-intervention distribution of the outcome. But achieving this is precluded in predictive extrapolation for lack of relevant information in the data from the target, i.e. information pertaining to how effective installed bed nets are at decreasing Z.

Of course, we could impute $P^*(z|w)$ from the experimental population's distribution $P(z|w)$ or $I_X(z|w)$ as a good approximation. But then we would fail to account for the differences induced in W by upstream differences in U created by the selection node S'. This would hence, at least in part, amount to naïve extrapolation that glosses over at least some known causally relevant differences between populations, and would come on pain of potentially significant epistemic and practical costs as our predictions of the effects of $do(x)$ in the target are again likely to be off the mark.

Let me briefly draw some interim conclusions from this section before I move on to some more general remarks on outlook.

The above discussion makes two things clear. First, it is difficult to support assumptions about identity in parameters, functional form, and basic causal structure from observational or experimental data without falling prey to the extrapolator's bind.

This puts a spin on B&P's remarks that when "the target domain does not share any mechanism with its counterpart [...] the only way to achieve transportability is to identify R [the causal relation of interest] from scratch in the target population" (B&P 2014, 588). The above arguments suggest, however, that the very activity of asserting *whether* populations share mechanisms in the first place may *itself* require or trivially permit learning the causal relation or effect of interest from scratch in the target.

Moreover, even if such assumptions were somehow independently supported in a way that does not require measurement and comparison of parametric features in both populations, predictive extrapolation nevertheless poses distinct obstacles for graph-based extrapolation. When the intervention of interest has not yet been experienced in the target, the quantities that transport formulae require to be estimated in the target may lead to mistaken conclusions about the target quantity, or, to estimate them 'correctly' we may need to introduce the intervention of interest in the target, implying once again that we fall prey to the extrapolator's bind.

These concerns will, of course, not apply to all extrapolation scenarios. But it seems that the conditions that give rise to them are relatively common. All that is needed are some, potentially mild, differences between populations, and a case where the intervention has so far not been experienced in the target. That being said, it is important to recognize that the liability of B&P's approach to fall prey to the extrapolator's bind is likely to be context-dependent: pertinent data from and knowledge about the target are sometimes available, sometimes not; sometimes costly, sometimes cheap to obtain; and obtaining them will sometimes fall prey to the extrapolator's bind, but will not at other times. This suggests that more research is needed on the conditions under which the graph-based approach can provide genuinely ampliative, and hence successful extrapolation. This is beyond the scope of the current chapter, but the arguments provided here can serve as a fruitful basis for such investigations by making clear that predictive extrapolation, which is common in EBP applications, might often not allow graph-based extrapolation to be successful.

7.5 Conclusions

The graph-based approach to extrapolation is promising: it seems to help make causal assumptions explicit; offers a powerful analytic machinery to help answer a wide range of causal queries about a target from a combination of experimental and observational

data; helps substantially reduce measurement costs in both populations; and provides all of this while promising to help achieve successful extrapolation that both manages to overcome causally relevant differences between populations and evade the extrapolator's bind. These virtues make it seem like a silver bullet, and, indeed, B&P seem to think along these lines when they suggest that the challenges that the problem of extrapolation has "[…] been given a complete formal characterization and can thus be considered 'solved'" (B&P 2016, 7352).

In this chapter, I have provided two criticisms that cast doubt on this conclusion. The first is that the graph-based approach is importantly limited in the kinds of causally relevant differences between populations it can represent. Selection diagrams can capture some causally relevant differences between populations, but not all, and some can only be captured on pain of introducing too many selection variables, and thereby likely precluding transportability. At the same time, it is important to recognize that transportability is not the last word on whether extrapolation is feasible – it is not necessary for successful extrapolation. The presence of endogenous causally relevant differences that cannot be plausibly represented by selection diagrams does not imply that extrapolation fails, but only that B&P's approach is not helpful for extrapolation in these cases as it will not allow deriving transport formulae. As I have suggested, some cases where this happens may still be handled successfully without the graph-based approach, however.

The second line of criticism I offered argues that the graph-based approach involves various substantive, but unspoken, assumptions about causally relevant similarities between populations. Any sophisticated strategy for extrapolation, of course, needs to make some such assumptions, and potentially many. But there are different ways of doing so. Petersen and van der Laan (2014), for instance, distinguish between knowledge-based and convenience-based assumptions in DAGs. Currently, it seems that B&P treat many of the assumptions required by their approach as convenience-based. At worst, they are simply assumed to hold true. At best, unhelpful suggestions to use causal discovery methods are made (B&P 2012, 700, footnote 7). As I have argued, however, causal discovery methods are unlikely to provide support for the causal assumptions needed for graph-based extrapolation. This begins with simple concerns about extensive data requirements and ends with more principled concerns specific to predictive extrapolation, where probabilistic information that is pertinent to settling

issues of similarity and difference in causal structure is simply not available or, if available, uninformative. Others have, at least in part, recognized these problems as well. Consider Hyttinen et al., who remark that:

> […] significant parts of the causal literature regard the problem of identifying the causal effect *given* the causal structure as entirely separate from the problem of discovering the causal structure in the first place. For example, the entire literature on algorithms applying the do-calculus assumes — generally without further discussion — that the causal graph is known [references suppressed]. In the general model space that the do-calculus allows for, the causal structure can hardly ever be uniquely determined from the passive observational distribution or even from the experimental distributions that Bareinboim and Pearl (2012) consider. Still, the algorithms rely on being able to check complicated features of the causal structure. (2015, 399)

It is clear that *justified* extrapolation requires causal assumptions be knowledge-based, not convenience-based. Even B&P recognize this need when they assert that: "[i]f knowledge about commonalities and disparities is not available, transport across domains cannot, of course, be justified" (2016, 7351; footnote). However, an account of how to obtain causal knowledge to justify the assumptions required by graph-based extrapolation cannot be a mere afterthought, or just hoped to be delivered by some supplemental approach. The question of how to turn assumptions from mere assumptions into adequately supported ones is far from trivial, but needs to be addressed before the graph-based approach can be claimed to be able to solve any real-world problem of extrapolation.

Yet, so far, B&P remain silent on how to achieve this, and specifically about how to do so without falling prey to the extrapolator's bind. It is true that, at face value, transport formulae seem to substantially reduce measurement costs, by highlighting that not all causally relevant differences matter for successful extrapolation and the quantities to be measured in both populations can substantially differ from, and be less costly to obtain than, what one might intuitively suspect. However, as I have argued, the epistemic requirements of justified graph-based extrapolation do not end with the quantities demanded by a transport formula. They also include support for the structural and parametric assumptions that are needed for the construction of selection diagrams in the first place. So, in many cases, the apparent measurement cost advantages of the graph-based approach may be entirely eclipsed by the extensive empirical demands involved in supporting its assumptions. As I have argued, these measurements may not

only extend significantly beyond what a transport formula demands, but may also threaten to undermine successful extrapolation by instructing us to measure uninformative quantities, or making our extrapolations redundant by falling prey to the extrapolator's bind.

The conclusion to be drawn from this chapter is simple: It is not helpful to leave the processes by which we come to construct selection diagrams and estimate the quantities demanded by transport formulae shrouded in mystery, nor assume that they have somehow been completed. The crucial question for the graph-based strategy is this: is it possible to construct selection diagrams and measure informative quantities from the target demanded by transport formulae without falling prey to the extrapolator's bind? My answer is: at least not always, and possibly rarely. This puts important constraints on the applicability and general success of the graph-based extrapolation strategy, and on which kinds of efforts may or may not be undertaken in constructing selection diagrams and using transport formulae to compute causal effects.

The arguments developed here suggest that there remains a significant theoretical and methodological gap between transportability and successful extrapolation. Transportability seems neither sufficient nor necessary for successful extrapolation. The project of formally deriving licences to transport causal effects does not, and cannot clarify how to achieve adequately supported extrapolation, and hence extrapolation that promises to be successful. Licences, as in licences to draw action-guiding and informative conclusions about real-world targets, are not provided by formal results only. They require support, too, which, jointly with a transparent framework and clever algorithms telling us which assumptions are in need of support, can help us achieve adequately justified extrapolation. On its own, however, the graph-based approach falls well short of this aim. These issues are currently unhelpfully glossed over by B&P, and relegated to footnotes, but, as I have argued, need to be addressed explicitly.

In *Chapter 8* I will attempt to make some constructive proposals for how at least some of the challenges discussed so far might be addressed, and how graph-based extrapolation might still be useful, alongside other approaches, and supplemented by theoretical and empirical resources to provide the causal knowledge it requires to function. These proposals will leave the conclusions reached in this chapter largely untouched, but nevertheless provide some reasons to think that the graph-based approach can, under some conditions, usefully complement our efforts to go beyond

mere transportability and achieve successful extrapolation, albeit playing a less significant role than perhaps hoped by its proponents.

References

Aalen, O. O., K. Røysland, J. M. Gran, R. Kouyos, and T. Lange. (2016). "Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms", *Statistical Methods in Medical Research*, 25(5): 2294-2314.

Bareinboim, E. and J. Pearl. (2012). "Transportability of causal effects: Completeness results", In: Proceedings of the Twenty-Sixth Conference on Artificial Intelligence (AAAI-12), Menlo Park, CA.

— **(2013).** "A general algorithm for deciding transportability of experimental results", *Journal of Causal Inference*, 1: 107-134.

— **(2014).** "Transportability from multiple environments with limited experiments: Completeness results". In: Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger (eds.), pp. 280-88, Curran Associates.

— **(2016).** "Causal inference and the data-fusion problem", *Proceedings of the National Academy of Sciences*, 113: 7345-52.

Cartwright, N. D. (1989). *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.

— **(2013).** "Evidence, Argument and Prediction". In: V. Karakostas, and D. Dieks (eds.), EPSA11 Perspectives and Foundational Problems in Philosophy of Science, The European Philosophy of Science Association Proceedings. Cham, Switzerland: Springer International Publishing Switzerland.

Deaton, A., and N. D. Cartwright. (2018). "Reflections on Randomized Control Trials". *Social Science & Medicine,* 210: 86-90.

Elwert, F. (2013). "Graphical Causal Models". In: S.L. Morgan (ed.), *Handbook of Causal Analysis for Social Research*, pp. 245-73, Dordrecht: Springer.

Greenland, S., and B. Brumback. (2002). "An overview of relations among causal modelling methods", *International Journal of Epidemiology*, 31: 1030-37.

Hausman, D. M., R. Stern, and N. Weinberger. (2014), "Systems without a graphical causal representation". *Synthese,* 191(8): 1925-30.

Hausman, D., and J. Woodward. (1999). "Independence, Invariance and the Causal Markov Condition", *British Journal for the Philosophy of Science*, 50: 521-83.

Huang, Y., and M. Valtorta. (2006). "Identifiability in causal bayesian networks: A sound and complete algorithm". In: Proceedings of the Twenty-First National Conference on Artificial Intelligence, Menlo Park, CA, AAAI Press, 1149–54.

Huitfeld, A., S. A. Swanson, M. J. Stensrud, and E. Suzuki. (2016). "Effect Heterogeneity and Variable Selection for Standardizing Experimental Findings". Working paper, arXiv preprint arXiv:1610.00068v1.

Hyttinen, A., F. Eberhardt, and M. Järvisalo. (2015). "Do-calculus when the true graph is unknown", in: UAI'15 Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, 395-404.

Mooji, J. M., S. Magliacane, and T. Claassen. (2018). "Joint Causal Inference from Multiple Datasets", arXiv:1611.10351v3 [cs.LG].

Pearl, J. (1988). *Probabilistic Reasoning in Intelligence Systems*. San Mateo, CA: Morgan Kaufmann.

— **(1995).** "Causal diagrams for empirical research", *Biometrika*, 82: 669-710.

— **(2009).** *Causality: Models, Reasoning, and Inference*. 2nd edition. New York: Cambridge University Press.

— **(2014).** "The Deductive Approach to Causal Inference", *Journal of Causal Inference*, 2(2): 115–29.

— **(2015).** "Generalizing Experimental Findings", *Journal of Causal Inference*, 3(2): 259–66.

Pearl, J., and E. Bareinboim. (2011). "Transportability of causal and statistical relations: A formal approach". In: Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11), Menlo Park, CA.

— **(2014).** "External Validity: From Do-Calculus to Transportability Across Populations". *Statistical Science*, 29(4): 579–95.

Pearl, J., and T. Verma. (1991). "A theory of inferred causation". In: Allen, J.; Fikes, R.; and Sandewall, E. (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pp. 441– 452, San Mateo, CA: Morgan Kaufmann.

Petersen, M. L., and M.J. van der Laan. (2014). "Causal models and learning from data: integrating causal modeling and statistical estimation", *Epidemiology*, 25(3): 418-26.

Scheines, R. (1997). „An introduction to causal inference" In: McKim and Turner (eds.), *Causality in Crisis? Statistical Methods in the Search for Causal Knowledge in the Social Sciences*, pp. 185-99, Notre Dame, IN: University of Notre Dame Press.

Shpitser, I., and J. Pearl. (2006). "Identification of joint interventional distributions in recursive semi-markovian causal models". In: Proceedings of the Twenty-First National Conference on Artificial Intelligence, Menlo Park, CA, AAAI Press, 1219–26.

Spirtes, P., C.N. Glymour, and R. Scheines. (2000). *Causation, Prediction, and Search*. 2nd edition. Cambridge, MA: MIT Press.

Steel, D. (2010). "A New Approach to Argument by Analogy: Extrapolation and Chain Graphs". *Philosophy of Science*, 77(5): 1058-69.

— **(2013).** "Mechanisms and Extrapolation in the Abortion-Crime Controversy". In: Chao, Hsiang-Ke; et al. (eds.). *Mechanism and Causality in Biology and Economics*. pp. 185–206. Berlin/Heidelberg: Springer Science & Business Media.

Tian, J., and J. Pearl. (2003). "On the identification of causal effects". Technical Report R-290-L, Department of Computer Science, University of California, Los Angeles.

Verma, T., and J. Pearl. (1988). "Causal networks: Semantics and Expressiveness". In: Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence, Mountain View, CA, 352-59.

Vivalt, E. (2019). "How Much Can We Generalize from Impact Evaluations?". Unpublished manuscript, ANU, Canberra. Retrieved from: http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf (retrieved, Feb. 28, 2019)

Weinberg, C. (2007). "Can DAGs Clarify Effect Modification?", *Epidemiology*, 18(5): 569-72.

CHAPTER 8

Extrapolation – Where Next?

8.1 Introduction

Some authors have suggested that the problem of extrapolation has been 'solved' (B&P 2016, 7352; Marcellesi 2015, 1309), at least in the abstract. As discussed in *Chapter 4*, to underwrite this assessment, Marcellesi invokes the distinction between *abstract analyses* of the conditions under which extrapolation can be successful and *concrete methods* for extrapolation. But while it seems right to say that abstract analyses successfully clarify the general conditions under which causal effects can be extrapolated in principle, this achievement, by itself, does not imply that *successful* extrapolation is possible in any real-world context. According to my characterization of successful extrapolation offered in *Chapter 3*, real-world extrapolation is a problem that cannot be solved in the abstract alone, since it irreducibly involves overcoming the epistemic obstacles encountered in supporting whatever assumptions are required by any strategy for extrapolation, all while managing to steer clear of the extrapolator's bind. As I have argued throughout the preceding chapters, doing so is often not only difficult when empirical demands are extensive, but extrapolation can become a pyrrhic endeavour, where all too easily too much is asked in terms of what causal knowledge about the target we need to possess and we might sometimes be better off by not extrapolating at all, but attempting a different kind of inference to learn what we are interested in. In a nutshell, while offering abstract strategies for extrapolation is feasible, and sometimes straightforward, *successfully* extrapolating is often hard, and sometimes practically infeasible, even if abstract strategies (and some of their proponents) would like to suggest otherwise.

This is broadly coherent with more recently espoused views by Cartwright (forthcoming), which suggest that, to be successful, real-world extrapolation will often need to draw on so-called *middle-range theory* of the specific phenomena of interest and *programme theories* of the specific interventions of interest. On this view, one might also think that attempts to 'manualize' guidelines for evidence use and extrapolation are unhelpful, partly because too much of what is required to support an

extrapolation is tied to concrete extrapolation contexts, e.g. to particular local facts about populations and specific theories pertaining to the causal mechanisms that operate there, which together help us clarify issues of causally relevant similarities and differences. Yet, while I agree with Cartwright that contextual information and theory are important, it also seems possible to say at least some more general things about how to facilitate successful extrapolation, including on what roles theory can play and what to do when it remains unavailable and difficult to produce from scratch.

In this chapter, I aim to offer some suggestions to help us make progress on important practical epistemic problems that, so far, remain unaddressed by general and abstract strategies. I will aim to do so at an *intermediate* level, which is more specific than abstract strategies, but also more general than appeals to the importance of (potentially local) theory, which might not always be available. Perhaps unsurprisingly, the result of this discussion will not be a full-fledged alternative abstract strategy for extrapolation, nor a bottom-up empirical strategy that promises to evade the extrapolator's bind, but rather a series of suggestions for how to tackle some of the problems I have discussed so far.

Section 2 discusses different theoretical and empirical resources that could help underwrite extrapolative inference, and how they can interact in doing so. *Section 3* formulates several substantive desiderata for attractive (future) strategies for extrapolation. Here, I also comment on two general issues that this thesis has neglected so far. First, I make some suggestions for how reasoning about uncertainty and confidence in extrapolation could be improved by drawing on two existing frameworks developed in the philosophical literature. Second, I expand on several general suggestions for how EBP institutions could improve their methodological guidelines, specifically taking into account how diverse kinds of evidence might work together in supporting an extrapolation, as well as how different strategies for extrapolation perform under various conditions. *Section 4* summarizes how this chapter complements the contributions made in previous chapters.

8.2 The Extrapolator's Bind Revisited: Theory and Empirical Methods

The extrapolator's bind makes clear that successful extrapolation is not achieved by merely elaborating the abstract conditions under which a causal effect can be correctly predicted in the target in principle, or blindly following the epistemic demands imposed

by such conditions and supporting them by any available or feasibly producible background knowledge and supplementary evidence. The bind places further constraints on the success of extrapolative inference by demanding that the experimental result does not become (almost) irrelevant to our conclusion. If a significant part of an extrapolative inference is driven by a host of substantive assumptions, and the credibility of the conclusion hinges mostly on whether supplementary evidence and background knowledge for *these* assumptions can carry the required justificatory burden, then extrapolation becomes less and less successful the more one tries to justify it.

With this general insight in place, we can phrase the demands implied by the extrapolator's bind in a positive way. Successful extrapolation should be *ampliative* with respect to the supplementary evidence and background knowledge it requires. The extrapolative conclusion must extend beyond, preferably significantly, what could be inferred based on this supplementary material alone. This desideratum ensures that the experimental result remains relevant to our conclusion and that a key requirement for successful extrapolation is indeed satisfied.

How can this be achieved in practice? As outlined in *Chapter 3*, there are two distinct elements involved in supporting the empirical assumptions ***P*** required by strategies for extrapolation: supplementary empirical evidence ***S*** and background knowledge ***K***. It is important to note that the distinction between these is not always precise, as ***K*** can include both theoretical and empirical resources, and ranging from general in scope to highly specific, including pertinent local facts about, say, a particular feature of a causal mechanisms believed to govern an outcome of interest. Nevertheless, a useful distinction could be to say that background knowledge ***K*** prevalently encompasses existing resources, theoretical or empirical, which are pertinent to clarifying the validity of the assumptions required (by a strategy) for extrapolation; whereas supplementary empirical evidence ***S*** is typically not yet readily available before an extrapolation, and will need to be produced (including from pre-existing data) in the context of a specific extrapolation.

S and ***K*** can importantly interact, such as when background theory suggests a broad scaffolding covering both populations (e.g. by specifying important parts of a causal mechanism that is believed to be shared between them) and additional empirical evidence can fill in some remaining blanks (e.g. by clarifying the values of parameters

or the functional form of specific causal relationships). Yet, while the two will often be importantly intertwined in practice, it is still useful to keep them distinct, as, at least in principle, they can fully substitute one another. An extrapolation that proceeds in the absence of any prior theoretical or empirical resources will need to be supported by empirical evidence yet to be produced. Likewise, a rich theoretical background, such as strong theory asserting that experimental and target populations are sufficiently causally similar, full stop, may sometimes allow validating crucial extrapolation assumptions without acquiring much or any additional empirical evidence. Going forward, it hence seems useful to discuss these resources separately: we can place important constraints on what role background knowledge ***K*** can play and when it needs to be complemented by supplementary evidence. Likewise, recognizing that extrapolation will often need to rely on supplementary evidence ***S*** raises special questions about what evidence to produce, how to produce it, and what this implies for how effectiveness studies should be conducted. Let me address these issues in turn.

8.2.1 Background Knowledge and Theory

K so far has been largely a placeholder for a variety of theoretical and empirical resources[1] that might be helpful in supporting extrapolation. These include, but are not limited to, 'high theory' or general causal principles, 'middle-range theory', local licencing facts, and their interplay. 'High theory' or general causal principles, for instance, could be such things as the 'law of demand', where economists often (and often plausibly) assume that demand for a good decreases in its price (other things being equal), and that this holds generally across contexts. Middle-range theory, by contrast, is typically characterized as a type of (causal) theory that is more local in range and attempts to unify and explain a relatively small and specific range of phenomena (see Merton 1968[1949]), such as the consumption and saving behaviours of households of the rural poor; how a specific social norm governs the behaviours of individuals and when it does not (see e.g. Bicchieri et al. 2014); how and when economic agents are susceptible to 'loss aversion' (Kahneman and Tversky 1979; Gill et al. 2012; Gal et al. 2018), etc. Finally, local licencing facts are relatively immediately observable features of individuals or populations that can reliably indicate the presence or absence of certain

[1] It might also draw on *practical* resources, such as experience in implementing a certain kind of intervention and observing its effectiveness under different conditions.

causal features. For instance, in the spirit of process tracing, and following similar suggestions by Strevens (2007), there can be cases where overt observable features of individuals, units, or objects, such as the sociodemographic characteristics of a person, the species-membership of an animal, or the colour of a fruit, can afford important explanatory and predictive purchase on their behaviours. This works whenever the features are either closely and robustly correlated or are themselves causally produced, and preferably uniquely so, by a specific causal mechanism. If this is the case, then the presence of the feature of interest can be used as an inferential shortcut to help clarify issues of mechanistic similarity or difference. Such local facts can be helpful for extrapolation not just on their own but also in concert with theory, e.g. when they can be plugged into, as it were, available middle-range theory to make inferences about broader causal characteristics of a target. Here, rather than having to draw on 'thick' causal knowledge, i.e. knowledge specifically produced to clarify *whether*, say, Z plays a role in producing an X-Y-effect, we can draw on 'thin' causal knowledge: a licencing fact F is observed and background theory simply tells us *that* Z plays a role in an X-Y-effect whenever F is the case.

As an example, consider a case invoked by Cartwright and Hardie (2012) to illustrate some of the pitfalls to be expected in extrapolation. The case concerns an intervention (TINP) aimed at improving nutritional health in children. The intervention was first successfully tested in the Tamil Nadu province of India and later unsuccessfully exported to a novel context in Bangladesh (BINP). The reason invoked for the failure in Bangladesh was, in part, that, unlike in India, mothers in Bangladesh were often not in charge of procuring food and administering it to children. Rather, mothers-in-law often assumed this role. While the intervention in Bangladesh was still effective in increasing mothers' understanding of what promotes children's nutritional health, the different causal roles played by mothers in Bangladesh precluded the intervention from being effective there, as the knowledge and resources provided by the programme could not adequately translate into meaningful effects on children's nutritional health (White 2009).

Drawing on this and other, related cases, Cartwright and Hardie (2012) suggest that extrapolation can be greatly facilitated by two strategies, called *horizontal* and *vertical search*. These strategies are intimately tied to Cartwright's *effectiveness argument*, which demands that the causal principles governing the production of the outcome of

interest involve the intervention variable in both populations, as well as that the support factors necessary for the intervention to make a contribution to the outcome are distributed in the right way in the target. Horizontal search then means looking 'left and right' of the intervention variable to see what the necessary support factors are and whether they are instantiated in the target. Vertical search means investigating whether one has chosen the right level of abstraction to formulate the causal principle believed to govern the production of the outcome to then tell whether the intervention variable does indeed play the same causal role in both populations.

In the Tamil Nadu/Bangladesh case, the suggestion would be that vertical search could have helped formulate an adequately specified middle-range theory, which, complemented by local facts about both populations, might have allowed decision-makers to predict the failure of the intervention in Bangladesh or could have helped ensure its effectiveness by providing reasons to modify the intervention so as to target relevant agents, i.e. mothers-in-law instead of mothers. Specifically, a simplistic candidate for a middle-range theory here could say that distributing supplementary foods and nutritional health counselling to 'those who are in charge of distributing food to children' can be effective in increasing children's nutritional health, possibly conditional on a host of further background conditions. Together with local facts about which individuals in each population play the relevant social roles posited by the theory, the middle-range theory would have permitted successful prediction of causal effects, as well as perhaps successful intervention in Bangladesh.

Various further, and more detailed, examples for how middle-range theory, in interplay with local facts and previous expertise in implementing interventions, can greatly facilitate prediction and intervention can be found in the *realist evaluation* literature (e.g. Pawson and Tilley 1997; 2001; Astbury and Leeuw 2010; Pawson 2013). Here, there has traditionally been extensive emphasis on the role of theory, and, accordingly, effectiveness evaluation is supposed to begin with an explicit theory of how an intervention is envisioned to be effective (often called a 'programme theory', 'theory of change', or 'logic model'), proceeds to implementing the intervention, and ends with making suitable revisions to the theory, if needed, in light of the experiences made. This process, over time, is supposed to facilitate the development and refinement of middle-range theories of the phenomena targeted by policy interventions.

I agree with realist evaluators and Cartwright on the importance of theoretical resources for extrapolation. I also support their calls for incentivizing the production of middle-range theory, as well as for explicit and judicious *theorizing*, e.g. piecing available theories and local facts together, refining existing theory, and developing new theory tailored to specific use-cases.

Yet, while such calls are important, it is also crucial to think carefully about what we should do in cases where suitable theory is unavailable and difficult to produce. One concern to motivate this is to consider that developing middle-range theory often requires establishing empirically what the success of interventions eventually hinges on. So, very often, trial and error will be needed before the required theory becomes available. What is more, the timescales involved in how social mechanisms undergo structural change, as well as in trends determining what kinds of interventions are being studied in EBP, will put a natural upper bound on how developed middle-range theory can get. A sophisticated middle-range theory might take years to develop, and by the time it is available, fashions in EBP pertaining to what interventions are studied might have changed, and so might institutional and social backgrounds that importantly bear on agents' response to interventions. Promoting the development and use of middle-range theory is important, but hoping that this alone can help improve the practice of extrapolation seems too optimistic. What is equally needed are rough-and-ready empirical strategies that can be applied right here, right now, and that do not depend on the existence of theory that might be unavailable, and even if feasibly producible, might not be ready in time to advance our extrapolation efforts.

In what follows, I will highlight some empirical strategies which, based on the arguments developed in previous chapters, might be promising candidates for substantiating extrapolation assumptions in the absence of theory, while also providing resources to build such theory, and complementing it with important empirical information whenever available.

8.2.2 Empirical Strategies

How can we empirically learn about causally relevant similarities and differences between populations in a way that is practically feasible, efficient, and does not raise concerns about the extrapolator's bind? And how can we use such information to support crucial assumptions about causally relevant similarities?

In *Chapter 2*, I distinguished between different levels at which similarities and differences can obtain, i.e. variable distributions, structural parameters and functional form, and basic causal structure. It would seem interesting to organize the discussion of empirical methods around these levels. Unfortunately, this is difficult as empirical methods are faced with considerable ambiguities when investigating whether a difference in *effects* can be attributed to differences in some suspected *cause*. The problem here is that observing a difference in effects and associating it with a particular variable does not always tell us what the nature of the causally relevant difference with respect to that variable is, i.e. whether it is a difference in the distribution of the variable that matters, in whether and how this variable is involved in a causal mechanism, or whether it merely correlates with other such differences. Hence, with some exceptions, the methods to be highlighted here can rarely uniquely attribute causally relevant differences and similarities to one of the three levels identified in *Chapter 2*, and substantial uncertainty about what exactly is learned will often remain. With this caveat in mind, let me proceed to elaborate on some methods for detecting causally relevant *differences*, which can help inform us about where *similarities* might be important for underwriting extrapolation, as well as help support assumptions about such similarities.

The first step in clarifying whether populations are relevantly causally similar is to learn what things, beside the intervention itself, are involved in the production of an effect in the first place. Specifically, we need to identify relevant moderating and mediating variables, as well as support factors (*Chapter 4*, *Appendix 1*), that have the capacity to modify causal effects between populations (for terminological efficiency, I will simply call these variables *modifying variables* from now on). Identifying such variables helps clarify which variables might need to be involved in the same or similar ways in the causal mechanisms governing the outcomes of interest in both populations, i.e. with respect to functional form, structural parameters, and basic causal structure. As anticipated above, background theory can play important roles in suggesting which variables to look at. Yet, in the absence of theory, other means will be needed for discovery. Principally, there is a whole range of candidates here, including regression-based analyses performed on large-n observational datasets, collating data from multiple study populations (Allcott 2015; Dehejia et al. 2015), subgroup-analyses, factorial experiments, machine-learning-based approaches, qualitative comparative analysis (see e.g. Beach and Pedersen 2013; 2016; 2019), etc. Performing such investigations will usually not immediately yield a comprehensive understanding of all variables involved

in the production of an effect, but it is important to recognize that such an understanding, if it is ever to be developed, will not be achieved in one fell swoop, but in a piecemeal fashion. Let me expand on some specific methods in turn.

First, *subgroup analyses* can be useful for identifying significant differences in causal effects between subgroups of a population. These analyses typically proceed by splitting (post-intervention) the experimental population into subgroups according to observed characteristics, such as age, sex, and other potentially relevant features, and then estimating ATEs conditional on these subgroup characteristics. This can be a first, tentative step in learning what could be important modifying variables among the full gamut of candidates. It is tentative because it is well understood that subgroup analyses should be interpreted with caution, as they are liable to raise several methodological concerns. First, there are concerns about bias, e.g. where we might wrongly attribute causal significance to a variable that is merely correlated with a modifying variable, or the true causal significance of a modifying variable is masked by counteracting effects. Further, there are concerns about insufficient statistical power, where truly significant subgroup differences can remain undetected due to insufficient sample sizes. Finally, if subgroup analyses are conducted along a large range of candidate variables there are acute concerns about multiple hypothesis testing, which increases the likelihood of spuriously significant results. These concerns are acknowledged by many methodologists (see Varadhan and Seeger 2013; Fink et al. 2014) and it is often recommended that even highly significant subgroup differences should be taken only as reasons to form and pursue hypotheses about potentially important modifying variables, but not as sufficient by themselves for warranting definitive conclusions about what variables causally induce the estimated subgroup differences. Even so, subgroup analyses can be an easy and cheap-to-implement method to acquire information for subsequent investigations, as they can be performed on baseline data that can often be straightforwardly collected when conducting an RCT.

Another method, which that can evade several of the concerns leveled against subgroup analyses, is provided by *factorial experiments* (see Imai et al. 2011; 2013; see also Pearl 2011; 2014). In contrast to RCTs, factorial experiments test several interventions at once. For instance, if it is suspected that a variable $W\epsilon\{0; 1\}$ causally interacts with a treatment variable $X\epsilon\{0; 1\}$, factorial experiments allow us to estimate a parameter capturing this interaction in an unbiased way by randomly splitting the

experimental population into four groups and then assigning the following four treatment regimes to them: $[X = 1, W = 1]$; $[X = 1, W = 0]$; [X= $0, W = 1$]; and $[X = 0, W = 0]$. To the extent that assignment to groups and other methodological precautions successfully establish exogeneity of X *and* W, this allows unbiased estimation of several parameters at once, including the difference in the ATEs of X on Y conditional on $W = 1$ and $W = 0$ respectively, and provides an unbiased assessment of how differences in W causally induce differences in the magnitude of the *X*-*Y*-effect. Relatedly, factorial experiments can also aid in identifying mediating variables. For instance, Weinberger (2019) suggests that causal mediation analyses are helpful for separately identifying mediated and unmediated effects (also called *direct* and *indirect* effects; see e.g. Imai et al. 2011; 2013; see also Pearl 2011; 2014). Such analyses can be helpful for shedding light on what variables mediate the effects of interest in both populations, and whether important mediating pathways present in an experimental population might be disrupted in the target (determined by observational rather than experimental mediation analyses).

Of course, factorial experiments (as well as mediation analyses) are not without methodological problems either. First, they are clearly limited in scope. For one, there are features that cannot be easily manipulated by an investigator; think of psychological characteristics. These features present obstacles to unbiased identification of interaction effects, as inability to exogenously assign values to variables poses threats to the main identification assumption needed by the approach. Second, much like subgroup analyses, factorial trials may require significant expansions of sample size to ensure estimation of effects with the same precision as in a simple RCT that only measures the ATE of an intervention on *X* (Montgomery et al. 2003; Brittain and Wittes 1989). This suggests that factorial experiments are limited to settings where large populations are relatively easy to recruit and where there are only few important modifying variables to begin with. This puts constraints on the usefulness of factorial trials, as specifically the latter requirement is burdensome and highly restrictive. Here, it seems that without background knowledge that helps narrow down the list of candidate variables to investigate, one would need the very information that the factorial trial is supposed to produce prior to designing it. Yet, at least in some cases, it seems that atheoretical guidance for variable selection can be provided by subgroup analyses, suggesting that these analyses may be used fruitfully together with factorial trials, where the first can supply specific hypotheses and thereby constrain the menu of variables to look at, and

the second is used to further pursue such hypotheses in a way that is less prone to concerns about bias, causal misattribution, etc.

A further approach for identifying modifying variables comes from the machine learning literature, where algorithms such as CHAID have been proposed for the purpose of automated detection of interaction effects (Kass 1980). More recent proposals to use machine-learning methods can be found in the econometrics literature, where authors such as Athey and Imbens (2016; 2017) and Wager and Athey (2018) have proposed methods based on classification and regression tree (CART) algorithms. These algorithms use a brute-force approach to identify significant differences in CATEs in a (large-n) observational training sample, and then validate the learnt model by predicting outcomes of units from another sample partition that has not been used for model construction. According to Athey and Imbens (2016), this method can help identify modifying variables while evading several of the concerns about statistical power and multiple hypothesis testing that affect subgroup analyses. Like subgroup analyses, this method can be used on past trial data, at least to the extent that measurements of candidate variables have been obtained. At the same time, like with subgroup analyses, it is important to keep in mind that the interactions identified by machine learning approaches are statistical interactions, rather than causal ones. As such, they might often leave unclear whether one has identified a true modifying variable of an effect or rather a variable that is merely closely correlated with a true modifying variable.

Finally, there are also a range of qualitative methods and resources that might be useful for detecting potentially important modifying variables and other significant causal features that an extrapolation might need to consider. At least some of these resources could be useful even in the absence of strong background theory, including performing detailed case studies (see e.g. Crasnow 2011); considering expert knowledge; using anthropological and sociological methods such as interview-based methods and participatory observation emphasized in *Chapters 5* and *6*; and using qualitative comparative analysis (see e.g. Beach and Pedersen 2019). I will not expand on these methods in more detail here, as this might require a more extensive discussion of how they can proceed effectively in the absence of guiding theory. For now, it is enough to note that it might be promising to explore how such methods could complement the quantitative methods mentioned here.

The methods outlined above suggest that there are at least some promising bottom-up strategies that, even in the absence of strong background theory, might help generate evidence about modifying variables that could be useful for extrapolation. Of course, none of these methods provides a definitive, standalone solution for identifying modifying variables. Moreover, and perhaps more importantly, at least none of the quantitative methods is applicable if data on potential modifying variables are not available from both populations. This is already an acute concern at the stage of considering data availability from the experimental population, as in many cases effectiveness studies do not involve deliberate and extensive collection of data on potentially relevant modifying variables beyond a few generic baseline measurements. This suggests that widely circulated methodological guidelines specifying criteria for good evidence (e.g. as supplied by the What Works Clearinghouse, JPAL, the Campbell Collaboration, and others) should perhaps be extended to specifically encourage researchers involved in primary effectiveness studies to produce more extensive datasets that can be used for supplementary analyses. Any study where such data are not collected would seem to be a missed opportunity.

At the same time, it is important to keep in mind that even if such advice were widely followed, it will often only mean that suitable data are available from an experimental population, but not that corresponding data will also be available from the target. In line with the arguments developed in *Chapters 6* and *7*, data from the target are needed as well, however, to tell, for instance, whether a modifying variable in a study population also modifies the effects of interest in the target, and in the same way. At least in cases where mechanisms in the target are not 'dormant', and strong background knowledge and theory are unavailable, observational data from the target on potential modifying variables could be important for clarifying issues of causally relevant similarity and difference. But to play this role, such data must also be available.

Finally, it is important to stress again that what the methods outlined above have in common is that they have trouble disambiguating different forms of causally relevant differences, i.e. at the levels of variable distributions, parameters and functional form, and basic causal structure. The latter kinds of differences are particularly important, as it seems that differences at this level are more fundamental than others. Adjusting for differences in the distribution of a moderating variable is only useful if no differences at the level of basic causal structure obtain. Conversely, however, learning that two

populations differ importantly at this basic level would obviate the need to try and accommodate any differences at other levels. In light of this, it seems important to consider whether further empirical investigations can help clarify issues of similarity at this basic level first, i.e. before any such attempts at higher levels are made. Let me outline a tentative proposal for an empirical strategy that could help achieve this.

8.2.3 Comparing Causal Structures

Causal discovery methods (Spirtes et al. 2000; see also Eberhardt 2007; 2017 for an overview; see Chickering 2002 for an alternative approach), briefly discussed in *Chapter 7*, are methods that help us learn causal models from observational data by testing which potential causal models are consistent with the probabilistic independencies realized in the data. In a nutshell, this proceeds by drawing on the Causal Minimality, Causal Markov, and Faithfulness conditions, which are used to list all (Causal Markov) and only (Faithfulness) those probabilistic independence constraints implied by a causal graph. This list can then be compared with the probabilistic independence features exhibited by a dataset to disambiguate different possible models and constrain model selection to a smaller class of remaining models, called *Markov equivalence class*.

While these methods are largely used for discovering causal structures from data, my suggestion here is that they might also be useful for *comparing* causal structures between populations. Roughly, the idea is to use datasets from both populations to infer two sets of Markov equivalent models $\boldsymbol{M}$ and $\boldsymbol{M}^*$ (boldface indicates sets), one for each population, and then compare these model sets. If we were to find, say, that there is a significant overlap between $\boldsymbol{M}$ and $\boldsymbol{M}^*$, this could provide support for the assumption that two populations are relevantly similar at the level of basic causal structure.

A broadly related proposal, though not aimed at extrapolation, has recently been made by Eva et al. (forthcoming), who propose a series of metrics to express the similarity between two causal graphs. One of these metrics concerns *evidential similarity*, i.e. similarities in the probabilistic independence implications derived from two graphs G and G^*. Eva et al. proceed to develop this and other proposals in some detail, which I will not expand on here. For the present purposes it is enough to note that, based on the idea of evidential similarity, we might develop an additional

instrument that could help us learn, potentially in an atheoretic way, whether two populations might be similar at the level of basic causal structure. Let me expand.

Assume a set of measured variables $\boldsymbol{V}$ and two measured probability distributions over $\boldsymbol{V}$, $P(v)$ and $P^*(v)$, for the experimental and target population respectively. Based on this, we can list two sets $\boldsymbol{C}$ and $\boldsymbol{C}^*$ of probabilistic independence features that are exhibited by the data from each population. We can then use the identification conditions put forward by Spirtes et al. (2000) (or alternative conditions) to learn two sets of Markov equivalent causal models $\boldsymbol{M}$ and $\boldsymbol{M}^*$. Rather than directly comparing $\boldsymbol{C}$ and $\boldsymbol{C}^*$, the benefit of taking the detour through causal models is that the models can be inspected visually in a relatively straightforward fashion, which allows investigators to disambiguate them beyond what is afforded by the data, e.g. with the help of existing pre-theoretical intuition as well as, if available, background knowledge and theory. This can be helpful to further narrow down the sets $\boldsymbol{M}$ and $\boldsymbol{M}^*$ before any comparisons are made.

With $\boldsymbol{M}$ and $\boldsymbol{M}^*$ in place, we can then perform comparisons between them. For instance, we can express the *overlap* O between $\boldsymbol{M}$ and $\boldsymbol{M}^*$ as the proportion of their overlapping and non-overlapping parts. Formally,

$$O(\boldsymbol{M}, \boldsymbol{M}^*) = |\boldsymbol{M} \cap \boldsymbol{M}^*| / (|\boldsymbol{M} \cup \boldsymbol{M}^*| - |\boldsymbol{M} \cap \boldsymbol{M}^*|).$$

If O is larger, other things being equal, this might make us more confident that the two populations are similar at the level of causal structure.

What is more, similarity inferences do not always need to be undertaken concerning the full graphs contained in $\boldsymbol{M}$ and $\boldsymbol{M}^*$. We may additionally perform comparisons at the level of subgraphs included in the graphs in $\boldsymbol{M}$ and $\boldsymbol{M}^*$. Despite substantial uncertainties at the level of $\boldsymbol{M}$ and $\boldsymbol{M}^*$**,** we might learn, for instance, that all graphs in $\boldsymbol{M}$ and $\boldsymbol{M}^*$ share a certain subset of features. For instance, if all graphs in $\boldsymbol{M}$ and $\boldsymbol{M}^*$ exhibit a directed arrow $X \rightarrow W$, this can significantly increase our confidence that the mechanisms in both populations share this causal feature. Comparisons of the overall similarity of $\boldsymbol{M}$ and $\boldsymbol{M}^*$could then proceed in a bottom-up fashion, where our confidence in the similarity of overall causal structure can be expressed in terms of aggregating our confidence in the similarity of partial causal structures. Here, parts of the potentially shared overall graph structure can enjoy more confidence, with certainty of similarity and dissimilarity being the upper limit, and various degrees of uncertainty

could be expressed, including graphically, in a 'similarity-confidence map' over a union of the graph skeletons from $\boldsymbol{M}$ and $\boldsymbol{M}^*$**.**

Of course, in practice, substantial uncertainty will often remain, especially when $\boldsymbol{M}$ and $\boldsymbol{M}^*$ have large cardinality. This is because any of the Markov equivalent models in $\boldsymbol{M}$ and $\boldsymbol{M}^*$ could be the true model for the respective population. Even when $\boldsymbol{M}$ and $\boldsymbol{M}^*$ contain exactly the same models this would still be far from guaranteeing that populations are identical with respect to their mechanisms. It would merely make it *possible* that they are. Of course, as $\boldsymbol{M}$ and $\boldsymbol{M}^*$ become (significantly) smaller, potentially aided by ruling out models with the help of background knowledge and theory, we might approach cases where it becomes at least somewhat likely that populations are similar or identical. Further, equipped with $\boldsymbol{M}$ and $\boldsymbol{M}^*$**,** additional investigations can be performed, including collecting larger datasets, particularly time-series data, or performing additional interventions that can help disambiguate further (potentially decisively) between the remaining models (see Tong and Koller 2001; Murphy 2001; Eberhardt 2012; Hyttinen et al. 2013). All of this, of course, will need to proceed with a view towards evading the extrapolator's bind.[2]

It is important to note that, like the methods outlined earlier, the applicability of the method suggested here will greatly depend on the availability of large observational datasets, including perhaps time-series data, as well as, for intervention-based disambiguation efforts, the ability to intervene in both populations.

What is more, as elaborated in *Chapters 5, 6* and *7*, observation-based approaches to causal discovery will require that the mechanisms in both populations can be observed 'in action'. 'Dormant' mechanisms cannot write characteristic signatures of their features, including probabilistic independencies, into data that we might wish to use.

Despite the uncertainties that will often remain in practice, the method for causal structure comparison suggested here could play an important role to support basic assumptions involved in real-world extrapolation. Positively, if $\boldsymbol{M}$ and $\boldsymbol{M}^*$ enjoy significant overlap, and there is a set of subgraphs shared by all graphs in $\boldsymbol{M}$ and $\boldsymbol{M}^*$ (or an important subset), then this can increase our confidence in the similarity of populations at the level of causal structure and, in the limit, provide strong support for the similarity or identity of *partial* causal structures. Negatively, if two populations

[2] See e.g. Kocaoglu et al. (2017) for a method that allows formulating cost-constraints on potential interventions which could help automated search for optimal interventions avoid the extrapolator's bind.

share few, if any, graphs and subgraphs, this can also be highly informative, since it will often obviate the need to address further questions about similarity and difference at other levels of parameters, functional form, and variable distributions. If suitable data are available (or can be acquired at reasonable cost), it hence seems that causal discovery-based analyses can be an important first test to apply before further efforts to underwrite extrapolative inference are made. The details on the method suggested here, including test-cases, will of course need to be developed more fully before this proposal can be taken seriously. For now, the sketch provided here should merely be taken as a suggestion for a potentially interesting avenue for future research.

8.2.4 Putting the Pieces Together: Interactions Between Theory and Evidence

With the role of different theoretical and empirical resources for underwriting extrapolation clarified, it is important to say something about how we can use these resources most effectively, specifically with a view towards evading the extrapolator's bind.

Empirical methods for learning about heterogeneous causal effects and attributing heterogeneity to variables are an important resource. They have the principled ability to tell us what happens, causally, in the specific experimental and target populations of interest. Such methods hence promise that we do not always have to rely entirely on theory that might often be imperfectly informative about such matters and that will often leave considerable uncertainty as to whether a novel target is indeed covered by the theory. At the same time, various methodological concerns outlined above suggest that without prior theoretical guidance as to what variables to consider, on-the-spot empirical attempts to investigate issues of similarity and difference are highly limited. Concerns about data availability and the extrapolator's bind only add to these limitations. It hence seems unlikely that the methods sketched out above can individually, or in any combination, regularly provide sufficient support for extrapolation efforts from start to finish by themselves.

Good theory can undoubtedly be an important resource for underwriting extrapolation, too, but is not always available, and for it to become available it needs to be developed, potentially with the aid of some of the empirical methods outlined above. What is more, even highly developed middle-range theory will often not be able to get us all the way to an envisioned extrapolative conclusion. For instance, it might be able

to tell us that microcredit availability can increase consumption of durable goods by allowing agents to make lumpy investments and start small businesses, and that prior entrepreneurial experience can positively moderate this relationship. But it will, by itself, typically be unable to tell us, for instance, whether the marginal effect of microcredit availability will be quantitatively similar in a novel population as in those studied so far. So while theory, on its own, can be helpful for addressing some extrapolative queries (e.g. qualitative ones), it will often fall short of providing the support needed for more sophisticated extrapolation attempts, e.g. those involving quantitative queries.

Likewise, theory will also often be unable to tell us whether a specific causal relationship is suitably instantiated in a target, or rather severed. It might be able to tell us what happens *if* this relationship is or is not instantiated, such as when informing us that an intervention providing nutritional counselling and supplemental foods cannot be effective if it does not target 'those in charge of distributing food to children'. However, theory will often not be able to clarify on its own *whether* a certain relationship is or is not instantiated in a target, e.g. whether specific agents in the target satisfy the description of 'those being in charge of distributing food to children', so some empirical knowledge from the target will still be needed.

Hence, in many cases, theory and empirical methods will need to work together in underwriting extrapolation. What is more, it also seems that different kinds of empirical evidence, e.g. qualitative and quantitative, may fruitfully cooperate in cases where neither would be sufficient by itself to infer the envisioned conclusion. As suggested in *Chapter 6*, putting different ingredients together in the right ways can be successful in some cases, such as when an intervention-specific middle-range theory of how agents interact with bed nets can tell us that distributing nets to agents who experience strong incentives to use them for fishing is unlikely to help reduce malaria infection. Whether agents in a particular population do have such incentives may then be identified 'from a distance', such as when it is clear from observation that agents in a specific region largely depend on fishing for food and income, or from 'up close' when inquiring with agents what they would do if they were given bed nets.

Yet, there will also be cases where such productive interaction of background theory and supplementary empirical evidence is severely inhibited by basic features of the problems of extrapolation being targeted. Specifically, the distinction between

attributive and predictive extrapolation introduced in *Chapter 5* suggests that some problems of extrapolation are significantly more difficult to overcome than others. Let me briefly revisit, reflect more generally on, and slightly refine this distinction in order to extract some general recommendations for how it can help identify problematic cases of extrapolation.

8.2.5 When Mechanisms Sleep: Attributive and Predictive Extrapolation Revisited

In *Chapter 5*, I have argued that beyond the different kinds of problems of extrapolation and extrapolative inference outlined in *Chapter 2*, there is a further, orthogonal distinction between two general kinds of extrapolation, i.e. attributive and predictive. I have also clarified that the importance of this distinction lies in the kind of *evidence* typically available for supporting extrapolation in each setting and how informative this evidence can be on issues of causally relevant similarity and difference. Specifically, while attributive extrapolation allows us to observe the effects of interest and the mechanisms governing these effects 'in action' in the target, predictive extrapolation often does not permit this. This has important ramifications. If a mechanism cannot be observed 'in action', and has so far remained largely 'dormant', this makes it unlikely that we can reliably identify features of the causal mechanisms at issue.

Here, I want to offer some further refinements on this distinction. Specifically, so far, my discussion has treated the distinction as a dichotomy. This was sufficient for the arguments developed earlier. However, a broader outlook that is informative for guiding real-world extrapolation efforts should be refined to recognize that the distinction is one of degree and not of kind.

As the discussion in *Chapter 5* and *6* suggests, there are cases of predictive extrapolation where some information on causal mechanisms can still be obtained from observational and other methods. In *Chapter 6* I have argued that despite an intervention's effects not having been experienced and observed in the target so far, it might still be possible to observe the effects of related past interventions, or other past exogenous changes in the variables that figure in the mechanisms that are believed to govern the effects of interest. At least sometimes, such observations can give us some purchase on questions regarding features of relevant causal mechanisms in the target. However, the extent to which this is informative is a matter of degree.

Specifically, when using quantitative observational data pertaining to past interventions or exogenous changes in the target, we need to assume that the intervention of present interest is not structure-altering and, further, that the past intervention or exogenous change is sufficiently similar to the intervention of current interest, particularly that their effects are mediated by the same pathways and in the same ways. These assumptions will arguably be stronger in some cases than others (not all interventions are plausibly structure-altering), and supporting them can sometimes be easier than at other times. For instance, learning that a past public health awareness campaign has increased agents' understanding of health issue *A* can increase our confidence that related, similar aspects of our current efforts to decrease HIV infection rates might be similarly effective. By contrast, learning that agents have refrained from increasing their investments in durable goods after a lump-sum unconditional cash transfer in the past would not seem to justifiably increase our confidence that they will also refrain from doing so in response to microfinance products becoming available, even if both effects would plausibly transmit through largely the same pathways, including household income/endowment and other variables related to agents' decision-making pertaining to consumption and saving. What this suggests is that the extent to which quantitative observational evidence about past interventions or exogenous changes is available, and is indeed informative on our questions about causal mechanisms in the target, can be understood as a matter of degree.

Similarly, if one considers qualitative evidence produced by methods such as interviews and participatory observation, this can be less or more informative for answering questions about features of the causal mechanisms in the target depending on the extent to which agents have already experienced similar changes in the variables believed to be involved in governing the effects of interest, as well as other features that matter for the reliability of their self-reports and inferences based on observing their behaviours. Again, this suggests that the extent to which qualitative methods can inform us about causal mechanisms in the target is a matter of degree.

Moreover, at the level of mechanisms, we can say that besides the two extremes of 'dormant' mechanisms, whose important constituent parts have not, so far, expressed their causal powers in ways that allow observation to be informative about their characteristics, and 'live' or 'awake' mechanisms, there can be in-between states where, say, parts of a mechanism have been 'dormant', while others have, perhaps regularly,

been sufficiently 'active' to allow both quantitative and qualitative observational methods to get a handle on their details.

Finally, it is also important to recognize that 'dormant' mechanisms can not only constrain the usefulness of supplementary empirical evidence from the target, but can also hamper the usefulness of theory. If mechanisms in a target have so far remained 'dormant', this can make it more difficult to tell whether a background theory indeed applies to a specific target. For us to be confident that a theory applies to a novel context, we need to have criteria for applicability in place that allow us to determine from relatively straightforward observation whether a certain mechanism is instantiated in a population, and hence whether our theory will apply there. Yet, when such *labelling features*, as we might call them, are causally produced by a mechanism (as per Strevens 2007) this suggests that 'dormant' mechanisms might make such features invisible to us. Pace Strevens, agents are often not like lemons, so we cannot rely too much on the idea that labelling features can be observed 'from a distance'. Moreover, even if labelling features do reliably manifest when mechanisms are 'dormant', building sufficiently sophisticated theory that recognizes such features will often require an extensive history of successful and unsuccessful applications over a broad variety of settings, as well as suitable refinements being made to the theory itself and its (learned) conditions of applicability. So before strong background theory can help us achieve successful extrapolation, plenty of unsuccessful extrapolation will often need to take place, both in the form of inaccurate extrapolation that (potentially) helps us revise our theories, as well as in the form of extrapolation that falls prey to the extrapolator's bind (i.e. learning the effect of interest in the target to see if the target is indeed covered by our theory), before the conditions of applicability are suitably refined to allow more successful forms of extrapolation in future instances.

Going forward, and with a view to practice, it hence seems reasonable to suggest that analysts and users of evidence involved in extrapolation should think about the *degree* to which the mechanisms in the target have so far been 'awake', and would, ideally, permit quantitative and qualitative observational methods to help with settling issues of causally relevant similarity and difference. Intuitively, the more 'awake' a mechanism can be believed to be in the target, the more likely it is that pertinent evidence that bears on issues of mechanistic similarity and difference can be obtained. As the HIV prevention example from *Chapter 6* suggests, a good proxy for how 'awake' a

mechanism is, is to think about whether agents have prior experience with, and understanding of, the sorts of changes that an intervention is supposed to induce.

At the most general level, then, the attributive/predictive spectrum can provide us with an interesting atheoretical instrument that can 1) help us tell how likely it is that suitable information about causal mechanisms can be obtained from a target and 2) inform our confidence about whether available theory is likely to reliably speak to issues of causally relevant similarity and difference. Most importantly, predictive and attributive extrapolation (and gradual variants) can be distinguished *before* endeavouring to make an inference, and can hence allow us to anticipate whether specific instances of extrapolation might be considerably more difficult to handle than one would hope. The distinction can hence help analysts form more realistic expectations as to what is feasible in a concrete context, as well as allocate their efforts towards those strategies for learning about similarity and difference that seem most promising for the case at hand, or, in the limit, suggest that an altogether different approach is needed to learn what they are interested in.

8.2.6 Experimental Design, Sampling, and Overlapping Support

As the discussion of 'dormant' mechanisms suggests, it is important to observe mechanisms 'in action' in order to be able to infer something about their features. This is not only important when thinking about how to learn something about a target; it is important for experimental design as well. This is best appreciated in cases where heterogeneity in mechanisms manifests not only at the level of populations, but at the level of individuals, e.g. where X is causally relevant for Y in one way for individual i and in a different way for individual j, such as when a microfinance product increases household welfare by helping some agents make lumpy household investments (e.g. buying an energy-efficient cooking stove) and others (those with suitable opportunities and abilities) to start a small business. As argued in *Chapter 2*, such individual-level heterogeneity can apply both within and across domains and can hence be an important obstacle to successful extrapolation. In cases where such heterogeneity obtains, the likelihood of successful extrapolation can be greatly increased by ensuring that relevant variation in mechanisms is already *represented* in the experimental population, e.g. that a study of microfinance effects includes both individuals who would be likely to use microfinance loans to make lumpy household investments and those who would use

them to start a business. Here, *deliberate sampling* of individuals (cf. Cook and Campbell 1979, 75; Shadish et al. 2002) with both mechanism-types helps capture relevant mechanistic heterogeneity already in a trial population, thus allowing us to observe both mechanisms 'in action' there, and hence making it easier to make inferences about novel populations where either of these mechanisms are prevalent or where they are distributed in different proportions than in the experiment.

Let me expand on some ideas on experimental design and sampling, by first briefly reiterating some standard intuitions, before refining them with the help of the theoretical resources afforded by interactive covariate-based approaches.

A standard intuition, discussed briefly in *Chapters 2* and *3*, and occasionally mentioned but not further elaborated in some EBP methodological guidelines, is that the more similar an experimental population is to a target, the more reliable extrapolative inference will be (see e.g. Campbell 1986). Call this the *similarity thesis*. In terms of designing experiments, this thesis is sometimes taken to suggest that trial populations should be recruited in such a way that they are maximally similar to eventual target populations (assuming, of course, that one already knows what these targets are). For instance, when one suspects potentially important subgroup differences in a causal effect between women and men, and a target population consists of both women and men, one should try to recruit both into a trial. Similarly, if one suspects that implementers in a trial are significantly more skilled in implementing an intervention than those in a potential target, then one should make the trial more representative of actual implementation practice by involving implementers that are similarly skilled to those one believes would be involved in implementing the intervention when rolled out to novel settings.

Here, I want to draw on the *overlapping support assumption* involved in the interactive covariate-based approach by Hotz et al. (2005, 247; see also Muller 2015, 5, who coined the term) to refine these intuitions. In short, this assumption says that for any value of an interactive covariate W that is suspected to modify the effects of interest, there needs to be a non-zero probability for individuals in the experimental population to exhibit this value. As suggested in *Chapter 6*, at least in principle, this can be extended to cases where W is not a moderating or mediating variable that modifies an effect, but a proxy variable that, for instance, captures different mechanism-types prevalent in a population. Together with further assumptions outlined in *Chapter 6*, this

allows that even if experimental and target populations do not exhibit the same or similar distributions of W, we can still reweight CATEs to obtain a correct expectation of the effect of interest in the target.

The overlapping support assumption has important consequences for experimental design: it is not necessary to make trial populations as similar as possible to eventual targets. All that is required is that there is overlap in the causal features exhibited by both populations.

What is more, contra the similarity thesis, making experimental populations as similar to targets as possible is not only unnecessary, but can also be counterproductive. In many cases, making a trial population more similar to an envisioned target might involve not only permitting more variation in modifying variables, structural parameters, and basic structural features of mechanisms, but also in variables that merely co-determine the outcome, without interacting with the intervention of interest. This can make inference messy, as admitting such variation will generally induce more variance in outcome variables, which, other things being equal, makes it more difficult to detect causal effects, including CATEs. The overlapping support assumption can thus provide important guidance for experimental design: it tells us that what is important is not to make the experimental population as similar as possible to a target, which is especially difficult if potential targets are unknown at the time of design, but to sample widely enough to include causal features that might otherwise remain idiosyncratic to potential target populations. Conversely, if the latter fails, then this makes extrapolation significantly more difficult, as it is not guided by data that include these causal features as part of mechanisms that are rendered 'active' by an intervention.

Ensuring that there is overlapping support in the distributions of features that can modify a causal effect is not only helpful for interactive covariate-based strategies for extrapolation, but for other strategies as well. In favourable cases, where background knowledge and theory is available to guide our sampling efforts, following the desideratum of achieving overlapping support can also help with applying other strategies for extrapolation. At the same time, it is important to recognize that overlapping support by itself is neither easy to achieve if one does not know what to sample for, nor does it, by itself, strongly facilitate successful extrapolation – it should merely be viewed as one further measure that can be undertaken, if applicable, to make it more likely that successful extrapolation can be achieved.

With these suggestions pertaining to the roles and limitations of supplementary resources $\boldsymbol{S}$ and $\boldsymbol{K}$ in place, let me proceed to discuss some more general desiderata that can help facilitate successful extrapolation.

8.3 Desiderata for Extrapolation

In previous chapters I have discussed, at length, how existing strategies fail the desiderata for successful extrapolation outlined in *Chapter 3*, i.e. extrapolation that ensures that 1) an extrapolative conclusion is relevantly informed by an experimental result, 2) the conclusion is adequately justified, 3) it is accurate (and precise), and 4) it answers our query. Going forward, it seems that we can now formulate some additional desiderata for what a future strategy for extrapolation should, ideally, be able to do, either as a standalone strategy or as accommodated in a more general framework for extrapolation.

1) It should not have high *epistemic entry barriers*. The discussion of B&P's graph-based approach in *Chapter 7* usefully illustrated the importance of this desideratum. On B&P's approach the correct graph structure for both populations needs to be known before extrapolation can even proceed to considering what causally relevant differences can be adjusted for or 'conditioned away'. Such knowledge is needed to construct a shared causal graph G' on the basis of which a selection diagram D can then be constructed to capture causally relevant differences. Surely, confidence in the assumption that two populations indeed share a causal diagram G' can come in degrees. However, it seems that from the perspective of licencing inferences this assumption is an all-or-nothing *package deal.* Either it is true, in which case one can proceed to encode causally relevant differences in a selection diagram D and derive transport formulae, or it is not, in which case extrapolation terminates. But the approach does not seem to permit (or aim to do so) interesting ways in which it could be partly true, and inference could still proceed successfully. So while, at the epistemic level of underwriting extrapolation, it might be possible to further subdivide the general assumption of populations sharing a causal diagram G' into smaller, more manageable *component assumptions* pertaining to similarities involving specific variables and relationships (e.g. by requiring identity in some subgraphs or individual causal relationships), the construction of a selection diagram as per B&P's account still eventually requires a full overlap of the graphs G and G^*, meaning that

similarities/identities on the level of individual causal relationships or subgraphs are only useful if they indeed add up to full overlap of G and G^*. Yet, as I have argued, especially in predictive extrapolation, it will often be exceedingly difficult to validate assumptions about similarity/identity at the level of basic causal structure, particularly when pertaining to causal relationships that constitute the main causal pathways between intervention and outcome variables, which are arguably the most important parts of a graph for purposes of extrapolation. By contrast, Cartwright's effectiveness argument and Steel's mechanism-based approach, although at the cost of limited inferences, come with substantially milder assumptions.

In light of this, it seems reasonable to think that a strategy for extrapolation should be flexible concerning what inferences it permits and responsive to what causal knowledge is available. It would start with milder assumptions, which constitute epistemically more manageable entry barriers, and permit an appropriately limited set of inferences, before proceeding to permit more ambitious inferences whenever more knowledge is or becomes available. This way, successful extrapolation is not tied to high entry barriers, which might induce a cautious analyst to give up before trying to satisfy the extensive demands imposed by, say, the requirement to learn the correct causal graphs for two populations.

2) An attractive strategy for extrapolation should facilitate *incremental* and *cumulative* learning about causally relevant similarities and differences. This desideratum is intimately related to the concern about 'package-deal' assumptions. Incremental and cumulative learning means that the process of supporting substantive extrapolation assumptions should be such that each increment of supplementary evidence $\boldsymbol{S}$ and background knowledge $\boldsymbol{K}$ can bear individually on the validity of *component assumptions*, i.e. assumptions pertaining to similarities and differences in specific respects, such as individual variables and causal relationships. Here, both $\boldsymbol{S}$ and $\boldsymbol{K}$ should be able to change our confidence in a specific component assumption (e.g. pertaining to a specific causal relationship) and confidence in more encompassing, general assumptions (say about similarities between mechanisms more generally) should be a function of the confidence in their components. Moreover, each token of $\boldsymbol{S}$ and $\boldsymbol{K}$ should also be informative for our ultimate extrapolative conclusion $\boldsymbol{C}$, i.e. it should be able to change either the substantive content of our conclusion, the accuracy or the precision of our conclusion, or our confidence in that conclusion, by adding to the

evidential weight in its favour. Conversely, it should not be the case that we have to acquire a vast collection of ***S*** and ***K*** before we can obtain *any* prediction and express how confident we are in it.

3) An attractive strategy should also be pluralistic concerning the kinds of supplementary evidence and background knowledge it accepts for underwriting extrapolation. For one, if desired, it should accept all relevant evidence that is available (see e.g. the *principle of total evidence* in Carnap 1947; Good 1967). Following the arguments provided in previous chapters as well as in *Section 2*, it seems desirable that different kinds of supplementary evidence and background knowledge can be *integrated* to bear on an all-things-considered assessment of how likely a specific component assumption is to be satisfied. This is particularly relevant in light of concerns levelled about interactive covariate-based approaches, and extrapolation in econometrics more generally, which, as argued in *Chapter 6*, seems to prioritize quantitative observational evidence. There, I have argued that predictive extrapolation problems will often require considering other kinds of evidence, including, importantly, qualitative evidence of various kinds. More specific suggestions for frameworks to integrate diverse kinds of evidence will be made shortly.

4) Building on the arguments characterizing extrapolation as highly heterogeneous outlined in *Chapter 2*, an attractive strategy for extrapolation should be *context-sensitive*, i.e. it should permit various kinds of features of concrete extrapolation contexts to inform how extrapolation can and should proceed. As suggested above, this includes important differences in the envisioned kind of extrapolative conclusion, e.g. qualitative, quantitative, and others outlined in C*hapter 2*. Such differences should inform what kinds of assumptions are needed, and what supplementary evidence ***S*** and background knowledge ***K*** are required to support them. But it also extends to important differences in the resources available for underwriting an extrapolation, including those induced by time constraints, limited availability of theoretical and empirical support, as well as limited computational and analytic capabilities on the part of extrapolators. Here, analysts should be guided to select the goals that are realistically achievable, given their constraints, and presented with concrete pathways to extrapolation that are suitable for reaching these goals. Moreover, an attractive strategy for extrapolation should also be responsive to the desired fidelity of the inference, which in turn hinges, among other things, on the real-world non-epistemic risks involved in extrapolation,

e.g. wrongly accepting a prediction that turns out to be inaccurate, failing to accept a prediction that would have been accurate, and so forth. What is more, it also seems plausible to think that what, exactly, it means to fall prey to or avoid the extrapolator's bind can be a context-sensitive matter (e.g. depending on the generality of the conclusion), so this, too, should be acknowledged. Finally, whenever appropriate, an attractive strategy should also recognize important contextual information about the envisioned or plausible target(s) of extrapolation(s) from studies yet to be conducted and make suitable recommendations for how to design studies in a way that facilitates successful extrapolation, e.g. by guiding experimenters in producing suitable supplementary empirical evidence in accordance with the suggestions made in *Section 2*.

5) An attractive extrapolation strategy should also be able to tell us, as explicitly as possible, what assumptions about causally relevant similarity and difference are required. This desideratum is met to different degrees by the strategies examined in previous chapters, but all of them involve at least some unspoken assumptions as well. B&P's graph-based approach seems to come closest to satisfying this desideratum as its explicit graph-based representation allows analysts, once suitably trained in using the approach, to 'read off' the majority of substantive causal assumptions from graphs. Yet, as I have argued, even B&P's approach involves unspoken assumptions, such as when the way in which modifying variables bear on causal effect magnitudes needs to be parametrically identical between populations, as well as substantive ambiguities about what, exactly, needs to be assumed to construct a selection diagram. There is hence room for improvement on the part of all strategies examined so far, and the present work aims to contribute to making clearer what assumptions they require and how demanding they are. It would also seem helpful to recognize these demands more broadly, including in widely circulated methodological guidelines for using effectiveness evidence. I will say more on this shortly.

6) It should also be able to tell us, if desired, how causally relevant differences between populations bear on differences in the quality or magnitudes of causal effects to be extrapolated. Here, interactive covariate-based strategies and B&P's graph-based approach have a clear advantage, but the preceding discussion has also suggested that further clarification is needed as to which kinds of causally relevant differences exactly

can be handled and where extrapolative inference starts to break down as the difference become unmanageable.

7) An attractive strategy should tell us as early as possible whether (the desired kind of) extrapolation is feasible or not. This desideratum is important particularly as more involved extrapolations can require the collection of large amounts of supplementary evidence to support the assumptions involved. In the interest of allocating limited resources efficiently, it seems desirable to structure the acquisition of such evidence so as to prioritize information that can potentially be decisive as to whether extrapolation can proceed at all. For instance, as argued earlier, it seems that causally relevant similarities and differences at the level of causal structure are more fundamental than those potentially obtaining at other levels. For instance, if the mechanism in the experimental population is $X \rightarrow Z \rightarrow W \rightarrow Y$ and the relationship $Z \rightarrow W$ is disrupted in the target, then learning whether a suspected moderating variable of the $W \rightarrow Y$ effect is distributed in the same way in the target does not add anything to our conclusion that interventions on X will not have effects on Y. B&P's graph-based strategy makes some progress towards satisfying this desideratum by offering a complete algorithm for deciding transportability. What their approach is missing, however, is an *epistemic* layer that tells us, even if an effect is principally transportable under some assumptions, which assumptions should be validated first in order to learn as efficiently as possible whether to terminate extrapolation for lack of justification or to proceed instead.

8) An attractive strategy should allow us to express uncertainty about individual component assumptions as well as determine the strength of support/confidence that an extrapolative conclusion enjoys. Specifically, it is reasonable to think that in real-world cases we will rarely be able to establish the validity of substantive assumptions with certainty. Even if possible, it would be undesirable as such certainty could often only be achieved on pain of falling prey to the extrapolator's bind. It hence seems important that analysts are able to express how much confidence they have in each substantive extrapolation assumption, and conversely, how much uncertainty remains in whether they are satisfied. As suggested above, this could be greatly facilitated by trying, as much as possible, to break extrapolation assumptions down into smaller, more manageable components, e.g. assumptions pertaining to individual causal relationships, where each such component assumption is a unit of analysis in telling us what is in need

of support, how to acquire such support, and in expressing how much support is available. I will say more about approaches that could help accomplish this shortly.

9) Finally, an attractive strategy should be able to tell us which assumptions are most important, and which are less problematic to entertain as mere assumptions. This is essentially a call for the ability to perform sensitivity analyses. It seems clear that some assumptions are more important than others, e.g. those pertaining to similarities in the main causal pathways along which an effect is transmitted might be more important than those pertaining to a moderating variable that plays a subordinate role in inducing differences in effect magnitudes. In order to help us tell which assumptions are more important, it seems desirable that extrapolation strategies, or at least a more general framework accommodating such strategies, allow us to perform rudimentary sensitivity analyses, i.e. systematic explorations of how our ability to extrapolate at all, as well as the substantive content of our extrapolative conclusions and our confidence in them, change with respect to changes in the validity of and support for specific component assumptions.[3] This allows us not only to tell which assumptions are most in need of support, but also which assumptions might be entertained as mere assumptions, e.g. when certain conclusions (and our confidence in them) are robust over changes in the validity of (and confidence in) these assumptions. Finally, performing such analyses can also help with achieving 7), i.e. terminating extrapolative inference as soon as possible when not feasible, in that it can guide us 1) in prioritizing empirical efforts to learn about causally relevant differences and similarities in those features that are most important, most in need of support, and least costly to learn about, and 2) to focus on acquiring and producing evidence that has the most significant bearing on the content of our conclusions, their accuracy and precision, and the weight in their favour.

With these desiderata in place, let me expand on two broader themes that I have not touched upon so far. The first concerns issues of how to quantify uncertainty and express confidence in an extrapolative conclusion. The second concerns how EBP institutions, in particular those issuing methodological recommendations for evidence production and use, might respond to the arguments developed here, and how they

[3] See e.g. Rosenbaum (1995) who considers sensitivity analyses as a means to explore whether effect estimates are robust under changes in the validity of identification assumptions. See also Manski (1990; 2008) for a related approach.

might make important general contributions to improving the practice of extrapolation in EBP.

8.3.1 Uncertainty and Confidence

As the discussion in previous chapters suggested, substantive assumptions about causally relevant similarities are required by all strategies for extrapolation, but will often be difficult to support in practice. This not only raises concerns about the demandingness of strategies for extrapolation, but also suggests that significant uncertainty will often remain as to whether crucial assumptions are indeed satisfied. Yet, the fact that support for specific extrapolation strategies can vary greatly, both in the kind of support offered and in the strength of support afforded, is currently not captured by existing strategies. They tell us, sometimes more and sometimes less explicitly, which assumptions we need to bet on, but they do not tell us how to (best) support these assumptions, express the uncertainty surrounding them, or how confident, given some degree of support for specific assumptions, we can be in an extrapolative conclusion (see Reiss 2015 for related concerns). This is not a shortcoming of these strategies per se, but it suggests that additional layers of analysis pertaining to the uncertainties that invariably remain in extrapolation are needed to meet the epistemic and practical needs of analysts, practitioners, and policy-makers, who, in the face of real stakes, will often need to know not only what can be expected, but also how confident they can be in these expectations.

My suggestion here is that two additional layers of analysis are needed. The first concerns how confident we can be in the validity of specific extrapolation assumptions, given some support in the form of supplementary evidence ***S*** and background knowledge ***K***. The second concerns how the confidence in specific extrapolation assumptions compounds and propagates onto the confidence we are entitled to have in an extrapolative conclusion.

There is a substantial body of existing work to consider when thinking about these issues. Broad fields of study, including those concerning philosophers' darlings of Bayesian and error statistical approaches (see e.g. Bovens and Hartmann 2003; Mayo and Spanos 2011 for overviews), have been concerned with developing theory that can, among other things, help quantify and express uncertainty in scientific and everyday inference. This chapter is not the place for extensive reviews of such efforts, nor to

comment on their relative merits. Instead, I will only gesture towards some ideas that could be useful for addressing the needs identified above.

Let me begin with some suggestions for how to facilitate assessments of the support for specific extrapolation assumptions. In the context of pharmacology, Landes et al. (2018) offer a Bayesian framework for amalgamating different kinds of evidence to assess causal hypotheses about drug efficacy and harms. They aim to follow previous suggestions to consider broader varieties of evidence for this purpose (rather than, say, only RCTs), which is similar to the calls for considering broader varieties of evidence I have made in *Chapter 6*, while offering more concrete suggestions as to how diverse and possibly inconsistent bodies of evidence can be integrated in order to speak to specific causal hypotheses.

The framework used to achieve this is based on Boven's and Hartmann's (2003) Bayesian-network approach to modelling scientific inference. Here, evidential relationships between hypotheses, their observable consequences, observational reports about whether these consequences obtain, and information about the reliability and relevance of the reports are encoded in a directed graph G. To express how observational reports of observable consequences bear on the probability of specific hypotheses, investigators impute *conditional probability tables* that specify the conditional probabilities of nodes over the states that their parents can assume. Following this, a prior probability P over the variables, constrained by the conditional independencies encoded in the graph, is selected by an investigator. Then, as novel observational reports become available, a posterior probability for the hypothesis of interest can be computed[4] according to the rules of Bayesian inference (Landes et al. 2018, 25).

In the context of extrapolation, this framework could be useful for integrating diverse kinds of evidence (e.g. qualitative and quantitative), beliefs about their respective bearing on hypotheses concerning causally relevant similarities and differences, including assessments of their reliability and relevance, to form general assessments of the probability of such similarities and differences obtaining.[5] This could help express, on the level of specific component assumptions, how likely it is that specific similarities

[4] Analysts can make use of commercial software packages such as *Netica* for this purpose.

[5] Of course, as e.g. Reiss (2015) would caution, the framework proposed by Landes et al. remains unhelpful unless we know how the concrete material facts pertinent to an extrapolation context bear on the assumptions, i.e. how to construct the conditional probability tables required by Bayesian networks.

and differences are realized, and hence be useful for assessing, in a fine-grained way, what our available resources in the form of ***S*** and ***K*** say about specific instances of similarity and difference. This could be an important first step towards helping analysts navigate the unavoidable uncertainties involved in extrapolation.

However, importantly, while the approach developed by Landes et al. (2018) seems helpful for amalgamating different kinds of evidence and support for specific extrapolation assumptions and computing probabilities for whether these assumptions are satisfied, it does not, by itself, allow us to make assessments of the *weight of evidence* in favour of a specific assumption (see Peirce 1878; Keynes 1921; Good 1985). Put simply, it is one thing to have a probabilistic belief pertaining to a hypothesis, and another to have an idea of how strongly supported this belief is by the evidence involved in obtaining it. The latter is often thought to be a question of the quantity and quality of evidence in favour of a hypothesis, as well as its diversity or consistency (cf. Weed 2005). Assessments of evidence weight are important as analysts will typically not only be interested in first-order probabilities pertaining to whether crucial extrapolation assumptions are satisfied, but also in making second-order judgments about the confidence in these assessments and whether this confidence is sufficiently high to licence action.[6] This suggests that, while the evidence amalgamation framework proposed by Landes et al. is an important first step in the pursuit of investigating issues of uncertainty in extrapolation, the approach needs to be further complemented by strategies that help express considerations about the weight of evidence in favour of specific extrapolation assumptions.

My second suggestion here is that it seems important to provide analysts with the means to investigate how support for specific assumptions bears on support in favour of the extrapolative conclusion ***C*** reached by an inference that draws on these assumptions. Here, it seems unlikely that Bayesian approaches will be helpful, as the important questions do not concern how probabilistic information propagates through a network of nodes, governed by stipulated evidential relationships and updating rules. Rather, to be able to tell how the confidence in specific extrapolation assumptions compounds and propagates onto the confidence we are entitled to have in an extrapolative conclusion,

[6] This is not a new idea, of course. For instance, the IPCC (United Nations Intergovernmental Panel on Climate Change) has included in their Assessment Reports not only the primary estimates of pooled climate simulation results on e.g. quantities like predicted average surface temperature of the earth's climate system, but also communicated second-order confidence in these results. Predictions are accompanied by second-order assessments of the confidence in these predictions (see Wüthrich 2016).

we need to consider how different assumptions work together in warranting a conclusion. The relationships between these assumptions are not (merely) evidential. Rather, they will in large parts be governed by underlying theory pertaining to what assumptions are necessary to make specific kinds of inferences and how these assumptions interact, e.g. whether they need to be jointly satisfied in specific ways, are logically, causally, or probabilistically related, etc. Moreover, to assess how uncertainty concerning specific causal features of populations propagates onto uncertainty in an extrapolative conclusion, substantive causal models might be needed (e.g. along the lines of causal DAGs used by B&P). Such models can encode, among other things, how different causal features interact in producing a causal effect. In virtue of encoding such information, these models cannot only be used for deriving single predictions of causal effects in the target, but also for systematically studying how predictions would differ under variations in the causal model, e.g. in variable values/distributions, functional form and structural parameters, and basic causal structure.[7] This, together with some information pertaining to the evidential relationships between different kinds of support and specific assumptions, could then help us systematically study how the confidence in specific assumptions compounds and propagates onto an extrapolative conclusion.

Such investigations could be further supported by a framework recently developed by Roussos et al. (2019, ms.). Their aim is to address issues of decision-making under ambiguity in the context of using ensembles of climate models to predict rare climate events. In such contexts, analysts are often faced with substantive *model uncertainty,* i.e. uncertainty concerning the representational accuracy of models that might differ in parameterizations or indeed in deeper, structural features pertaining to how they represent the earth's climate system (see e.g. Parker 2013). In the context of extrapolation, similar uncertainties will often plague the analyst. Here, they are not about model uncertainty, i.e. whether specific models accurately represent their targets, but rather about *assumption uncertainty,* i.e. whether we are justified in assuming that suitable relationships of similarity at different levels and in specific respects obtain between two populations, as well as *feature uncertainty*, i.e. whether a population indeed exhibits certain causally relevant features. These different kinds of uncertainty

[7] This is similar to proposals from the literature on *computational model validation*, where, among other things, simulations on (ensembles of) computational physics models are used to quantify the uncertainty involved in making predictions of the behaviours of real-world engineering systems (see Roy and Oberkampf 2011). Similar to the present case, such simulations typically involve substantial uncertainty as to whether the computational models are accurately parameterized or adequately represent their targets at more fundamental, structural levels.

encountered in the context of extrapolation seem similar enough to those encountered in using model ensembles to think that Roussos et al.'s framework could potentially be usefully applied to issues of extrapolation as well.

The framework that Roussos et al. employ is a modified version of the *confidence approach* (see Hill 2013; Bradley 2017), which explicitly models how the weight of a body of evidence, as well as context-specific features (such as the stakes involved in a decision-making context, the desired certainty of an analyst/decision-maker, and their attitudes towards uncertainty), bear on the confidence that one is entitled to have in a model prediction (or, in our case, an extrapolative conclusion), i.e. a second-order assessment of how strongly a first-order prediction is supported by the evidence in its favour. Roussos et al. use the idea of *nested intervals*, i.e. predictions of varying precision (typically imprecise probability estimates) of the quantities of interest derived under slightly different conditions (e.g. from only a part of a model ensemble). Their approach allows analysts to gauge how much confidence differently precise probability intervals each enjoy given how many predictions from a model ensemble underwrite them. A highly precise prediction interval, for instance, will typically only be supported by a narrower range of models, so the weight in favour of these precise intervals, and the confidence they might enjoy, will be lower.

Applied to extrapolation, using this framework could proceed by deriving a series of (imprecise) predictions starting from the case where all extrapolation assumptions are assumed to be satisfied, and departing systematically from this ideal to cases where the support for specific component assumptions dwindles. In essence, this would amount to a sensitivity analysis that investigates how changes in the support for specific component assumptions can induce changes in the support for a conclusion. The idea of nested intervals could be useful here, too, as it might allow analysts to tell how confident they can be in each of a series of nested prediction intervals for the quantities of interest. Conversely, if the confidence afforded for a specific interval (say, one that would be sufficiently precise to be action-guiding) is insufficient, this would help analysts recognize that additional support for crucial assumptions might be needed. An attractive feature of the confidence approach in this regard is that it explicitly models decision-makers' uncertainty attitudes and the non-epistemic stakes involved in a decision-making context. In doing so, it moves us closer to practice as it can help study how differences in the support for specific assumptions bear on whether the confidence

in an extrapolative conclusion would be *sufficiently* high, as judged by a decision-maker. Taking such features into account is not only useful in the ways sketched out above, but also because it may allow analysts to terminate an extrapolation early on if it becomes clear that difficult-to-justify assumptions are too important for a specific conclusion.

In sum, the approach used by Roussos et al. can offer an important improvement to existing views, which say little on these issues. Cartwright's *Argument Theory* is an exception here, but it too offers little detail given it only suggests that an extrapolation is only ever as strong as the weakest link (see Cartwright and Stegenga 2011), i.e. as strong as the assumption that is least well supported. While this is helpful for making analysts more cautious about extrapolation, it does not seem correct in all cases. For instance, if the question is whether the mechanism $X \rightarrow Z \rightarrow Y$ is instantiated in the target, then learning that $X \rightarrow Z$ is instantiated there, but being entirely uncertain as to whether $Z \rightarrow Y$ is also instantiated there, does not mean that one should have no confidence whatsoever that $X \rightarrow Z \rightarrow Y$ is instantiated. The *weakest-link thesis* seems inadequate on the level of specific extrapolation assumptions. It seems more plausible when assumptions are aggregated, their interrelations are already accounted for at more abstract levels, and inference can be cast in terms of effectiveness arguments and their premises. Here, it would seem more plausible to think that premises pertaining to orthogonal issues can sometimes be thought of as working together in the way that Cartwright suggests, e.g. when we separate issues of whether mechanisms are sufficiently similar from whether causal support factors are distributed in the right way. Yet, at the level of specific component assumptions pertaining to individual variables and causal relationships, it seems that more complicated interactions will need to be taken into account. It hence seems that more work pertaining to the dynamics of uncertainty and confidence in extrapolation is needed to allow analysts to make finer-grained assessments of how well their extrapolations are supported.

Taking some steps backwards from these specific suggestions, it seems that, at least in principle, approaches used for modelling evidence amalgamation for specific extrapolation assumptions, forming second-order assessments of confidence in these assumptions, and investigating the dynamics of evidence weight propagation can be linked up in a neat way. In such an integrated framework, support for an extrapolation propagates 'upwards' from background knowledge $\boldsymbol{K}$ and supplementary evidence $\boldsymbol{S}$ of

various sorts onto specific assumptions ***P*** pertaining to causally relevant similarities between populations, and then propagates 'forwards' from these assumptions onto the extrapolative conclusion ***C***, mediated by various contextual features, such as uncertainty attitudes, the nature of the real-world stakes involved in an extrapolation, etc. As suggested, such an overarching framework could also be useful for answering a range of further questions, including: Which extrapolation assumptions are the most important ones? Which additional supplementary evidence and background knowledge should we obtain to yield the highest increase in confidence in our extrapolative conclusion? We might also be able to further specify details about the relative costs of obtaining such evidence, as well as encode, at the level of specific kinds and tokens of evidence, whether obtaining them would raise concerns about the extrapolator's bind (e.g. by setting costs arbitrarily high).

To be sure, the suggestions made here are, of course, no more than that, and plenty of additional work is needed to tell whether there are indeed ways of adding these levels of analysis to primary extrapolative inference in order to provide important second-order information about uncertainty and confidence, without thereby overburdening the analyst (e.g. asking them to write down conditional probability tables for a Bayesian network, or to contemplate how different extrapolation assumptions interact). The need to say more on these issues, however, is uncontroversial, and, I believe, will play an important part in moving the academic debate surrounding extrapolation closer to addressing the practical needs and constraints arising in the context of real-world extrapolations.

8.3.2 Institutional Desiderata: Making Recommendations and Testing Strategies

Here, I briefly want to make two suggestions concerning the widely circulated methodological recommendations made by key EBP institutions, such as 3ie, JPAL, the What Works Clearinghouse, the Campbell Collaboration, GRADE, Consort, and others. So far, the methodological recommendations made by these institutions for how to produce, amalgamate, and use effectiveness evidence for purposes of informing policy often remain silent on issues of extrapolation. When they do comment on extrapolation, the advice often takes the form of encouraging authors of effectiveness studies to comment on the suspected applicability of their results and alerting evidence users to the need to ensure that experimental and target populations are appropriately similar.

But little effort is made to propose systematic approaches to extrapolation, mention those examined here, or make more specific recommendations for what kinds of problems users of evidence need to be aware of.[8]

This is not surprising, as at least some of these institutions are primarily concerned with issuing guidelines for what constitutes high-quality effectiveness evidence and proposing hierarchies to distinguish evidence with respect to how reliable it is in informing us about the effectiveness of interventions where they were studied. This is an important first step, but, in the spirit of existing criticisms and the arguments developed here, it is equally important to recognize that these measures fall radically short of addressing the concrete needs experienced by decision-makers who wish to apply effectiveness evidence potentially far outside of the domains where it has been obtained. Here, it is easy to confuse high-quality evidence of causal effects *somewhere* with high-quality evidence that speaks strongly to questions about the target of interest (cf. Cartwright 2013; Cartwright and Hardie 2012, ix.). There is clearly room for improvement, therefore, in how issues of extrapolation are currently handled by institutions that issue methodological guidelines.

One obvious, but rather thin, suggestion is to call for more explicit commentary on problems of extrapolation, including what different kinds of problems there are and how these are constituted; how study design can facilitate extrapolation; what methods are available to perform extrapolation, what assumptions they involve, and what resources might be needed to support them.

Another, more substantive suggestion takes issue with criteria for what makes evidence 'high quality', i.e. evidence-hierarchies. In light of the arguments developed here and elsewhere (e.g. Khosrowi 2019; Reiss and Khosrowi 2019, ms.), it seems important to distinguish between the quality of evidence in establishing causal effects of interventions in a study and the *applicability* of evidence to novel settings (including the quality or reliability of different kinds of *supplementary* resources needed to underwrite extrapolation to these settings). As argued in *Chapter 6*, for instance, it seems that different kinds of evidence, such as quantitative and qualitative, can sometimes productively interact in underwriting extrapolation in ways that would not be possible

[8] JPAL is a notable exception here, as it recently advertised postdoctoral research positions, including on extrapolation, where research might include "[…] simulating and comparing extrapolation methods, or pooling data for meta-analyses across different RCTs with similar interventions and outcomes"; see https://www.povertyactionlab.org/careers/postdoctoral-associate-j-pal-global-102760.

with either type of evidence alone. Hence, it would seem desirable to investigate in more detail how such productive interactions come about, how they can be facilitated, and how certain combinations of evidence might be especially suitable for underwriting extrapolative inference. Moreover, it would also seem helpful if methodological guidelines for evidence production were complemented by attempts to provide *use-centric* grading schemes for what types of supplementary resources are helpful for underwriting extrapolation; how the quality or adequacy of *bodies* of evidence, rather than individual tokens, might be assessed; and how quality can differ over various dimensions, such as the type of problem targeted, the type of conclusion envisioned, the desired fidelity of the inference, etc.

The second suggestion I want to make here relates to an important limitation of the present project: as a largely theoretical and critical endeavour, it cannot assess the real-world capabilities of different strategies for extrapolation beyond what can be argued from the discomfort of an office chair. While it seems clearly important to improve our theoretical understanding of strategies for extrapolation in order to help us determine when they are sensibly applicable, it is also clear that this can only go so far, and that facilitating a comprehensive understanding of which strategies work for what kinds of problems will, in part, need to involve trying them out in their envisioned domain of application. Hence, it seems reasonable to suggest that, adding to the calls for developing and using middle-range theory pertaining to specific interventions, we also need complementary efforts to systematically study and form general conclusions about the conditions of applicability of different extrapolation strategies. Here, we may wish to address various general questions, including: how much causal information is needed to facilitate different kinds of extrapolation and how much is typically available in different settings? What kinds of causally relevant differences tend to upset specific extrapolation strategies most? How fast do different combinations of empirical learning methods and extrapolation strategies converge on the true value of a causal effect in a target?

These questions can be addressed in different ways. One strategy is empirical and simply proceeds by conscientious and systematic efforts to document the success of using different extrapolation strategies under varying conditions. Another approach is to study the performance of extrapolation strategies virtually, in simulation studies. Here, much like in econometrics and many other disciplines involved with causal and

statistical inference, Monte Carlo simulation studies can be used to generate datasets for hypothetical experimental and target populations from mechanisms that the investigator specifies themselves. Since the ground-truth about features of these mechanisms, and hence about their similarities and differences, is known, this makes it easier to assess the performance of different strategies in handling different kinds of causally relevant differences between the populations so specified, as well as the ability of at least some empirical methods to adequately recover features of these mechanisms (such as the ability of machine-learning-based approaches to detect and attribute heterogeneous causal effects, or the ability of causal discovery methods to correctly identify causal structures from data).

Some efforts to study the performance of extrapolation strategies and supplementary empirical methods have already been undertaken. For instance, a large part of the discussion in Hotz et al. (2005) consists of reports on empirical tests of their proposed strategy. Similarly, the performance of machine-learning approaches is also routinely gauged in simulation studies.

Here, I want to call not only for proponents of specific extrapolation strategies to invest more heavily in demonstrating the effectiveness of their proposed methods, but for broader efforts to be made on the part of major EBP institutions to guide analysts by offering more details on available extrapolation strategies, to study their applicability and performance, and to widely disseminate the results of such analyses so that practitioners can use these resources to inform their extrapolation efforts.

8.4 Positive Proposals Summarized

In sum, the proposals made in this chapter make clear that it is possible to add interesting insights pertaining to how we might underwrite extrapolative inference, and in a way that is more general than pointing to the importance and usefulness of concrete contextual resources. This helps build a basis for a future, more general *framework* for extrapolation, i.e. one that, among other things, provides thick layers of analysis that can connect abstract strategies for extrapolation with the concrete epistemic challenges encountered in real-world settings. To this end, I have highlighted a variety of empirical methods that 1) can be used in the absence of theory to learn about causal features that might be important to consider, 2) can be helpful in conjunction with theory by supplying information that theory alone cannot deliver, and 3) help us build such theory

if it is so far unavailable. Moreover, I have proposed several desiderata that one might consider in further developing existing strategies for extrapolation. While they aim to improve the responsiveness of general strategies to the concrete features of particular extrapolations, the desiderata themselves are general in spirit. Finally, I have made some suggestions for how a general framework for extrapolation might be enriched by additional efforts to explore issues surrounding uncertainty and confidence in extrapolation. These, too, are issues that can be explored at a general level, while at the same time linking abstract strategies for extrapolation with the concrete details of specific extrapolation contexts, particularly with a view towards meeting important epistemic needs experienced by extrapolators.

The arguments and analyses provided in previous chapters have begun to build the basis for a general framework for extrapolation that can make recommendations for 1) how analysts and decision-makers can identify different kinds of problems of extrapolation and how they can tell, before any further efforts are undertaken, how challenging these problems are likely to be; 2) what general extrapolation strategies might be useful for addressing these problems; 3) what assumptions will need to be made to enable extrapolative inference of the desired kind; 4) what additional resources might be needed to support these assumptions; and 5) what kinds of problems of extrapolation are unlikely to be overcome by drawing on such resources, at least without raising concerns about the extrapolator's bind.

The present chapter complements these contributions by further clarifying: 6) how to make use of existing resources for underwriting extrapolative inference; 7) when one might need to acquire additional evidence; 8) what methods are available for doing so; 9) what frameworks should be considered when integrating different sources of support; and 10) how to think about and express the uncertainties that will remain in most kinds of extrapolation. In virtue of providing details on these important issues, the present chapter has made some further contributions to systematizing the theory and practice of extrapolation in EBP.

References

Allcott, H. (2015). "Site Selection Bias in Program Evaluation", *The Quarterly Journal of Economics*, 2015, 130(3): 1117–1165.

Astbury, B., and F. Leeuw. (2010). "Unpacking Black Boxes: Mechanisms and Theory Building in Evaluation", *American Journal of Evaluation*, 31(3): 363-81.

Athey, S. and G. W. Imbens. (2016). "Recursive partitioning for heterogeneous causal effects", *PNAS* 113: 7353–7360.

Athey, S. and G. W. Imbens. (2017). "The state of applied econometrics: causality and policy evaluation", *Journal of Economic Perspectives*, 31(2): 3–32.

Bareinboim, E. and J. Pearl. (2016). "Causal inference and the data-fusion problem", *Proceedings of the National Academy of Sciences*, 113: 7345-52.

Beach, D., and R. B. Pedersen. (2013). *Process-tracing methods: Foundations and guidelines*. Ann Arbor: University of Michigan Press.

— **(2016)**. *Causal case study methods: Foundations and guidelines for comparing, matching and tracing*. Ann Arbor: University of Michigan Press.

— **(2019)**. *Process-Tracing Methods - Foundations and Guidelines*, 2nd edition. Ann Arbor: University of Michigan Press.

Bicchieri, C., J. W. Lindemans, and T. Jiang. (2014). "A structured approach to a diagnostic evaluation of practices". *Frontiers in Psychology*. 5(1418): 1-13.

Bovens, L., Hartmann, S. (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Bradley, R. (2017), *Decision Theory with a Human Face*. Cambridge: Cambridge University Press.

Brittain, E., and J. Wittes. (1989). "Factorial designs in clinical trials: The effects of non-compliance and subadditivity", *Statistics in Medicine*, 8: 161-171.

Campbell, D. T. (1986). "Relabeling internal and external validity for applied social scientists". In W. M. K. Trochim (ed.), *Advances in quasi-experimental design and analysis*, pp. 67–77, San Francisco, CA: Jossey-Bass.

Carnap, R. (1947). *Meaning and Necessity*. Chicago, IL: University of Chicago Press.

Cartwright, N. D. (2013). **"**Knowing what we are talking about: why evidence doesn't always travel". *Evidence & Policy*, 9(1): 97-112.

— **(Forthcoming)**. "Lullius Lectures 2018: Mid-level theory: Without it what could anyone do?" In: C. Martínez Vidal and C. Saborido (ed.), *Nancy Cartwright's Philosophy of Science*, Special Issue of *Theoria*.

Cartwright, N. D., and J. Hardie. (2012). *Evidence-Based Policy: A Practical Guide to Doing It Better*. New York: Oxford University Press.

Cartwright, N. D. and Stegenga, J. (2011). "A Theory of Evidence for Evidence-Based Policy". In P. Dawid, W. Twining., and M. Vasilaki (eds.), *Evidence, Inference and Enquiry* (Proceedings of the British Academy), New York: Oxford University Press.

Chickering, D. M. (2002). "Optimal structure identification with greedy search", *Journal of Machine Learning Research*, 3: 507-54.

Cook, T., and D. T. Campbell. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton-Mifflin.

Crasnow, S. (2011). "Evidence for use: Causal pluralism and the role of case studies in political science research", *Philosophy of the Social Sciences*, 41(1): 26-49.

Dehejia, R., C. Pop-Eleches, and C. Samii. (2015). "From Local to Global: External Validity in a Fertility Natural Experiment", Working Paper 21459, National Bureau of Economic Research.

Eberhardt, F. (2007). *Causation and Intervention*, PhD thesis, Carnegie Mellon University, Pittsburgh, PA.

— **(2012)**. “Almost optimal intervention sets for causal discovery ”, *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI2008*), arXiv:1206.3250.

— **(2017)**. "Introduction to the foundations of causal discovery", *International Journal of Data Science and Analytics*, 3(2): 81-91.

Eva, B., R. Stern, and S. Hartmann. (forthcoming), “The Similarity of Causal Structure”, *Philosophy of Science*, https://doi.org/10.1086/705566.

Fink, G., M. McConnell, and S. Vollmer. (2014). “Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures”, *Journal of Development Effectiveness*, 6: 44-57.

Gal, D., and D. D. Rucker. (2018). "Loss Aversion, Intellectual Inertia, and a Call for a More Contrarian Science: A Reply to Simonson & Kivetz and Higgins & Liberman", *Journal of Consumer Psychology*, 28 (3): 533–39.

Gill, D., and V. Prowse (2012). "A structural analysis of disappointment aversion in a real effort competition", *American Economic Review*, 102(1): 469–503.

Good, I. J. (1967). “On The Principle of Total Evidence“, *The British Journal for the Philosophy of Science*, 17(4): 319-21.

— **(1985)**. “Weight of Evidence: A brief survey”. In Bernardo, J. DeGroot, M. Lindley, D. and Smith, A. (eds.), *Bayesian Statistics*, 2nd edition, pp. 249-69 Amsterdam: North Holland.

Hill, B. (2013). “Confidence and Decision”, *Games and Economic Behaviour,* 82: 675–92.

Hotz, V. J., G. W. Imbens and J. H. Mortimer. (2005). “Predicting the efficacy of future training programs using past experiences at other locations”, *Journal of Econometrics* 125: 241–70.

Hyttinen, A., F. Eberhardt, and P. Hoyer. (2013). “Experiment selection for causal discovery”, *Journal of Machine Learning Research*, 14: 3041–71.

Imai, K., L. Keele, D. Tingley, and T. Yamamoto. (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, K., D. Tingley and T. Yamamoto. (2013). “Experimental designs for identifying causal mechanisms”, *Journal of the Royal Statistical Society A*, 176: 5–51.

Kahneman, D., and Tversky, A. (1979). "Prospect Theory: An Analysis of Decision under Risk", *Econometrica*, 47(2): 263–91.

Kass, G. V. (1980). “A Exploratory Technique for Investigating Large Quantities of Data”. *Journal of the Royal Statistical Society C*, 29 (2): 119-127.

Keynes, J. M. (1921). *A Treatise on Probability*. London: Macmillan.

Khosrowi, D. (2019). “Trade-offs Between Epistemic and Moral Values in Evidence-Based Policy”, *Economics & Philosophy*, 35(1): 49-78.

Kocaoglu, M., A. G. Dimakis, and S. Vishwanath. (2017). “Cost-optimal learning of causal graphs”, *Proceedings of the 34th International Conference on Machine Learning.*

Landes, J., B. Osimani, and R. Poellinger (2018). “Epistemology of Causal Inference in Pharmacology. Towards a Framework for the Assessment of Harms”, *European Journal for Philosophy of Science*, 8(1): 3-49.

Manski, C. (1990). “Nonparametric Bounds on Treatment Effects”, *American Economic Review*, 80: 319-23.

— **(2008).** *Identification for Prediction and Decision*. Princeton: Princeton University Press.

Marcellesi, A. (2015). “External Validity: Is There Still a Problem?”. *Philosophy of Science*, 82(5): 1308-17.

Mayo, D. G., and A. Spanos. (2011). “Error Statistics”, in: Gabbay, D., P. Thagard and J. Woods (eds.), *Handbook of Philosophy of Science*, vol. 7, Philosophy of Statistics, pp. 151–196, Amsterdam: Elsevier.

Merton, R. K. (1968 [1949]). *Social theory and social structure*. Enlarged edition. New York: Free Press.

Montgomery, A. A., T. J. Peters, and P. Little. (2003). "Design, analysis and presentation of factorial randomised controlled trials", *BMC Medical Research Methodology*, 3:26.

Muller, S. M. (2015). "Interaction and external validity: obstacles to the policy relevance of randomized evaluations", *World Bank Economic Review*, 29(1): 217-25.

Murphy, K. P. (2001). "Active learning of causal bayes net structure". Technical report, Department of Computer Science, U.C. Berkeley.

Parker, W. (2013). "Ensemble Modeling, uncertainty and robust predictions", *Wiley Interdisciplinary Reviews: Climate Change*, 4(3): 213-23.

Pawson, R., and N. Tilley. (1997), *Realistic Evaluation*. London: SAGE Publications.

—— (2001). "Realistic Evaluation Bloodlines". *American Journal of Evaluation*, 22: 317-24.

Pawson, R. (2013). *The science of evaluation: a realist manifesto*. London: SAGE Publications.

Pearl, J. (2011). "The Causal Mediation Formula -- A Guide to the Assessment of Pathways and Mechanisms", *Prevention Science*, 13: 426-36.

—— (2014). **"**Reply to Commentary by Imai, Keele, Tingley, and Yamamoto Concerning Causal Mediation Analysis", *Psychological Methods*, 19(4): 488-92.

Peirce, C. S. (1878). "The Probability of Induction." *The Popular Science Monthly, Illustrations of the Logic of Science*, XII.

Reiss, J. (2015). **"**A Pragmatist Theory of Evidence", *Philosophy of Science*, 82(3): 341-62.

Reiss, J., and D. Khosrowi. (2019, ms.), "Evidence-Based Policy and Its Hidden Costs of Justice". Unpublished manuscript, Durham University.

Rosenbaum, P. (1995). *Observational Studies*. New York: Springer.

Roussos, J., R. Bradley, and R. Frigg (2019). "Making confident decisions with model ensembles". Unpublished manuscript, LSE.

Roy, C., and W. L. Oberkampf. (2011). "A Complete Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing", *Computer Methods in Applied Mechanics and Engineering*, 200: 2131-44.

Shadish, W. R., T. D. Cook, and D. T. Campbell. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Spirtes, P., C. Glymour, and R. Scheines. (2000). *Causation, Prediction and Search*, 2nd ed., Cambridge, MA: MIT Press.

Strevens, M. (2007). "Why Represent Causal Relations?" In A. Gopnik and L. Schulz (eds.), *Causal Learning: Psychology, Philosophy, Computation*. New York. Oxford University Press.

Tong, S., and D. Koller (2001). "Active learning for structure in bayesian networks". *Proceedings of the International Joint Conference on Artificial Intelligence.*

Varadhan, R and J. D. Seeger. (2013). "Estimation and reporting of heterogeneity of treatment effects. In P. Velentgas, N. A. Dreyer, P. Nourjah, S. R. Smith and M. M. Torchia, *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*, pp. 35–44, Rockville, MD: Agency for Healthcare Research and Quality.

Wager, S., and Athey, S. (2018). "Estimation and inference of heterogeneous treatment effects using random forests", *Journal of the American Statistical Association*, 113(523): 1228-42.

Weed, D. L. (2005). "Weight of Evidence : a Review of Concept and Methods", *Risk Analysis*, 25(6): 1545–57.

Weinberger, N. (2019). "Path-Specific Effects", *The British Journal for the Philosophy of Science*, 70(1): 53-76.

White, H. (2009). "Theory-based impact evaluation: Principles and practice". Working paper 3 of International Initiative for Impact Evaluation. New Delhi: International Initiative for Impact Evaluation.

Wüthrich, N. (2016). "Conceptualizing Uncertainty: An Assessment of the Uncertainty Framework of the Intergovernmental Panel on Climate Change", In: EPSA15 Selected Papers: The 5th conference of the European Philosophy of Science Association in Düsseldorf.

CHAPTER 9

Final Conclusions

This thesis has taken issue with problems of extrapolation as routinely encountered in EBP. My main aim was to critically evaluate the extent to which existing strategies for extrapolation can persuasively address real-world extrapolation problems. To this end, *Chapters 2* and *3* have developed the necessary theoretical background. *Chapters 4*, *5*, *6*, and *7* have focused on specific strategies for extrapolation, what assumptions they involve, and how they are vulnerable to the extrapolator's bind. Finally, *Chapter 8* provided a positive outlook concerning how to address the challenges identified in earlier chapters. Let me briefly revisit each chapter to emphasise the distinctive contributions made there.

Chapter 2 provided an extensive discussion of problems of extrapolation. Here, I presented a detailed and novel analysis arguing that problems of extrapolation, as well as extrapolative inference, are highly heterogeneous and can differ along a variety of important dimensions. This helped clarify that some problems of extrapolation are considerably more challenging than others, and provided a useful general background for further investigating what kinds of problems of extrapolation can realistically be overcome by specific strategies.

Chapter 3 laid out some basic assumptions involved in extrapolation, irrespective of the particular strategy adopted. This helped recognize that strategies for extrapolation share a common core of assumptions, which can motivate the guiding ideal that extrapolation can, in principle, be successful. In addition, I provided a working analysis of what extrapolation is, at the most general level, and took issue with Steel's extrapolator's circle. Building on Steel's work, I offered some conceptual refinements, making clear that the extrapolator's circle is better understood as a special case of a wider challenge which I called the *extrapolator's bind*. The extrapolator's bind, in turn, was further fleshed out as a gradual affair, reflecting the idea that whether the relevance of an experimental result is displaced by supplementary evidence and background knowledge is not a dichotomy but a matter of degree. Against this background, I proposed a novel analysis of *successful* extrapolation, which integrates the extrapolator's bind. According to this analysis, successful extrapolation is a function not

just of whether our extrapolative conclusions answer to the question at hand, of whether the answers are accurate, precise, and well-justified, but also of whether the experimental result remains relevant to our conclusion. Framed as an imprecise desideratum (further informed by contextual details), successful extrapolation requires that we steer clear of the extrapolator's bind as much as possible and ensure that our inference remains ampliative with respect to the additional resources we draw upon.

Chapter 4 was the first chapter to take a closer look at proposals for how to extrapolate, focusing on Cartwright's *Argument Theory of Evidence*. Here, I argued that the Argument Theory is a useful framework for thinking about extrapolation in general, as well as for casting different forms of extrapolative inference in explicit form and thereby promoting their critical appraisal. I then took issue with the *effectiveness argument*. While it is best understood as a mere illustration of how the Argument Theory is supposed to work, it is also useful for demonstrating how even a simplistic extrapolative inference can involve surprisingly strong assumptions that raise important concerns about the extrapolator's bind. Moving away from these particulars, I argued that the Argument Theory, beyond its concrete illustration in the form of the *effectiveness argument*, does not provide further details on alternative, more sophisticated extrapolation arguments. These details, I have suggested, might need to be borrowed from other strategies for extrapolation. I have argued that this is not a shortcoming, however, as the *Argument Theory* can nevertheless figure as an important framework to facilitate not only (some) concrete instances of extrapolation, but also, and more importantly, critical assessment of other *strategies* for extrapolation, by helping make their assumptions explicit and enabling the scrutiny of these assumptions. The remainder of the thesis has largely followed the Argument Theory in this spirit and ambition.

Chapter 5 focused on Steel's *mechanism-based strategy* for extrapolation, which employs *comparative process tracing* to help extrapolate claims of causal relevance. Here, I flagged some basic concerns about the applicability of CPT to social science extrapolation. Moving towards more fundamental concerns, and building on Steel's detailed investigations, I proposed a distinction between *attributive* and *predictive* extrapolation, where the latter is common in EBP settings. Drawing on this distinction, I argued that CPT exhibits important limitations in predictive extrapolation settings where the causal mechanisms of interest are 'dormant' in the target, as this can severely complicate the acquisition of evidence that bears on issues of causally relevant

similarity and difference. So while Steel's CPT-based strategy provides an effective inferential shortcut for facilitating attributive extrapolation, at least some kinds of predictive extrapolation pose severe obstacles to it. This, together with important limitations in scope, suggested that Steel's strategy, at least on its own, is unlikely to permit successful extrapolation in a wide range of cases in EBP.

Chapter 6 examined *interactive covariate-based strategies* proposed by econometricians. Here, I argued that, while promising to address a wide range of extrapolative queries, including those of central interest to EBP, these strategies involve substantive but unspoken assumptions about causally relevant similarities at the level of parameters, functional form, and basic causal structure. Entertaining these assumptions as mere assumptions would be undesirable, as assuming such similarities is often unwarranted. However, supporting these assumptions is by no means straightforward either and raises important concerns about the extrapolator's bind. As a potential remedy, I considered whether Steel's CPT can usefully complement interactive covariate-based extrapolation, at least by clarifying issues of similarity in basic causal structure. In contrast to the concerns about 'dormant' mechanisms raised in *Chapter 5*, here I sketched out conditions under which the problems surrounding 'dormant' mechanisms in predictive extrapolation might still be overcome. The main conclusion was that, at least in some cases, it seems that integrating different kinds of strategies for extrapolation, as well as different kinds of evidence, i.e. qualitative and quantitative, can offer important ampliative leverage, i.e. it allows us to reach conclusions that would be inaccessible from either type of method or evidence alone. This suggests that at least some kinds of predictive extrapolation in EBP could be strongly facilitated by revising standards of what supplementary evidence and other sources of justification are relevant.

Chapter 7 focused on the *graph-based extrapolation* approach offered by Bareinboim and Pearl. Although the graph-based approach has many important virtues, including the ability to graphically encode causal knowledge and assumptions, permit non-parametric inference, encode different kinds of causally relevant differences, and provide an effective algorithm to decide the transportability of causal effects, there are also important shortcomings. I focused on three problems in particular: 1) it remains unclear how selection diagrams can encode several important types of causally relevant differences, potentially limiting the scope of the graph-based approach. 2) Selection diagrams involve substantive parametric identity assumptions that are difficult to

support empirically, raising concerns about the extrapolator's bind, especially in predictive extrapolation. 3) Transport formulae, including those derived in B&P's examples, can sometimes demand observational quantities from the target that are not meaningfully measurable in predictive extrapolation cases where the mechanisms are 'dormant'. This discussion further supported the idea that 'dormant' mechanisms are an important concern, and that the extrapolator's bind remains a serious problem even for approaches that promise to evade it by only requiring observational evidence from the target.

Drawing on these contributions, the present work can be understood as an attempt to complement existing abstract extrapolation strategies by building a basis for a more general and practice-oriented *framework* for extrapolation, i.e. one that recognizes and takes issue with important features of the wider problem space encountered in real-world extrapolations, in particular the epistemic needs arising there and the challenges involved in meeting these needs. My contributions towards building such a framework span from providing a general analysis of problems of extrapolation and extrapolative inference, to formulating a guiding ideal and important normative strictures on successful extrapolation; highlighting specific classes of extrapolation that are especially problematic; making suggestions for how these might be overcome by integrating different methods and kinds of evidence; and identifying broad classes of problems that are unlikely to be overcome.

Chapter 8 added to these contributions by making several more concrete suggestions for how extrapolation might be underwritten by different theoretical and empirical resources, as well as how these resources can productively interact. In addition, I formulated a substantial list of general desiderata for more attractive, future strategies for extrapolation, and general frameworks that might accommodate such strategies. Finally, I identified two further areas where additional research is needed in building a general framework for extrapolation, aiming to improve both our theory of extrapolation as well as its practice. The first pertains to issues concerning uncertainty and confidence. Here, I made suggestions for how we might think about expressing our confidence in specific assumptions, how our confidence in specific assumptions propagates onto the confidence we are entitled to have in a conclusion, and what existing frameworks we might draw upon to achieve this. Second, I have called for EBP institutions to amend existing evidence-ranking schemes so as to accommodate the role of supplementary evidence needed for extrapolation and important interactions between

different kinds of evidence. Moreover, I have argued that additional efforts to investigate and document the performance of existing extrapolation strategies in concert with different supplementary resources might be undertaken to help extrapolators put available resources to (better) use.

In summary, the present work suggests that the attractiveness and real-world applicability of existing strategies for extrapolation is limited. Strategies that only get things right in the abstract are of little use for practice, and leaving the substantial epistemic problems encountered in extrapolation unaddressed just means that abstract strategies can only tell us when extrapolation is possible in principle. From the point of view of practice, this is unsatisfactory, as it falls short of providing a recipe for *successful* extrapolation and undermines the main promise of EBP: that evidence libraries are informative for predicting the effectiveness of interventions in novel settings. The strategies examined here involve substantive assumptions about causally relevant similarities, often going well beyond those explicitly discussed, and none of them, by itself, provides a compelling recipe for how to evade the extrapolator's bind. This is hardly surprising. Extrapolation, we might say, is like induction. The very nature of the problem suggests that asking for a definitive solution is elusive. Real-world extrapolative inference is piecemeal and many of the things required to support it will be tied to the concrete contexts of specific extrapolations.

Yet, as I have argued, some general things can be said, too, e.g. that ensuring the relevance of an experimental result is a key normative requirement for successful extrapolation; predictive extrapolation poses distinct challenges; 'dormant' mechanisms are particularly problematic; and integrating different kinds of methods and evidence can be important. Building on these insights, it seems possible to further develop a useful, more general framework to facilitate extrapolation. As the present work suggests, such a framework could make important contributions to structure the epistemic activities taking place in extrapolation at an intermediate level, i.e. between abstract strategies and the concrete resources available in each setting. Complementing both, this could help close the substantial gap between abstract strategies for extrapolation and its concrete epistemic practice, and help EBP live up to its ambitions of informing policymaking by drawing on high-quality evidence and, in the future, high-quality *inferences*, too.