ECONOMIC SCIENCE FOR USE:

CAUSALITY AND EVIDENCE IN POLICY MAKING

Economische wetenschap voor de praktijk: over causaliteit en bewijs in beleidsvraagstukken

THESIS

to obtain the degree of Doctor from the Erasmus University Rotterdam by command of the rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board The public defence shall be held on

Friday 2 September 2016 at 11:30 hrs

by

LUIS MIRELES-FLORES

Born in Xalapa, México

Erasmus University Rotterdam

Doctoral Committee

Promotors:

Prof.dr. I.U. Mäki
Prof.dr. J.J. Vromen

Other members:

Prof.dr. J.M. Reiss
Prof.dr. R.E. Backhouse
Dr. M.J. Boumans

Copromotor:

Dr. C. Marchionni

To Mark Blaug

TABLE OF CONTENTS

PREFACE AND ACKNOWLEDGEMENTS

This thesis has been work in progress for so long that now it is difficult for me to recognize the outcome in light of the plans and ideas I initially had. My friends and family have expressed their happiness about me finally completing it. I appreciate that. From a practical perspective, I also feel somewhat happy about getting it done. But, from an intellectual perspective, I must admit I am not entirely satisfied with the result. It barely touches the surface of a topic in the philosophy of the social sciences that needs a lot more systematic investigation: the methodology of economic policy-making.

I feel I actually have worked on four different dissertations at different stages during these years. The first was a fantasy. Rather over-ambitiously, I wanted to develop a universal account of science-based policy. I would conceptually analyse all its elements, agents, patterns, mechanisms, disturbing factors, and so forth. I would then present them in a majestic philosophical framework that could be used to characterise and assess all policy-making instantiations. Mark Blaug was my main supervisor back then. He told me that the topic was pretty good, but that I was crazy, and that the only possible way to go was to use case studies.

This led me to a different, more realistic, version of my project. Together Mark and I developed a plan for the whole thesis in which most of the chapters were cases from the history of economics related to policy issues, e.g., the classical theory of markets, Hume's writings on money, the early 20th century debates on monetary policy, privatization, and free trade. It was a great plan. I still look at it and think somebody should get it done at some point. Mark used to tell me he would do it himself in case I wouldn't. Anyway, I worked on that project for some time, until Mark had to retire for health reasons.

I got a bit stuck after that, so I turned to my other supervisor, Uskali Mäki, and visited Helsinki for the first time. I started reading everything I could on models and scientific representation, and revised my plan again. The new project focused on how models in economics can be, and are used (and misused) for policy purposes. There are probably many interesting aspects yet to be analysed about how scientific representation in the social sciences affects and is affected by the policy-making process, but in retrospect I think I was simply unprepared to give that project the treatment it deserved. Most of the literature on modelling and the people working on it at the time were focused on the epistemic virtues of models, and on whether and how they represent real or ideal targets. I wrote two chapters trying to connect that literature with notions like 'policy relevance' and 'effectiveness'. They did not make it to the final cut.

Then Julian Reiss arrived to teach at Rotterdam and with him the knowledge and motivation that got me writing the fourth and final version of my dissertation. The philosophical analysis of causality and causal inference was (not surprisingly) the missing link for my project to start making sense. The issue was how the findings of scientific economics can be used for supporting and achieving effective policy, and more generally how science and policy connect. As I became more familiar with the literature on causality, it seemed obvious that the policy-making process essentially consists in making use of knowledge about causal influences, causal structures, and causal inference, with the aim of affecting socioeconomic reality in specific ways. Thus the key was to analyse those notions in relation to policy. Whether causality is "the cement of the universe" might still be an open metaphysical question, yet it definitely turned out to be the cement of my project. Causality and causal inference play a role in policy-making that is methodologically distinct in a number of meaningful respects from their typical role in scientific practice. My whole thesis argues for this, so I won't say more here.

The final outcome is narrow, fairly specific, and utterly preliminary in several respects. Moreover, my illustrations are limited to a couple of case studies, each of which could provide, in my view, enough philosophically intriguing material to be developed into whole dissertations of their own. I am aware that such an outcome is the kind of highly focused analysis that is commonly expected in a PhD dissertation, still seeing how little my manuscript actually manages to uncover about such noteworthy and intricate topic is what inevitably makes me feel unsatisfied. In my view, this thesis mostly points towards what I believe to be the right account of the methodology of economic policy-making. Hopefully, I will have more opportunities to continue analysing the topic and to help promote it further for discussion at the philosophical and methodological arenas.

§

As the preceding remarks should make clear, Mark Blaug, Uskali Mäki, and Julian Reiss are the primary individuals that should be held responsible for my writing this PhD. Mark was not only the most amazing mentor anyone can ever wish for, but also the main reason why I came to EIPE in the first place. No matter how harsh he was when commenting on my ideas, he always managed to make me feel excited about thinking harder and about the whole business of doing academic research. I sat three years in a row in his history of economics seminar just out of pure pleasure (the very last three years he ever held it). Now that I come to think of it, I don't entirely understand why he tolerated me and my endless babbling so gracefully and patiently. I am glad I had the chance to tell him in person how grateful I am to him.

From Uskali, I have learnt not only about philosophy but also about all sorts of potentially relevant stuff for life. I am sure I fell short of his radically high standards of clarity and precision in all I have written, yet thanks to his critical eye and feedback, I have ended up at least being able to say what I mean clearly enough to get a PhD. Much more important to me than all the philosophical discussions, knowledge, skills, and generous support for research funding, which Uskali has selflessly shared with me, is the fact that throughout the years he has made me feel like part of his family—an extended academic family, that is. He knows very well how thankful I am for all he has done for me.

I have experienced something similar with Julian. I value his guidance, his care for my project, all the discussions about philosophy, economics (and science in general), his seminars, the reading groups, workshops, and so on. But what I really value the most is that rather than merely becoming another supervisor, he became (probably entirely against his better judgement) more of a good friend to me. Crucial parts of my dissertation were discussed, revised, and probably even written during the many after-jazz-concert sessions we had at the old Lantaren-Venster and other music venues. The huge intellectual debt I have to Julian should be perfectly recognizable to anyone who reads my monograph and knows his work.

The influence of these three guys on my thinking has been so significant at so many levels that, contrary to traditional academic disclaimers, I would feel quite tempted to say that most of the remaining mistakes in my dissertation might very well be theirs.

In addition to my supervisors, Caterina Marchionni is the person who has helped me the most with developing and finishing this project. She has read carefully and repeatedly all of my chapters. She probably knows many, if not all, of my arguments by heart. On top of that, she often attended seminars and conference sessions in which I would present different versions of my work. It felt almost as if she kept giving me a chance over and over again just hoping to be positively surprised one day. I know I have not yet managed to impress her, but I definitely cannot imagine finishing this thesis, or even being a philosopher, without her continuous unconditional support, trust, and friendship. My whole decision to move to Rotterdam would have still been totally worthwhile, even if it were only because there I got to meet Caterina.

Many people have read and commented on my work or have contributed to the completion of this thesis in one way or another. Some of them might not even remember doing so, but I do.

Frank Hindriks has contributed a great deal to my philosophical development, yet perhaps much more significant has been his contribution to my survival in the Netherlands during the first years. He kindly provided me with shelter whenever I faced housing problems (which curiously happened more than once). My first discussion ever about the existence of a Dutch soul was with Frank. He claimed that humans in general have no soul. He is now one of my best friends and, the more I know him, the more I am convinced that he is the very "black swan" of his own generalisation.

Since the first day I arrived to EIPE, Jack Vromen received me with a huge smile and a welcoming hug. He became a sort of professional tutor to me in all thinkable and unthinkable academic respects. Little did he know that he was going to be dealing with me for so long. I appreciate very much all his advice and support.

Attilia Ruzzene and François Claveau were my main partners in the studies of causality. Attilia is one of the most easy-going and attentive listeners I have ever encountered. She is incredibly sharp even (or specially) when she seems a bit spacey. François is like a super-athlete of intellectual activities; he can do whatever he wishes seemingly with no effort. I worked on an article with him (chapter 4 in this dissertation) and during that process I discovered that he is also extremely patient and tolerant. I believe he is physically unable to swear and perhaps even to think of a curse word. I sincerely wish I could give back at least a small portion of all that I have learnt from both of them.

Clemens Hirsch was my PhD brother. Among many other things, we shared the taste for Herzog's films. We use to speculate that Herzog's book on the making of *Fitzcarraldo* could be interpreted as describing the writing a PhD dissertation. In fact, in one of our most pessimistic moments, we thought that the title of that book, *Conquest of the useless*, would be perfect for a PhD thesis. Well, I did it! I finalised my conquest, dude!

Ana Bobinac is my soulmate from another life. She is also one of the best economists that I know in person. She introduced me to the highly policy-oriented area of health economics, and when she was not spending time indiscriminately smiling around or rolling her eyes up, she shared with me a lot of useful insights about economics and her well-grounded understanding of economists' practices in relation to policy-making.

Jaakko Kuorikoski and Aki Lehtinen are my current colleagues in Helsinki, yet I know them and have been inspired by their work almost since I started thinking of my thesis. They have become really close (and, contrary to all expectations, rather talkative) friends of mine. One of my personal projects is to manage to surprise them one day with some pretty cool philosophical idea that they had not thought of already.

More people who have directly contributed to my dissertation or to make my life in Rotterdam and Helsinki a fascinating and joyful experience are Kwela Hermanns, Paolo Palamiti, Mary Robertson, Anil Divarci, Joshua Graehl (there is only one TooL!), Daniel Vargas, Tyler DesRoches, Sine Bagatur, Pedro José Llosa (truly one of my masters of life and also one of the very best writers I know), Melissa Vergara (my Colombian sister!), Gerdien van Eersel, Johanna Thoma, René Lazcano, Tom Wells, Julie D'Hondt (she housed me and helped me out many times during the first years), Menno Rol (he kindly wrote the extremely indispensable summary in Dutch that can be found at the end of this book), Viktoria Edlund and Martin Skullerud (they housed me for two splendid seasons in Oslo), Till Düppe, Alessandro Lanteri, Job Daemen, Georgios Papadopoulos, Emrah Aydınonat, Michiru Nagatsu, Marion Godman, Alessandra Basso, Samuli Pöyhönen, Emilia Westlin (to you not only thanks, but also sorry for any inconvenience), Päivi Seppälä, Ilmari Hirvonen, Carlo Martini, Manuela Fernández Pinto, Judith Favereau, Tomi Kokkonen, Matti Heinonen, Till Grüne-Yanoff, Petri Ylikoski, Chiara Lisciandra, Miles MacLeod, Pekka Mäkelä, Raul Hakli, Rogier De Langhe, Arthur van der Laan, Ticia Herold (she was always so kind that I thought she was in love with me), Fred Muller, Frans Schaeffer, Tim de Mey, Conrad Heilmann, Constanze Binder, María Jiménez Buedo, David Teira Serrano, Jesús Zamora Bonilla, Federica Russo, Gustavo Marqués (and all the members of CIECE!), Adolfo García de la Sienra, Leonardo Espinosa, Sergio Castrillón, Yesmith Sánchez, Jose Cañada, Álvaro Muñoz, Carlos Llamas, David Gambarte, Ian and Liisa Bourgeot, and the Locus Publicus (so many good ideas by many EIPEople have been conceived and deeply discussed in that place).

I also received helpful comments and criticisms on different sections of this thesis from Marcel Boumans, Deirdre McCloskey, Bert Leuridan, Koen Swinkels, Bradley Turner, Lorenzo Casini, Mary Morgan, Harro Maas, David Colander, Wade Hands, and Roger Backhouse.

I had the opportunity to help developing and editing the *Erasmus Journal for Philosophy and Economics*. I spent easily a third of my time on issues related to the journal and it was entirely worth it! A highlight of that job is that I got to interact with an outstanding team of co-editors: Tyler DesRoches, Tom Wells, François Claveau, Joost Hengstmengel, Philippe Verreault-Julien, and Willem van der Deijl. To be part of such team has been a privilege and a pleasure. My best wishes to the current and forthcoming teams of editors.

Every visit to Mexico, no matter how brief, always got me refilled with love and energy. There is not enough space to mention here all my beloved family members (my New-Year's-Eve gatherings with family range from 60 to 80 people), but I would like to say thanks to my parents and my siblings (consanguineous or not), Araceli, Luis Tomás, Ariel, Angel, Sony, Sonia, Laura, Fausto, Rodolfo, Javier, Rapé, Carlos, Josué, Jonathan, Adrián, and El Robert. I have no words to express how much you all mean to me. It has been not easy to live most of the time so far away from all of you.

Finally, I gratefully acknowledge the financial support that I have received from CONACYT (México), Erasmus University Rotterdam, CIMO (Finland), and TINT Centre of Excellence in the Philosophy of the Social Sciences (Finland).

LMF
July 2016, Helsinki, Finland.

PART I
THE ART OF ECONOMIC POLICY MAKING

CHAPTER 1

TOWARDS A METHODOLOGY OF ECONOMIC POLICY MAKING: THE POSITIVE-NORMATIVE DISTINCTION RECONSIDERED

Until very recently, philosophy of economics had little to say about the particular practices and methodology involved in the process of policy making. Admittedly, there is a large amount of philosophical literature studying the scientific standards of economics *indirectly* motivated by policy issues or by the idea that policy should be based on "reliable" scientific knowledge. Likewise, there is a vast literature on how to evaluate whether concrete public goals are "desirable" for society, and on the development of methods for measuring and aggregating social preferences and wellbeing. So the words "policy" and "practical relevance" actually appear in numerous investigations in philosophy of economics (though sometimes only in the title, abstract, or concluding remarks of the texts). Still, few studies engage with a detailed philosophical analysis of these notions or with specific methodological aspects of how economic science is, can, and should be used for the achievement of concrete policy goals.

The relative deficiency of methodological research on the process of economic policy making has been already pointed out by a few authors. David Colander has commented extensively on this problem (see Colander 2001; and 2010) and argued that methodologists do not focus on the process of policy making because they have inherited a theory-biased attitude currently dominating the economics discipline as a whole.[1] To Roger Backhouse, the issue actually "presents methodologists with an opportunity", since they "might be able to help bridge this gap through helping economists understand better the processes whereby economic ideas are applied to policy" (Backhouse 2004, 188).

The lack of research on the processes whereby economic ideas are applied to policy, or what I refer here as the methodology of economic policy making, can be partially explained by the fact that philosophers of economics have implicitly accepted a standard understanding of the positive-normative dichotomy when developing their research agendas. There is no consensus among philosophers or economists on the precise meaning or significance of

[1] Colander is a noteworthy and exceptional case, since he has been urging for and working on developing a methodology of economic policy making—or what he calls 'the lost art of economics'—for about 20 years already (see, e.g., Colander 1992; and Colander 1994). I will comment on Colander's approach below.

this dichotomy. Quite the opposite, debates about the positive or normative nature of economics are rather unsettled (see, e.g., Blaug 1992, chapter 5; Dasgupta 2005; Hausman and McPherson 2006; Kincaid, et al. 2007; Douglas 2009; Hands 2009; Putnam and Walsh 2011). But regardless of the ongoing disputes, the focus of authors interested in philosophy of economics shows a clear segregation onto two general areas of research along the lines of the positive-normative dichotomy.

Broadly put, during the more or less 30 years of philosophy of economics as a self-standing field, there is, on the one hand, a branch including authors mainly interested in the methodology and epistemology of economic science (or of 'positive economics', even if probably no author would be keen to use that label anymore). This research focuses on topics such as, the nature of economics, the scope and methods of the science, problems of theory appraisal, representational relations between abstract economic theory and reality, the nature and types of economic modelling, the methods of causal inference that are proper to economics, the nature of economic explanations, and so forth. On the other hand, there is another branch in philosophy of economics that includes authors mainly interested in moral and political philosophy, and who would often explicitly label most of the topics they deal with as part of 'normative economics', e.g., topics such as justice, welfare, equality, income distribution, poverty, and other economic issues with a significant bearing on ethical concerns.

As historians of economics know well, the modern positive-normative distinction is the outcome of a corrupted interpretation of what was originally intended as a threefold distinction: scientific economics, "normative" economics, and the "art" of economic policy making (see, e.g., Mill 1844 [1830]; Keynes 1917 [1890]; Machlup 1978; Colander 1994). This threefold distinction was based on three different aims of research, namely 1) to find and establish scientific laws, 2) to evaluate the desirability of socio-economic goals, and 3) to help bring about concrete economic outcomes in the real world. During the 20th century, accounts of the distinction switched from a methodological discussion about different forms of enquiry to a debate about whether there could be pure statements of fact and whether scientific knowledge could be value-free. In particular, after Milton Friedman's (1953) essay on the methodology of positive economics, the evaluation of the desirability of goals and the policy making were conflated into a single branch and labelled "normative economics", so that these two branches could be more easily contrasted as a whole with the distinctive features of scientific "positive economics" (Friedman 1953, 4). The positive-normative distinction has since then been commonly presented as a dichotomy between a positive branch of economics dealing with value-free (objective) factual statements and a normative branch studying value-laden (subjective) practical judgments.

Without attempting to resolve all open issues in the current debates over the positive-normative distinction, in what follows, I revise the history of the distinction to single out a few aspects that are relevant to understanding why economic policy making is not generally recognised as a *distinct* branch of economic enquiry. In addition, the interpretation of the distinction that I put forward in this chapter highlights some of the potential topics of research towards a well-developed methodology of economic policy making.

In section 1.1, I present the standard version of the positive-normative distinction commonly held by economists and identify three of its main shortcomings. In section 1.2, I introduce some key elements of the classical threefold version of the distinction that have been chronically ignored, and which I will subsequently use as the basis for my proposal. In section 1.3, I comment on the problem of trying to characterise statements according to the standard version of the positive-normative distinction. In section 1.4, I argue that endorsing an interpretation in line with the classical threefold distinction makes clear which elements of the process of economic policy making have remained unexplored from the philosophical and methodological perspectives. In section 1.5, I briefly revise David Colander's suggestions for a "methodology" of economic policy making. And finally, as an illustration, in section 1.6, I present a study in the methodology of economic policy making, which I argue arrives at some mistaken conclusions precisely as a consequence of confusing the branch of scientific economics with that of economic policy making.

1.1. The standard version of the dichotomy

The standard interpretation of the distinction between positive and normative economics was made popular by Milton Friedman in his 1953 article on "the methodology of positive economics". Most of Friedman's essay is devoted to discuss the "methodological problems that arise in constructing the 'distinct positive science' [of economics]" (Friedman 1953, 3). But first, to make clear the difference with non-positive issues, in the initial pages he offers "a few remarks about the relation between positive and normative economics" (p. 3).

The key distinction in Friedman's account is about *types of statements*. Accordingly, positive statements are said to have two fundamental characteristics: first, they can be appraised "objectively", which to Friedman means that they are "to be accepted or rejected on the basis of empirical evidence" (p. 6); and second, they can be "in principle independent of any particular ethical position or normative judgements" (p. 4), which is usually interpreted as their being value-free. Once positive statements are characterised in this way, then by elimination any statement which could not be evaluated empirically or independently of value judgements is to be classified as normative.

Establishing a strict dichotomy between two types of statements allows Friedman to distinguish then between, on the one hand, positive economics devoted to investigating the economic world "as it is" by formulating and testing positive statements; and on the other hand, normative economics concerned with investigating non-positive claims, or as he puts it, "questions of what ought to be done and how any given goal can be attained" (p. 4), which are taken as value-laden claims which allegedly cannot be empirically investigated.

Notice that Friedman here puts together into the normative branch both "questions of what ought to be done", i.e., evaluation of goals, and questions of "how any given goal can be attained", i.e., evaluation of means. This point is important because commentators (concerned with other aspects of the essay) tend to take for granted that in Friedman's view policy-related statements are supposed to (or can) be part of positive economics in as much as they can be empirically supported. The fact is that when one considers Friedman's definitions, he explicitly keeps questions about policy making *out* of the positive branch, most likely because they are commonly not independent of value judgements in the way he intended. Yet the main idea he wants to convey is that all prescriptive and all policy making issues are ultimately *dependent upon* to the positive science:

> Normative economics and the art of economics […] cannot be independent of positive economics. Any policy conclusion necessarily rests on a prediction about the consequences of doing one thing rather than another, a prediction that must be based—implicitly or explicitly—on positive economics (Friedman 1953, 5).

Thus "policy conclusions" from the art of economics *rest upon* the positive science, but they are not part of it. Admittedly, at the end of the day it is rather unclear, in Friedman's essay taken as a whole, whether statements about how goals can best be attained—such as policy recommendations—should be understood as having a positive or a normative status (this issue will be discussed in more detail below in section 3). For the purposes of this chapter it should suffice that, at least in principle, Friedman thought of policy making as being part of what he labelled ‘normative economics’. This interpretation explains, for instance, why he offers the following proviso: “There is not, of course, a one-to-one relation between policy conclusions and the conclusions of positive economics; if there were, there would be no separate normative science” (Friedman 1953, 5). Why would he offer this clear qualification if he thinks that policy conclusions are part of positive economics?

It can still be argued that given the specific purposes of Friedman’s methodological essay, the conflation of ethical and policy issues into the “normative” branch of economics might not be highly problematic. After all, his explicit goal is to analyse the methodological aspects of the positive branch of economics. However, his interpretation of the distinction as a dichotomy has permeated without misgivings into the discussions of virtually all later commentators. After Friedman’s influential essay, the common way of understanding the distinction has been to take positive economics as dealing with objective (descriptive) factual statements, and normative economics as dealing with (prescriptive) statements involving value judgements, but with no straightforward account about where exactly policy-making issues are supposed to be located.

Friedman’s presentation of the distinction can be especially problematic if taken as a categorical and rigorous distinction between economic statements referring to phenomena of a different nature. Without proper qualifications, the standard interpretation of the distinction could suggest a strict dichotomy between the realm of facts and the realm of values (see Blaug 1992, 112-114).

Unfortunately, it is precisely a bold and unqualified version of Friedman’s distinction that has become the standard version spread among the economics profession as the core of the debates after the 1950s (see Hands 2009). Textbooks of economics certainly present the distinction in a standard and unqualified way (see, e.g., Samuelson and Nordhaus 1989, 9, 12; Begg, et al. 1994, 11; Baumol and Blinder 1997, 2, 16; Mankiw 1998, 26, 32; Parkin 2000, 12-13; Taylor 2003, 36). Not surprisingly, it is also the ruthlessly polarised version of the dichotomy that is usually attacked and rejected by critics of the distinction (e.g., Putnam 2002; Kincaid, et al. 2007; Putnam and Walsh 2011).

There are at least three aspects of the standard interpretation of the positive-normative distinction that are problematic:

1) It suggests a dichotomy between the scientific and ethical branches of economics, leaving no distinctive place for the specific methods and practices related to the endeavour of economic policy making. Yet, as I will argue, the

process of policy making deserves separate methodological consideration. A reinstatement of some essential features of the classical threefold distinction can contribute to a more accurate understanding of the peculiarities of economic policy making.

2) It implicitly conflates different pairs of "opposite" notions, namely positive/normative, objective/subjective, value-free/value-laden, descriptive/prescriptive, as if they could serve as unequivocal criteria for demarcating between the realm of facts and the realm of values. Yet, as will be clarified below, there is no clear way of achieving a definite twofold distinction on the basis of such criteria.

3) It presupposes that a demarcation about types of statements is necessary as an essential foundation for a meaningful typology of distinct economic endeavours. As I will argue, a distinction among different types of enquiries requires neither a definite distinction of types of statements nor a strict distinction between the realms of facts and values. Positive economics, normative economics, and economic policy making are distinct endeavours mainly because of their aims and uses of evidential support are different.[2] I will elaborate on these three (related) issues respectively in the following three sections.

1.2. The classical threefold distinction

The origins of the positive-normative distinction in economics can be traced back to the notions of 'science' and 'art' that were widely in vogue among classical economists since the beginning of the 19th century. John Stuart Mill in particular presented the difference between these notions in the following terms:

> Science is a collection of *truths*; art a body of *rules*, or directions for conduct. The language of science is, This is, or, This is not; This does, or does not, happen. The language of art is, Do this; Avoid that. Science takes cognizance of a *phenomenon*, and endeavours to discover its *law*; art proposes to itself an *end*, and looks out for sneaks to effect it (Mill 1844 [1830], 88-89; emphasis in the original).

[2] Admittedly, it might be the case that distinct types of aims somewhat relate to distinctive linguistic presentations. Statements *could* certainly be typified in various ways, for instance, according to their logical form. An example of this is Fritz Machlup's characterisation of statements, to be commented below (see section 1.3). My argument in relation to this point is simply that a distinction of different branches of economics need not (and should not) be based primarily on a typology of statements, but rather on their different aims and the procedures proper to such aims.

Mill's characterisation of science and art is based on two different types of aims: 1) investigating "truths" about phenomena, and 2) putting forward "rules" or "directions for conduct". Mill uses the expressions "language of science" and "language of art" to illustrate an implication of the distinction, however this linguistic aspect is not meant as a crucial criterion—when considered in the context of his essay—for distinguishing between science and art. Any difference in the linguistic mode or in the logical form of propositions can be understood as a symptom (rather than a cause) of the fact that the deeds of science and the deeds of art are different in the first place.

It may be the case that statements of science (axioms, premises, hypotheses, conclusions, and the like) are commonly presented in the indicative mode (e.g., "A is a B"), while statements of the art (prescriptions, advice, rules of conduct, and the like) are frequently presented in the imperative mode (e.g., "C should be D"). Since the aim of science has to do with discovering and describing phenomena, then no wonder that the indicative mode turns out to be (pragmatically) well suited to expressing its assertions. And since the aim of art has to do with prescribing rules of action, then it is also no surprise that the imperative mode is often used to express such rules. Yet these linguistic uses are contingent consequences of the difference in aims.

As I will argue further below, the linguistic form of propositions cannot alone determine what belongs to science and what to art. For now, the underlying idea to be taken and reconsidered from Mill and the classics is that *describing* "the laws" of phenomena is a fairly distinct endeavour from *prescribing* which practical ends to pursue and how to attain them. And aiming at different goals suggests the need of using different procedures.

A reading of the distinction on the basis of *distinct aims* is somewhat reinforced when Mill describes science and art in relation to causes and effects. To Mill a science is mainly concerned with the study and classification of *causes*, whereas an art is concerned with the classification and production of *effects* (Mill 1844 [1830], 117-118, n11). Hence, the goal of science is to uncover and describe "the laws" of the causes of phenomena, and the conditions under which such causes operate. Whereas the goal of art is to propose which effects should be attained, and to delineate precepts for action to achieve them.

Once more, the account so far is well known by historians of economics. But there is an essential aspect that usually eludes most commentators of Mill's account. After distinguishing between science and art in general, Mill makes a further distinction between two kinds of art which in turn have different aims (Mill 1874 [1843], 6.12.2). Being types of art, both are *prescriptive* and both are *concerned with the study of effects*, yet both focus on effects from entirely different perspectives.

Following Mill (1844 [1830], 95; 1874 [1843], 6.12.2), on the one hand, there is a type of art whose main goal is to define ends and evaluate their desirability (which commonly, but not necessarily, involves a strong reliance on ethical or moral considerations). Let us call this: Mill's *art of setting goals*. On the other hand, there is an art whose main aim is to define "maxims of policy" for bringing about given desired effects based upon the causal knowledge obtained from one (or many) of the sciences. Let us call this: Mill's *art of determining means*. The first type of art is concerned with defining and choosing a practical end to go after, whereas the second type of art deals with how a particular end could best be achieved or brought about in reality (in a concrete context) given the knowledge one possesses about its causes and the relevant background characteristics of the target situation.

A crucial difference between the *art of setting goals* and the *art of determining means* is thus that the former is based for the most part on some form of assessment of reasons (including non-empirical reasons) about *which end* should be followed, whereas the latter is based on an inquiry into *which means* are the most effective in producing a *given end*, and its results can in principle be entirely grounded on empirical evidence. But then, one may ask here, if the art of determining means can be based on empirical (causal) knowledge, then how is it different from the science? Is it not the aim of science to investigate causal relations precisely in order to find out what works and what does not for the production of intended effects?

I will elaborate on the differences in more detail below, but the essential contrast between Mill's 'science' and 'the art of determining means' can be broadly put as follows: to find out what are the causes of phenomena, science requires methods of causal inference that trace in isolation one cause to its effects, e.g., Mill's "methods of experimental inquiry" (see Mill 1874 [1843], 3.8); whereas to bring about concrete desired effects, the art of determining means "traces one effect to its multiplied and diversified causes and conditions", and thus requires a more pluralistic methodology, one that "brings together from parts of the field of science most remote from one another the truths relating to the production of the different and heterogeneous conditions necessary to each [desired] effect" (Mill 1874 [1843], 6.12.5). To analyse and understand a causal relation in isolation (in *science*) is a different aim, and requires different procedures, from tracing all the causes and heterogeneous conditions (in *the* *art of determining means*) that are relevant for bringing about a desired effect in a concrete situation.

In general, no strict distinction between *facts* and *values* is required whatsoever for Mill's account to hold. The basic demarcation is entirely *methodological*: three different purposes—either discovering causes, or setting desirable goals, or finding effective means for producing concrete desired effects—require distinct procedures and distinct appraisal criteria.

In 1890, John Neville Keynes published *The scope and method of political economy*. In addition to its many merits, Keynes's book offers a clear illustration of how common Mill's understanding of the three kinds of enquiries—science, the art of setting aims, and the art of determining means—was among economists up to the turn of the century.

Keynes's version of the distinction is essentially a rephrasing of the classical interpretation. Still, he can be credited with the first straightforward exposition of the distinction as an explicit characterisation of economics into *three* distinct types of research that he labelled: 'positive economics', 'normative economics', and 'the art of economics', which in effect map on Mill's notions of 'science', 'the art of setting goals', and 'the art of practice'.

Firstly, a positive discipline in general "may be defined as a body of systematized knowledge concerning what is" (Keynes 1917 [1890], 34). The aim of such a positive science "is the establishment of uniformities" (p. 35), such as "the law of supply and demand, the Ricardian law of rent, Gresham's law, and the like" (p. 36n). The positive branch of economics is then concerned with the theoretical development of economic knowledge and such a development consists in formulating theoretical hypotheses, "theorems of pure reason", and claims about economic causal relations. Accordingly, the traditional scientific activities of postulating, testing, revising, or refuting causal claims are all part of positive economics.

Secondly, a normative discipline could be defined "as a body of systematized knowledge relating to criteria of what ought to be, and concerned therefore with the ideal as distinguished from the actual" (pp. 34-35). In Keynes's terms, the aim of a normative enquiry is "the determination of ideals" (p. 35). Normative economics is then concerned with the ways of determining ideal economic goals, for which the main grounds of evaluation are moral precepts. Ethical judgements in normative economics consist in the evaluation of the social merit (or desirability) of particular economic effects, for instance: deciding whether inflation and unemployment are good or bad for society, or whether the consequences of the free trade of a particular commodity are desirable or not, or deciding which group within a society deserves to be, or should not be, affected by a particular tax policy (see Keynes 1917 [1890], 32-33).

Finally, the art of economics can be defined "as a system of rules for the attainment of a given end" (p. 35). The art is also 'normative' in so far as it aims at prescribing rules for action. However, as an endeavour, the art aims at "the determination of maxims or precepts by obedience to which given ends may be best attained" (p. 32). Hence, the standards for evaluation in the art of economics are instrumental (rather than ethical). Practical recommendations are to be evaluated in terms of the effectiveness of specific practical strategies with regard to the production of concrete (already defined and agreed upon) desired effects. From the point of view of the art of economics, the ethical evaluation of practical ends is taken as a given, and the focus is on the evaluation of the effectiveness of a specific means to achieve a given end.[3]

1.3. On the criteria of demarcation in the standard approach

As mentioned in the first section of this chapter, by the 1950s economics textbooks were already oblivious of the classical threefold distinction and of its original interpretation. Moreover, during the second part of the 20th century, the modern positive-normative dichotomy based on a distinction between types of statements became entangled with other alleged strict dichotomies such as: is/ought, facts/values, objective/subjective, descriptive/prescriptive (see Blaug 1992, chapter 5; Hands 2009). The conflation among all these notions in the standard positive-normative distinction has made obscure the original meaning and intention of the distinction in its classical threefold version. As I will argue in this section, a careless mixing up of these notions has strongly contributed to shaping the views not only of the endorsers of the standard version of the dichotomy, but also of the detractors of any positive-normative distinction.

Hume's guillotine, that is, "the principle that only factual statements can follow from exclusively factual statements" (Black 1970, 24), was originally coined by philosopher Max Black as a warning in logic against deducing nonfactual statements—connected by 'ought' or 'ought not'—from premises containing only statements about facts—connected by 'is' or 'is not'.

[3] The notion of 'effectiveness' is a key and distinctive aspect in the art of economic policy making. It is also one of the central concepts studied in the present dissertation. I explore its definition in chapter 2, yet all of the remainder chapters can be considered as different attempts at characterising and illustrating the forms and conditions of effectiveness.

The principle is clearly a restatement of David Hume's observation that one should be cautious to not infer "ought" from "is" (see Hume 1975a [1739], III.1.i). The image of a guillotine serves well to illustrate a severe logical slash between factual and non-factual statements, meaning that no 'ought' statement can logically follow only from 'is' statements (see Black 1964). However, the principle has hitherto received many different interpretations that go far beyond the formal distinction between normative and factual statements in logic.[4]

As Mark Blaug has pointed out, in economics the distinction between positive and normative has been often tied to Hume's guillotine "implying as it does a watertight logical distinction between the realm of facts and the realm of values" (Blaug 1992, 113). In Blaug's view, the positive-normative distinction coupled with the is-ought strict divide has been further conflated with the following pairs of "antonyms" and used as the typical list of criteria to decide whether a proposition is either positive or normative:[5]

positive	**normative**
is	ought
facts	values
objective	subjective
descriptive	prescriptive
true/false	good/bad

According to this interpretation, a statement is *positive* if it is about "what is", refers to facts, is objective, is descriptive, and can be true or false. Whereas a statement is *normative* if it is about "what ought to be", refers to values, is subjective, is prescriptive, and cannot be given a truth value on empirical grounds, but rather can only be assessed on the basis of ethical reasons (see Blaug 1992, 112-113).[6] The problem with this understanding of the distinction

[4] The is-ought divide (or Hume's guillotine) can be connected to—and it is sometimes mistakenly referred to as—Hume's fork. However, the two notions refer to quite different distinctions motivated by Hume's ideas. While the guillotine refers to the problem of deducing 'ought' from 'is', Hume's fork refers to a distinction between "Relations of Ideas" and "Matters of Fact" (see Hume 1975b [1777], IV.1): statements about ideas are *analytic*, *necessary*, and can be known *a priori* of any experience; while statements of fact are *synthetic*, *contingent*, and can only be known *a posteriori*.
[5] The following list describing antonyms related to the positive-normative distinction and Hume's guillotine is taken from Blaug 1992, 113.
[6] Blaug's account is not meant as a position that is explicitly held among economists. He simply suggests that when economists defend the positive-normative distinction as a strict dichotomy between statements of facts and statements of value, they usually employ one or several of these pairs of antonyms as the demarcation criteria (see Blaug 1992, chapter 5).

is that there is no unambiguous way of characterising statements as purely positive or purely normative using the pairs of antonyms above as criteria of demarcation.

To see this, take the following generic propositions about relations between events A and B, as presented by Fritz Machlup (1978, 430):

Positive:	If A, then B; that is, B is the effect of cause A.
Normative:	B is good; that is, you ought to get (strive for) B.

It is not difficult to find apparently uncomplicated instances of these generic propositions, take for example:

Positive:	Taxing a commodity causes a reduction in its consumption.
Normative:	Smoking is bad for society, and should be avoided.

The first proposition 'Taxing a commodity causes a reduction in its consumption' is an example of what most economists would take as a positive statement. The claim posits that there is a causal relation, and thus it allegedly refers to a matter of fact, it is descriptive, it can be true or false, but perhaps more importantly: to be positive it should be objective and value-free in Friedman's sense. Is this really the case?

Recall that according to Friedman, the two requirements for a statement to be positive are: A) that they can be evaluated on empirical grounds, and B) that they can be assessed independently of any value judgement (see Friedman 1953, 4-6). In the example at hand, whether it is true that 'taxing a commodity causes a reduction in its consumption' can indeed be decided on the basis of empirical evidence (e.g., on the basis of historical data, natural experiments, statistical studies, case studies, expert opinion, and so on). Of course, the conventions upon which the available empirical evidence is interpreted as *adequate* and *decisive* for supporting the validity of a scientific statement are inescapably based on value judgments of sorts.

Richard Rudner's (1953) original version of this argument goes as follows: whenever the truth of a scientific hypothesis is to be evaluated on the basis of empirical evidence, a subjective judgement has to be made about how much evidence is proper and sufficient for accepting or refuting the hypothesis. Since there is no such thing as absolute certainty, there is no way of setting an ultimate evidential benchmark for the empirical appraisal of any posited hypothesis. Still in practice scientists "as scientists" do accept and reject hypotheses after considering the available empirical evidence. Thus, whenever a scientist decides that a certain amount of evidence is enough to establish the truth of a hypothesis, such decision is not based on empirical grounds, but it is the result of a subjective evaluation. This in turn could be an evaluation at least indirectly founded on ethical considerations, e.g., considerations about how bad the consequences of an erroneous appraisal of the posited hypothesis could be (see Rudner 1953; Sen 1981; Dasgupta 2005; Douglas 2009).

According to Rudner's argument, then, the notions of objective/subjective as a demarcation criterion for statements do not necessarily coincide with a demarcation based upon the notions of value-free/value-laden. In Friedman's version of the distinction positive statements must be 'objective', meaning empirically grounded; but they also must be value-free. However, if Rudner is correct, value considerations are inescapable in the empirical evaluation of scientific (positive) hypotheses. And hence it would seem that the set of scientific statements satisfying Friedman's positive standards *strictly understood* would be empty.[7]

Consider now the proposition 'Smoking is bad for society'. As presented above (and according to Friedman's criteria) this is a normative statement. It is a claim about the "desirability" of smoking (i.e., whether it is good or bad) for society, and accordingly it is to be founded explicitly on certain value considerations, it is subjective in as much as the reasons behind the judgement are not only empirical (e.g., reasons determining what "bad for society" means), and it is also prescriptive given that undesirable ends are to be avoided. But can these features be taken as consistent and definite criteria for normative statements?

Without going too deep into the existing debates on how different theories of morality account for the assessment of value judgements (e.g., utilitarianism, Kantian ethics, virtue ethics, and the like),[8] there is one particular issue

---

[7] It can be argued that Friedman's (1953) standards for positive statements actually leave enough room for subjective judgement (see Mäki 1986, 132-136; and Mäki 2009, 109-112). In this section I intend to illustrate an (admittedly) extreme position in which the criteria for positive statements include that they be entirely independent of value judgements, which is precisely the type of extreme belief or hope that Rudner's argument was intending to refute. For a much more balanced and refined position on the interactions between scientific empirical standards and values, see Koertge 2000.

[8] In this chapter I am taking for granted that there is no substantive theory of morality that can offer an unambiguous appraisal of all ethical propositions. There seem to be qualms and

concerning the standard characterisation of normative statements on which I will focus here: It is not at all clear that the ethical appraisal of so-called normative claims such as 'Smoking is bad for society' is (or can be) done independently of any empirical consideration. Even if the appraisal of the claim is to be mainly grounded on ethical reasons or morals norms, there is some "factual" information about smoking that is included into the social assessment of the desirability of smoking. For instance, claims such as 'smoking causes cancer', 'smoking causes coughing', 'smoking causes stained teeth', 'smoking affects the health of smokers and non-smokers in such and such way', and so on and so forth, which are all empirically testable claims.

Deliberations about whether an economic outcome is desirable for society usually consist of an evaluation of the reasons presented in support the posited end. These reasons could be non-empirical, e.g., based on mystical, or traditional, or irrational beliefs, yet especially in relation to social ends commonly some empirical reasons take part into the evaluation of the desirability of ends. That unemployment is not desirable for society is not only justified by arguing that it generates evil and suffering to society, but by offering reasons referring to empirical facts and causal implications of unemployment. That education is good for society is only true given a certain set of reasons, which may range from non-empirical claims such as 'education enriches the human soul' to empirical ones such as 'education increases economic income'.

I do not mean to claim here that the evaluation of normative statements necessarily implies or requires empirical considerations. I am just making the weaker suggestion that—especially in relation to socioeconomic ends—many value judgements are inferentially connected with (implicit or explicit) empirical claims. In this sense, supporting or rejecting value propositions often commits one to a certain stance about other propositions which can be empirical (see Brandom 2007; Peregrin 2012). Understood in this way, normative statements in economics can be, and often are, not only ethically but also empirically grounded. Differently put: empirical considerations can be, and are often, included among the reasons given to support the desirability a posited social outcome.

counterexamples affecting all existing proposals (see Baron, et al. 1997). I also take for granted that any kind of value judgement is justified or supported by reasons, and thus that the method for appraising a value judgement (broadly conceived) consists in evaluating how "good" its reasons are. These ideas are essentially in line with James Rachels's proposal of a "minimum conception of morality" (see Rachels 2003).

Indeed, the ethical evaluation of whether smoking is bad for society has historically been based on a huge amount of empirical research (see, e.g., Doll and Hill 1950; Levin, et al. 1950; Doll and Hill 1954; White 1990; Brandt 2007; Peto, et al. 2012). And it seems that the same has happened in most debates on the desirability of social outcomes traditionally discussed in economics, such as: unemployment, poverty, inflation, recessions, economic growth, economic competition, trade liberalisation, financial crises, and the like. But if normative statements are value-laden as well as (at least to a certain extent) empirically based, and positive statements are also empirically established and (to a certain extent) value-laden, then what exactly is the difference between positive and normative statements?

One of the main points I want to make in this chapter is that there is no clear-cut way to base the positive-normative distinction on a strict demarcation about types of statements. Thus instead of trying to answer the previous question, I will show in the next section why I think this is an unnecessary question, at least from a methodological point of view. In this section, my claim so far is simply that *positive and normative statements cannot be unambiguously told apart using the fact/value, the objective/subjective, or the value-free/value-laden dichotomies as demarcation criteria*.

The project of demarcating positive and normative statements becomes even more convoluted when one tries to find a place in the purported standard dichotomy for yet another sentential form that Machlup labels 'instrumental'. In such a case, events A and B relate as follows (Machlup 1978, 430):

Instrumental:	If you want B, A will get it for you; that is, A will be the means for the end B.

And thus—still following Machlup—the three kinds of statements phrased in terms of the same example used before would look like this:

Positive:	Taxing a commodity causes a reduction in its consumption.
Normative:	Smoking is bad for society, and should be avoided.
Instrumental:	Given that avoiding smoking is a desired goal in a society and given that taxing causes reductions in consumption, an effective way to achieve this goal is to tax the consumption of cigarettes.

Do instrumental statements of the form “If you want B, A will get it for you” belong to positive or to normative economics? Are they empirically or ethically grounded? Are they about the realm of facts or about the realm of values?

The statement “An effective way to reduce smoking is to tax the consumption of cigarettes” can indeed be treated as a scientific hypothesis. As such, it requires evidential support in order to be accepted or refuted; therefore, Rudner’s argument also applies to it. Subjective evaluations are ultimately what determines how much evidence is enough to take the instrumental hypothesis as a true claim or, in other words, to trust that the recommendation is reliable. In this sense, the instrumental claim is not value-free.

Are the values involved in the appraisal of instrumental claims ethical values? In so far as instrumental claims presuppose claims about the desirability of ends (e.g., that smoking is bad for society), and in so far as ethical values were involved in the deliberation about the desirability of the proposed end, then instrumental claims presuppose ethical considerations about the “goodness” of the intended practical end. Nevertheless, the appraisal of the instrumental claim need not be an ethical evaluation of the given desired end, but only an evaluation of the effectiveness of the proposed strategy to achieve the desired end.

In as much as hypotheses refer to strategies that are to be implemented in (and will affect) real and concrete situations, the evidential standards for accepting or rejecting the claims—as Rudner also explicitly suggests—are dependent on the practical consequences that the posited strategies are expected to have for society. Thus if taxing smoking would have no expected disturbing effects for society apart from its intended effect on smoking, then the evidential standards for accepting the hypothesis could be set at a relatively low level (since from a subjective perspective there would not be much to lose if the hypothesis is mistaken). In contrast, if it is the case that there are many potential negative by-products that can obtain in a society after the implementation of the taxing policy, then the evidential standards for the appraisal of the hypothesis had better be high (since, again from a subjective perspective, there would be much more to lose if the hypothesis is mistaken).

So an instrumental hypothesis of the form ‘If you want B, A will get it for you’ can be empirically tested. Supposing that B is a desired end, empirical evidence can be used to support the claim that A is effective for the achievement of B in a concrete situation. Nevertheless, values enter the process of evaluation as a subjective decision has to be made about which and how much evidence is enough to take the instrumental claim as true. Furthermore, the values involved in such subjective decision can include ethical considerations about the practical relevance (or practical risks) related to a mistaken assessment of the instrumental proposition. Thus what type of statements instrumental statements are supposed to be? Again, there is no clear way to strictly classify instrumental statements as either positive or normative by using the criteria of the standard distinction. Instrumental statements are value-laden (to a certain extent), but also objective (to a certain extent) and empirically testable.

To sum up, the fact/value distinction, the subjective/objective distinction, the descriptive/prescriptive distinction, the value-free/value-laden distinction, and so on, cannot properly serve as criteria for unambiguously classifying distinct types of statements as belonging to positive or normative economics. Moreover, value considerations and empirical considerations enter the evaluation of all sorts of hypotheses at different stages and in different ways depending on the particular purpose of the hypotheses.

It is noteworthy that most attacks against the positive-normative distinction commonly consist in attacking (or finding counterexamples to) one or a few of the pairs of antonyms related to the is-ought divide (Blaug 1992, 116-118). Philosophical debates have especially focused on the value-free/value-laden dichotomy and are often phrased in terms of the possibility or impossibility of a value-free science (also called the value-free ideal). Then criticisms against the value-free ideal usually proceed by arguing that subjective or ethical judgements (e.g., Rudner 1953; Douglas 2009) are unavoidable elements in the process of establishing scientific statements, and thus that a strict fact/value dichotomy is simply not tenable (e.g., Putnam 2002).

A common feature among the approaches against the positive-normative distinction is to argue that there are no pure value-free statements or that there is no definite fact/value dichotomy (see, e.g., Sen 1981; Solomon 2001; Putnam 2002; Dasgupta 2005; Kincaid, et al. 2007; Putnam and Walsh 2011). But notice that the arguments I have reviewed in the course of this section are actually against the following two ideas: A) that it is possible to single out positive statements that are entirely independent from value considerations, and B) that ideally science should consist in investigating and establishing value-free empirically grounded statements. To show that A and B are false is problematic for a positive-normative distinction only if such distinction is to be founded on a strict categorisation of individual statements.

In general, most participants in the debate presuppose that a distinction between positive (scientific) and normative (non-scientific) endeavours in economics is necessarily dependent upon the possibility of a demarcation between value-free and value-laden statements. But one can also hold a view outside this 20th-century standard framework that tends to either support strict demarcations or reject any distinction that is not definite enough. For instance—and this is another central claim I want to make in this chapter—one could hold a distinction similar to the classical distinction which, as explained in the previous section, need not be dependent on a strict demarcation of types of statements or on any harsh disconnection between facts and values.

1.4. A DISTINCTION ABOUT TYPES OF RESEARCH

As noticed above, focusing on classifying *statements* rigorously as a basis for the positive-normative distinction is the wrong approach. In this section, I argue that one can get a better picture of economic practice by focusing on distinct *types of research*, and then on what roles all sorts of propositions play in each of them. The impossibility of rigorously demarcating statements as belonging to either the realm of facts or the realm of values need not be in conflict with characterising distinct types of endeavours in economics.

My proposal amounts to restating the classical threefold distinction while avoiding any emphasis on a strict categorisation of statements in isolation, and rather focusing on particular methodological features related to each of the three classical kinds of endeavour, namely: establishing scientific knowledge, defining and setting economic goals, and designing and assessing economic policy prescriptions. The basic features that characterise each of these types of enquiry can be elucidated by asking on which basis, how, and why their respective posited hypotheses are investigated or—to be more specific—by answering the following three questions: a) What are the testing criteria for assessing the hypothesis under investigation given its semantic content? b) What is the logic of the testing methodology that conforms to such criteria? And c) what is the connection between the purposes of the investigation and the hypothesis under investigation? By looking at these features (on which I will elaborate below), the investigation of hypotheses in economics can be classified at least into three broad branches that can be labelled: 'economic science development', 'ethical or normative economics', and 'economic policy making'.

This understanding of the threefold distinction allows for the body of knowledge comprised in each of the branches to interact and overlap without getting into any big muddle. Empirical claims, subjective claims, ethical claims, prescriptive claims, and so on, can all occur intermingled whenever a hypothesis is being assessed within each of the three types of enquiry. What distinguish the inquiries are the distinct aims and the use of different assessment criteria and methodologies. In as much as one keeps clear the particular purposes of each investigation, it will be clear which type of economic endeavour one is doing. To see this more in detail, consider again the three claims previously presented as an example:

1. Taxing a commodity causes a reduction in its consumption.
2. Smoking is bad for society, and should be avoided.
3. Given that avoiding smoking is a desired aim in a society and given that taxing causes reductions in consumption, an effective way to achieve this goal is to tax the consumption of cigarettes.

These propositions are three different types of hypotheses, but not in the standard-distinction sense, i.e., in terms of some underlying or more fundamental fact-value dichotomy. They are different types of hypotheses viewed from a methodological perspective: the content of these claims differs such that rather distinct criteria and testing procedures are required for their assessment.[9]

The first proposition is about *causation*, the second is about the *goodness* of a social feature or outcome, and the third is about the *effectiveness* of a policy recommendation. The specifics of what they are about—even if the distinction need not be absolutely precise—are relevant to understand the differences in the research required for their evaluation. Causality, goodness, and effectiveness are notions that call for different assessment criteria. As a consequence, the methodologies for testing either a causal relation, the goodness of an end, or the effectiveness of a practical strategy present significant differences in their logic.

A causal hypothesis states that there is a genuine causal connection (rather than a spurious relation) between the posited causal relata. Criteria for genuine causation are usually embedded in the specific account of what causation amounts to (e.g., regularities, probabilistic relations, counterfactuals,

[9] It might seem that I intend to trick the reader by stating on the one hand (in the previous section) that it is not possible to strictly and unambiguously demarcate types of statements, and on the other (in the present section) that there are different types of hypotheses. To be clear, a threefold distinction of economic endeavours is compatible with a distinction of types of hypotheses, as long as the criteria for classifying hypotheses have nothing to do with any fixed intrinsic nature of facts and values. As it will become clear, the classification of the three hypotheses considered here is rather pragmatic, conventional, and mutable.

mechanisms, manipulations),[10] however the general logic involved in the appraisal of causal hypotheses is well captured by Mill's methods of causal inference (see Mill 1874 [1843], 3.8). Broadly put, most methods of causal inference consist in controlling or fixing all causal factors relevant to the causal outcome of interest, except of the posited cause under evaluation, so as to shield or isolate the operation of the causal relation from all potential sources of disturbance. In the philosophy of economics, this method has been referred to as the method of isolation (see Mäki 1992a). As will be made clear in subsequent chapters, the general gist of the method of isolation is analogous to the logic of empirical methods for causal inference commonly employed in economics, i.e., methods that control for (as many as possible) sources of error to test for causality (see Reiss 2008).

In contrast, a hypothesis of the second type states the goodness of an outcome and thus requires a criterion that allows for the discrimination between good and bad. Again, the specifics of the criterion depend on the particular theory of goodness that is adopted. For instance, whether good is what brings happiness, or good is what is compelled by moral categorical imperatives, or good is what is virtuous, and so on. For the purposes of this chapter, the methodology employed to evaluate hypotheses about the goodness of things can be broadly characterised as an argumentative method: a method in which a set of reasons are put forward in an argumentative way so that they are considered together to either justify or undermine the posited hypothesis (see Rachels 2003). Moral hypotheses can certainly be true or false (see Lynch 2009, chapter 8), yet the logic of the argumentative method employed in their appraisal is nothing like the method of isolation or the inferential methods of controlling all potential sources of error which are used in order to assess the truth of causal hypotheses.

Finally, a hypothesis of the third type states that, given a desired practical end, there is a strategy that is effective in order to attain it. This is the type of hypotheses that concerns the "art" of economic policy making. Just like in the previous cases, these hypotheses can be true or false. So what is the proper appraisal criterion? And what are the proper methods to test them?

Since the main content of the present dissertation can be taken as a book-length elaboration on the answers to these two latter questions, here I will just briefly introduce a few distinctive characteristics of the policy making

[10] In subsequent chapters, I will elaborate in more detail on the different accounts of causality and their relation to practical relevance (chapter 3), as well as on how causal accounts implicitly suggest a particular testing criteria (chapters 5 and 6).

endeavour. One part of economic policy making consists in using the results of science to ground reliable recommendations, hence causality also plays an important role in the process of policy making. Nevertheless (as I will explore further in the following chapters), the main criteria for the truth of a policy hypothesis is not primarily meant to discriminate between genuine causation and spurious relations, but rather between *effective* and *ineffective* practical strategies. Even if a policy hypothesis (or policy recommendation) involves usually a causal concept, testing for causation is not the same as testing for effectiveness. Causation can be tested in relation to no specific practical end, whereas effectiveness always has to be tested in relation to a given and concrete intended outcome.

As J. S. Mill also suggested (1874 [1843], 6.12.5), the art of policy making brings together causal knowledge from as many scientific sources as necessary, and takes into account all disturbing factors that are relevant to the case at hand with the exclusive aim of securing the production of the intended practical end. As I will argue (especially in chapters 3, 4, and 6), the methods proper to testing the *effectiveness* of concrete practical strategies involve a different and often much more complex logic than the one characterising the appraisal of scientific causal claims or the evaluation of hypotheses about the goodness of social ends. For instance, instead of isolating a single causal relation to evaluate whether it is causally efficacious, policy making requires (at least in principle) the inclusion and consideration of all relevant causal and non-causal factors that can affect in any way the occurrence of the intended practical outcome (as well as potential by-products).

I take it that the difference between testing causal relations (the scientific endeavour) and assessing the goodness of practical ends (the normative endeavour) is not controversial. Furthermore, as suggested in the introduction of this chapter, economists and philosophers of economics have no problem in studying separately the methodology of the scientific branch of economics (that investigates causal relations) and the methodology of the moral branch of economics (that investigates the goodness of socioeconomic outcomes). The claim I made at the beginning of this chapter—that there is not much systematic research on the methodology of policy making—can now be understood more precisely as the statement: there is a shortage of methodological research on the branch of economics that investigates hypotheses about the effectiveness of practical economic prescriptions and about the specifics of its appraisal procedures.

1.5. Towards a methodology of economic policy making

During the last two decades, David Colander has undoubtedly been the most forceful advocate of reviving the classical threefold distinction in order to better understand economics (Colander 1992; 1994; 2001). He has explicitly endorsed J. N. Keynes's distinction of three kinds of enquiry, namely the science, the ethics, and the art of economics. According to Colander, the art of economics is in modern terms "applied economics" (1992, 20), which he sees as "the engineering branch of economics" (1994, 36). Following the classical interpretation, the art of economic policy making "relates the lessons learned in positive economics to the normative goals determined in normative economics" (1992, 20).

There is no elaboration on how Colander exactly interprets the classical distinction, yet he openly favours Keynes's threefold distinction over Friedman's strict dichotomy, essentially because the latter leaves out "what most economists do, which is applied policy economics" (Colander 1994, 35). Yet, the most important reason for distinguishing among the three branches "is that different methodological rules apply to each" (p. 36). This is very much in line with my argument in the previous sections, i.e., that enquiries with different aims require different methods.

An important difference with my view, however, is that Colander thinks that the methodology of the art of economics consists of a set of general normative rules. In his view, the methodology of economic policy making amounts to the stipulation of "methodological rules for the art of economics [...] meant as rough guides to approaching issues of applied policy" (1994, 41). Then he offers a list of six of such methodological rules (pp. 41-45):

1. Try not to violate the law of significant digits.
2. Be objective and use the rational person criterion to judge policy.
3. Use the best economic theory available.
4. Take in all dimensions of the problem.
5. Use whatever empirical work sheds light on the issue at hand.
6. Do not be falsely scientific and present only empirical tests that are convincing to you.

Colander's approach is indeed an interesting step towards trying to show why stressing some methodological aspects of the distinct art of economics might be expedient. Supplementary methodological studies of actual cases of economic policy making (such as those presented in chapters 3 and 6 of this thesis) could serve to explore in more detail the actual practical conveniences or inconveniencies of following his rough proposed guides for applied policy. Nevertheless, Colander goes no further than presenting this list of six rules. Overall, there is no further explanation in Colander's approach about the empirical or philosophical foundations of the rules proposed. There is no suggestion about how they follow from any specific alleged characteristic of the art of economics in general. The only justification for them seems to be Colander's lengthy experience as an economic practitioner and common sense.

Consider, for instance, the rule of "take in all dimensions of the problem"; it sounds intuitively like the right thing to do, but there is no elaboration on why and how it should be done. Moreover, why is this rule stated as if it were specific to the art of policy making? Would not it be recommendable as well to take in all relevant dimensions in the scientific and the normative endeavours? Or is there perhaps a special way of doing so in relation to the art of economics? Again, a more detailed analysis of actual cases of economic policy making focusing on how theoretical claims are enriched with policy experience and then used along the relevant dimensions of a practical problem can help to get more meaningful methodological guidelines (see chapter 5 in this thesis).

Rules like "use the reasonable person criterion" or "do not present false or biased results" seem to be somewhat common sense recommendations adequate for any branch of economics (or any other scientific discipline for that matter). In any case, Colander does not offer any explanation of why these particular rules would be more important for policy purposes than for scientific or normative purposes.

The rule of "use the best economic theory available" is also a bit ambiguous as it stands. To which type of investigation does the appraisal of economic theories correspond? The rule seems to imply that economic policy practitioners should seriously consider the problems of theory choice or theory appraisal before using any theoretical claim as a basis for practical recommendations. But those are problems to be considered by the scientific (or positive) branch of economics, as Colander himself is eager to argue:

> In the art of economics one accepts the general laws and models that have been determined by the profession and one tries to apply the insights of economic models to real-world problems. Applied policy economics has nothing to do with testing a theory; it has to do with applying the insights of that theory to a specific case (Colander 1994, 36).

To be fair, Colander's approach has two clearly established goals. One is to make the developers of economic theories in the positive branch of economics more conscious about the character and the practical needs of applied economists and economic policy makers. And the second (perhaps the main objective of his overall project) is to modify economics education so as to reshape "the way economics is taught", since the "appropriate methodology for the art of economics is much broader, more inclusive, and far less technical than the methodological approach for positive economics that underlies current teaching practices" (Colander 1992, 23).[11]

My own proposal can be considered as an elaboration on Colander's methodological approach. The following chapters are an attempt to offer a more substantive philosophical account of the ideas, methods, and interactions involved in at least two issues about policy making hinted at by Colander. To illustrate the issues, I rephrase two of Colander's guidelines in the form of research questions:

1. How can policy makers take in all the relevant dimensions of a policy problem?

2. How can empirical evidence be properly used to shed light on the policy issue at hand?

I answer the first question by exploring the different roles that causal relations play in economic policy making. Different types of causal claims confer different sorts of practical power. When causal relations are studied and established in the scientific branch of economics, they usually follow a method that isolates the relations from all other potential disturbing causal factors. Thus, in principle, all relevant dimensions are controlled for in order to achieve a reliable causal inference. In economic policy making the way of "taking in all relevant dimensions of the problem" follows a different procedure: it consists in finding a way of evaluating which are all the potential disturbing causes in relation to a particular concrete effect, and then proposing a way of understanding the concurrent interactions of all the relevant causes for the production of the effect. Causal pluralism and some forms of causal modelling offer some philosophical answers to these issues. Understanding the meaning of so-called causal generalisations and their relation to particular instances is also relevant in order to understand how causal knowledge can be used to effectively bring about effects in the real social world (I elaborate on these topics in chapters 2, 3, and 4).

[11] On the reform of economics education, see also Colander 2005; 2008; Ñopo and Colander 2007.

I answer the second question by exploring the distinct roles that empirical evidence can play in supporting policy claims, in contrast to the particular role evidence plays in supporting causal claims. The main point to be stressed in this part of my project is that the evidence that properly supports a causal generalisation (in the positive branch of economics) is not necessarily sufficient (or the same) as the evidence that is required to support a particular policy recommendation. It is not any general type of evidence what is required if one wants to shed light on a policy issue, but the "right kind" of evidence in relation to the specific dimensions of the particular policy goal (I argue for this in chapters 5, 6, and 7).

The issues of the methodology of economic policy making that I explore in this thesis are not exhaustive at all, but they are central for understanding the particular features of the process whereby scientific economic knowledge is (and can be) used to generate reliable economic policy recommendations.

1.6. How keeping the distinction matters in methodology: an illustration

There are already some few cases in the literature which can be taken as instances of methodology of economic policy making (e.g., Milberg 1996; Klemperer 2003; Angner 2006; Swann 2006; Reiss 2008; Grüne-Yanoff 2011; Hansen 2011; Svetlova 2013). A fairly good illustration occurs in a relatively recent article by Anna Alexandrova (2006) on the use of economic knowledge for economic policy design. I will focus on this case because it shows how one might arrive at potentially mistaken methodological conclusions by conflating the scientific branch of economics with the art of economic policy making. And so this example illustrates why it is helpful to do philosophy of economics with a proper interpretation of the threefold distinction in mind.

Alexandrova's article is meant to challenge the next two claims: Firstly, that "economics, when successful, is necessarily done according to the method of isolation" (2006, 174); and secondly, that "models in economics supply claims about tendencies" (p. 174). In order to reject these claims, she uses as a case study "perhaps the most successful case of the application of game theory to date" (p. 174): the construction of the Federal Communications Commission (FCC) spectrum auctions.

Alexandrova starts by briefly describing the general ideas behind the method of isolation in economic science, with reference to the work of Uskali Mäki (see, e.g., Mäki 1992a; 2004a; 2004b; 2006). According to Alexandrova's interpretation, contemporary economic theory assumes that economic modelling "help us isolate important portions of social reality" (p. 174), and that the "essential elements isolated in an idealized model can help us to explain actual real world phenomena" (p. 174). Additionally:

> This view is often supplemented with (something like) John Stuart Mill's account of political economy as a science of tendencies, studying the behaviour of economic agents in the absence of disturbing factors and then combining these tendencies for the purposes of explanation and prediction (Alexandrova 2006, 174).

After such introductory remarks on the method of isolation in economics, Alexandrova proceeds to argue against the first claim she sets up at the beginning of her article. Her claim is that economics "when successful" is not necessarily done according to the method of isolation. And to show this, she uses the case of the FCC spectrum auctions as an example. The main steps of her argument matter so I summarise them in what follows.

First, she describes auction theory—a branch of game theory—as the kind abstract economic theory that is usually considered as providing the relevant economic knowledge for policy recommendations in order to design an actual auction. The basic theoretical auction model in relation to the FCC auctions case is the so-called "private value auction with two bidders" (pp. 174-176). In theory, this model postulates two players competing for a good in a first-price sealed-bid auction. They both know their own valuation of the good in question, but not the valuation of the other player, and the bids are simultaneously submitted.

This theoretical setting is complemented with some additional standard game theoretical assumptions about distribution of values, probabilities, and information. The question to be investigated by the economic theorists is then: how should the players bid? The mathematical optimization of the problem leads to a solution in which each player bids half of his own valuation of the good. As Alexandrova puts it: "Since both players are maximizing their expected utility given their beliefs about actions of the other, we have a Bayesian Nash equilibrium" (p. 175).

Such is the way in which game theorists build up their scientific knowledge: by developing "a typology of auctions (open sealed bid, second or first price, with or without reserve price, etc.) and types of information known to bidders (e.g., whether or not they receive it from the same source)" (p. 176) in order to solve theoretical games. In other words, game theorists build up their theories employing the method of isolation.

According to Alexandrova a rather different picture accounts for the case of the FCC design: "In 1993, the FCC commissioned the help of game theorists and experimentalists to design an auction that would privatize licenses for bands of electromagnetic spectrum in accordance with Congress's requirements" (p. 186). The aim of the team was to design "an actual institution that would satisfy the government's constraints" (p. 186). Alexandrova argues that the method followed by this team of experts "was different from the method of isolation" (p. 186). Instead of using the method of isolation, the designers "sought to find out facts about one material system as a whole" (p. 186). The method they actually used is depicted as follows:

> The process by which these facts were established was a mixture of modelling and experimentation, where the former provided only indications of possible causal relations and the latter revealed a material implementation of the desirable effects within the environment of the auction (Alexandrova 2006, 186).

Alexandrova's account of the whole process of designing the auction is intended to point out that: theoretical results from the positive branch of economics could serve as a basic material to start working with, but when "it comes to advising a policy maker, the most difficult task is to figure out the implications of all these different theoretical results for the actual task at hand" (Alexandrova 2006, 188), and then a mixture of alternative "highly applied, local, and empirical" procedures had to be used (p. 191) instead of the method of isolation. Alexandrova then hastily jumps to the conclusion that such a case study constitutes "a methodological benchmark" because it presents an instance of successful economics that was not "necessarily done according to the method of isolation" (p. 174).

There is a confusion here that springs from taking the process of designing the FCC spectrum auction as a methodological benchmark in the positive branch of economics, when it is actually a clear example of the complex methodology proper to economic policy making. The whole point of restating the classical threefold distinction is precisely to make clear (among other things) that the method used when aiming at assessing theoretical hypotheses is quite different from the methods required when the aim is the formulation and evaluation of economic policy recommendations. The latter case is precisely what Alexandrova's study supports and illustrates.

The method employed in the FCC policy making process actually echoes J. S. Mill's speculations about the different types of knowledge, in addition to scientific knowledge, that the method proper to the art of practice requires:

> No one who attempts to lay down propositions for the guidance of mankind, however perfect his scientific acquirements, can dispense with a practical knowledge of the actual modes in which the affairs of the world are carried on, and an extensive personal experience of the actual ideas, feelings, and intellectual and moral tendencies of his own country and of his own age. The true practical statesman is he who combines this experience with a profound knowledge of abstract political philosophy (Mill 1844 [1830], 109).

As long as everybody involved in the methodological analysis of economics clearly understands the distinction between the positive branch of economics and the art of formulating practical recommendations, then it is possible to consistently hold that the method of isolation is the method of successful economic theorising, and to hold simultaneously that a more complex procedure involving the consideration of the very same theoretical insights, plus additional empirical data, experience in doing policy, non-economic knowledge, heuristic procedures, and so forth, is required in cases of successful performance of the art of economics, just as it seems to have been the case in the design of the FCC spectrum auctions.

1.7. Concluding remarks

The lesson to be learnt from the classical approach is that it is not necessary to come up with an ultimate and pure characterisation of statements as positive or normative, or as value-free or value-laden, or as objective or subjective. The classical distinction was essentially a distinction about *types of inquiry* with different aims, rather than about statements with distinct natures. Furthermore, the three types of inquiry have enough distinctive methodological features, so that it is useful to retain a properly qualified version of the classical threefold distinction—instead of the two-fold standard version—rather than to dismiss the distinction altogether, as some authors have suggested (e.g., Sen 1981; Solomon 2001; Putnam 2002; Dasgupta 2005; Kincaid, et al. 2007; Putnam and Walsh 2011).

The issue at stake is to properly identify the purpose of the investigation that is being carried out: whether it is an investigation into economic causes and the ways and conditions in which they obtain (positive economics); an investigation about the goodness or desirability of economic outcomes to society (ethics of economics); or an investigation into the most effective ways for attaining a given economic end (economic policy making). A proper methodological appraisal of economics requires this identification, since any posited hypothesis (either causal, moral, or instrumental) is to be evaluated using different criteria and testing procedures depending on the type of inquiry in which they occur.

Once the distinction between the three branches of economics has been interpreted in the proposed way, it becomes obvious that philosophy of economics has not been paying enough attention to the methodological study of the art of economic policy making. Some basic ideas about how this methodology of the art could be developed and what type of issues it is supposed to engage with have been also put forward. There seems to be a huge and complex (and interdisciplinary) area which is still highly unexplored from a methodological point of view: investigations about how to use scientific economic knowledge to make given normative goals happen, i.e., investigations on the art of economic policy making.

PART II
CAUSAL KNOWLEDGE AND ECONOMIC POLICY MAKING

INTRODUCTION TO PART II

The idea that knowledge about causal relations can be exploited for the attainment of practical goals seems uncontroversial: knowing that oxygen is a cause of fire can be used to produce or to extinguish fire; knowing that exposure to asbestos is a cause of cancer can be used to provoke or to prevent cancer; knowing that excess demand is a cause of price increases can help determine whether it is lucrative to sell or to buy certain commodities; or knowing that education is a cause of economic growth can be used to guide private or public policy.

This insight about the practical usefulness of causal knowledge has been commonly taken for granted in the literature on causation. For instance, Nancy Cartwright has explicitly justified her probabilistic account of causality with a formalised version of this basic intuition. In "Causal laws and effective strategies" (Cartwright 1979), she writes that there is "a natural connection between causes and strategies that should be maintained: if one wants to obtain a goal, it is a good (in the pre-utility sense of good) strategy to introduce a cause for that goal" (p. 431). By 'the pre-utility sense of good' she means "effective" (p. 420). In line with this view, the way causal knowledge can be exploited for practical purposes is condensed as follows: if 'X *causes* G' is true, then bringing about X "will be an effective strategy for G in any situation" (p. 432).

Bert Leuridan, Erik Weber, and Maarten Van Dyck (2008) have labelled this position the "standard view on the practical value of causal knowledge" (p. 298). According to them:

> [The standard view stands for] the thesis that the practical value of causal knowledge lies in the fact that manipulation of causes is a good way to bring about a desired change in the effect (Leuridan, et al. 2008, 299).

As these authors also point out, many philosophers working on causation have simply taken the standard view as given (pp. 298-299). It is one of the key motivations behind the scientific aim of distinguishing *genuine* causes from *spurious* ones, since claims about genuine causation "are needed to ground the distinction between effective strategies and ineffective ones" (Cartwright 1979, 420). Consequently, this view is (somewhat silently) embedded in most of the literature on methods for causal inference (see, e.g., Simon 1954; Spirtes, et al.

1993; Scheines 1997; Glymour 1997; Pearl 2000; Hoover 2001; Shadish, et al. 2002; Steel 2004; Guala 2005).

The standard view is often illustrated with some evocative example involving smoking and lung cancer roughly as follows: suppose that "S is a variable that codes for smoking behavior, Y a variable that codes for yellowed, or nicotine stained, fingers, and C a variable that codes for the presence of lung cancer" (Scheines 1997, 188). Suppose further that the actual causal structure among these variables is: S *causes* C, S *causes* Y, but C and Y are not causally related to each other, and all three variables regularly obtain together. The practical relevance of genuine causal relations in contrast to spurious relations is then shown to follow automatically: avoiding smoking would be an effective strategy for avoiding lung cancer and yellowed fingers, but steering clear of nicotine stained fingers (say, by wearing protective gloves while smoking) would be a plainly ineffective strategy for avoiding lung cancer. But is detecting genuine causation all that is required to let us identify effective strategies?

As a result of the common acceptance and use of the standard view in rather abstract terms, little attention has been paid to the details of what it actually amounts to. What kind of practical power is causal knowledge in fact capable of conferring? How exactly is one meant to exploit or make use of causal knowledge for the effective attainment of practical goals in concrete and specific situations? Is it the same kind of practical power present in all forms of causal knowledge? Could there be any general methodological guidelines concerning the use of causal knowledge for practical purposes?

Instead of providing answers to these types of questions, authors interested in causality have primarily directed their attention to some other (no less important) inquiries, for instance, conceptual or semantic questions such as "what does *cause* mean?" (e.g., Russell 1912-1913; Ducasse 1926; Lewis 1973) or "what is the logical form of causal claims?" (e.g., Davidson 1967); ontological questions like "are there causal relations in the real world?" (e.g., Salmon 1980; Menzies 1989; Dowe 2000); and epistemological questions such as "how can one distinguish, find, or learn about causal relations?" (e.g., Simon 1954; Suppes 1970; Spirtes, et al. 1993). Hence a serious philosophical effort to investigate pragmatic questions related to how scientific knowledge is, can, and should be employed to support, guide, or implement practical decisions and policy recommendations is still missing.[12]

[12] In relation to this longstanding gap in the philosophical research, there have been a few authors who have openly suggested that philosophy of science could and should play a much more significant role in the investigation of how science is actually applied and used for

To be clear, I do not want to claim that the standard view is mistaken. It is in principle (and quite obviously) correct, since causal knowledge can indeed be exploited for practical purposes. The problem is that it is commonly presented and discussed at a high level of abstraction—see, e.g., the quotations offered by Leuridan, Weber, and Van Dyck (2008, 298-299)—and thus, in practice, it leaves in the shadow all the concrete features and details of the process by which causal knowledge is and can be used to do things in the world. So the standard view can be taken as a starting point for more substantial and thorough philosophical investigation of the practical relevance of causal knowledge.

In the following chapters, my goal is to pursue such a systematic approach. Although I adopt a philosophical standpoint, a better understanding of the practical aspects of causation should not be of interest exclusively to philosophers, but could also help policy makers engage with the tasks of formulating, evaluating, and implementing reliable policy prescriptions on the basis of available scientific knowledge.

In **chapter 2**, I discuss the notions of 'practical relevance' and 'effectiveness' in relation to causal knowledge. In the standard view the notion of 'practical relevance' is confined exclusively to one of its possible meanings, namely a specific form of intervening power, while in fact causal knowledge can contribute to the attainment of practical goals in many different equally important ways. In particular, I discuss how predictive power and intervening power relate to each other and to causal knowledge. The complexity of the notion of practical relevance becomes clearer by distinguishing between the 'practical potential' of causal knowledge (which depends on its predictive and intervening power), and the notion of 'effectiveness' (which refers to actual accomplishment in relation to concrete practical goals). The aim of this chapter is to argue that practical potential is not sufficient for the effectiveness of causal knowledge.

In **chapter 3**, I discuss the relation between causal pluralism and practical relevance. Causation is a plural notion at several levels: there are distinct theories about what causality is (e.g., those put in terms of regularities, probability raising, counterfactual dependence, physical processes, invariance under interventions, and so forth), and there are distinct causal concepts which refer to different configurations of causal factors within a given causal

---

practical (and policy) purposes; see, e.g., Cartwright 1974; Suppes 1984; Kitcher 2001; Douglas 2009; Mitchell 2009. My position here takes this suggestion seriously and elaborates on its possibilities.

structure (e.g., net cause, component cause, average cause, preventer, and so forth). As a result, a generic causal claim of the form 'X causes Y' can have a variety of meanings and interpretations. Yet the standard view says nothing about how different meanings of a causal claim could relate to different forms of practical relevance. The aim of this chapter is to argue that a proper philosophical disambiguation of the meaning of causal relations should complement any attempt to use causal knowledge for practical purposes.

In **chapter 4**, (co-authored with François Claveau) a case study is used to illustrate the main points of the previous chapter. The OECD research on unemployment in the 1990s was presented to the members of the OECD as an up-to-date scientific study including theoretical and empirical findings from economic science on the causes of, and the effective strategies against, the problem of unemployment. Most of the findings were presented in the form of causal generalisations. It is not clear, however, what the precise meaning of these claims is when they are analysed taking into account the plurality of interpretations of causal claims as presented in chapter 3. more specifically, the example shows (1) that the meaning of causal claims is not always clear when they are presented as guides for policy; (2) that ambiguities in the meaning of a causal generalisation can, in practice, come from several sources, namely: the specific meaning of the causal concept in the claim, the meaning of the causal relata, and the specification of the population of application; and (3) that it is rather unclear from the results of the OECD study which are the proper policy recommendations that can be reliably inferred from causal generalisations about population average effects to be implemented in particular target units.

CHAPTER 2

Causation, practical relevance, and effectiveness

There are several distinctions that have to be made in order to understand the ways in which causal knowledge can be exploited for the achievement of practical goals. As the standard view suggests, the manipulation of causes in order to bring about their effects is obviously one aspect that makes knowledge about causal relations useful for practical purposes. However, the role that causal claims play in the process of designing and implementing effective strategies has more facets in addition to simply allowing potential interventions.

In this chapter, I present a distinction between the *potential* of causal knowledge to be used for practical purposes and its *effectiveness* in relation to concrete strategy or policy implementations. Knowing that "X causes Y" is true indeed grants a certain potential for practical use, yet to infer an effective strategy, further dimensions related to the practical relevance of causal knowledge have to be taken into consideration.[13]

First, the practical potential of causal knowledge has two basic components: predictive power and intervening power; I will elaborate on these two notions in section 2.1. Secondly, the effectiveness of a strategy or policy is dependent upon the practical potential and some additional contextual components; I will discuss the contrast between practical potential and effectiveness in section 2.2. Some concluding remarks are presented in section 2.3.

2.1. Prediction and intervention as practical virtues

According to John Stuart Mill, causation "is co-extensive with the entire field of successive phenomena" because "every fact which has a beginning has a cause" (Mill 1874 [1843], 3.5.1). While elaborating on these ideas, Mill makes a brief remark about the practical relevance of causal knowledge:

> Of all truths relating to phenomena, the most valuable to us are those which relate to the order of their succession. On a knowledge of these is founded every *reasonable anticipation of future facts, and whatever power we possess of influencing those facts to our advantage* (Mill 1874 [1843], 3.5.1; italics added).

[13] Causal claims are among the most basic ingredients of larger pieces of scientific knowledge. Sometimes they are the main target of models, sometimes they are assumptions, and sometimes they are implications of those models. In what follows, as a working hypothesis, causal claims are taken as the most basic bearers of practical relevance.

Reasonable anticipation (prediction) and influence (intervention) of facts to our advantage are indeed two crucial aims for which causal knowledge can be practically relevant. But how exactly are these aims related to causality, and to each other? I call "practical virtues" the qualities that make knowledge useful for practical purposes, so that the practical relevance of causal knowledge can be thought as relying on *at least* two distinct practical virtues:

> **Predictive power**: 'X *causes* Y' has *predictive power* if, provided that the claim is true, the observation of an occurrence or change in X allows a reliable forecasting of an occurrence or change of state of Y.
>
> **Intervening power**: 'X *causes* Y' has *intervening power* if, provided that the claim is true, the production or manipulation of X is feasible, and intervening on X can produce or change the state of Y.

In general, one and the same causal claim 'X *causes* Y' might carry both practical virtues. Nevertheless, there are three differences between predictive and intervening power that are crucial for evaluating the practical relevance of causal knowledge.

First, in many cases where there is no intervening power there can still be predictive power. Suppose 'X *causes* Y' is true and we know it. Then intervening on X can be an effective strategy for affecting Y, but only when it is actually *possible* to manipulate or affect in any desired way the cause X. Arguably, knowing that 'X *causes* Y' is true does not necessarily entail the capacity of producing or affecting X. This means that intervening power is not always granted by genuine causation.[14] Yet, a causal claim without intervening power can still have predictive power, since knowing that 'X *causes* Y' is true could be used to successfully forecast the occurrence of Y, given that X has been observed, regardless of any possibility of intervening on X.

For instance, knowing the causal claims in Newtonian mechanics allows one to meaningfully predict (ceteris paribus) future positions of the planets and

[14] This is the case even if the causal relation 'X *causes* Y' was established as true by means of an intervention on X in accordance with the manipulationist theory of causality (see the following chapter, section 3.1), since modern versions of this account (e.g., Woodward 2003) only require for the truth of the claim that the putative cause X be affected via *some ideal* intervention (fulfilling certain particular requirements) so as to produce a change in Y. However, if the causal claim is meant to guide the attainment of a specific practical goal Y in a real world situation, the intervention required needs to be *feasible* (not only *ideal*), and the *appropriate* one (not only *some*) so as to affect Y in the specific desired way. The difficulty of inferring concrete practical recommendations from abstract and general causal claims is one of the main problems of economic policy making. Chapter 5 below is entirely devoted to discuss this issue.

other celestial bodies, whereas the same causal knowledge does not immediately grant the power to manipulate the putative causes in order to alter planetary positions. Weather forecasting amounts to a similar case: knowing the relevant causal relations about natural phenomena makes it possible to predict (with a certain degree of accuracy) weather changes in the short run, like chances of precipitation, the wind's expected strength, or the possibility of storms, but it does not allow one to intervene or control the occurrence of such phenomena. Similarly, consider a causal claim stating that gender is a cause of the consumption of a particular good. The manipulation of gender so as to control its effect on consumption preferences is, at the very least, extremely complicated; still that would be independent of the significant predictive power the causal claim could still convey, e.g., to the professionals of marketing in relation to the expected sales of their product.[15]

Secondly, intervening power is more dependent upon genuine causation than predictive power. Let me elaborate on this point. Knowledge about genuine causation is not always a necessary requirement for making reliable predictions (even if it can be argued that it is desirable). Think for example of a job interview in which a good score on some standardised test is employed to forecast the future performance of a prospective employee. The employee's abilities are the genuine causes of the future performance, and at the same time they are the causes of the test scores, yet it is the correlation between scores and expected performance—and not any genuine causal relation between them—which is used to justify a decision about whether to hire the applicant or not.

Following this same line of thought, it has been suggested that knowledge of genuine causation is not necessary for attaining practical goals, since spurious correlations can be quite useful on their own (see Leuridan, et al. 2008). However, correlations have to be relatively stable in order to be reliable for prediction. And if stability is important for prediction—say, in the sense that the more stable a spurious relation is, the more reliable the predictions based on it will be—then it can be argued that the stability of a correlation depends precisely on the indirect influence of some (unknown) common cause or of a more complex (unknown) underlying causal structure. If this is the case,

[15] Some methods used to infer causality in the social sciences (like those following the logic of the potential outcomes framework; see chapter 5) restrict their understanding of a 'cause' exclusively to variables that *can be affected* in one way or another: "causes are only those things that could, in principle, be treatments in experiments" (Holland 1986, 954-955). Hence, attributes of a population such as "gender", which cannot be "experimentally" changed, are not considered as causal variables in the first place. Yet, as explained above, knowledge of these attributes end their relations with other variables can convey significant predictive power.

then ultimately genuine causality is (at least indirectly) required for making reliable predictions.

There is still a meaningful difference between the predictive and the intervening power of a stable correlation between X and Y, whenever the underlying common cause or causal structure is unknown. A claim of this sort can still provide an agent with reliable predictive power, even if the agent knows that the relation between X and Y is not genuinely causal. On the same grounds, the fact that the claim is not genuinely causal is enough to undermine any potential intervening power, at least in the straightforward form of: manipulating X is an effective strategy to affect Y. The correlation between test scores and future job performance is stable and can be trusted to yield reliable predictions insofar as the common cause (employee's abilities) has a stable influence on both correlated effects. Any user of the claim about the correlation can make reliable predictions out of it, even without knowing the underlying common causal structure.[16] However, since the relation is not genuinely causal, it is simply not possible for any agent to affect the future job performance by directly intervening on the test scores. This is what I mean by saying: intervening power is more dependent upon genuine causation than is predictive power.

The third difference is perhaps more controversial. There are cases in which causal knowledge can have intervening power, and yet for practical purposes have a very ambiguous or null predicting power. Consider a causal structure in which a putative cause X has an influence on its effect Y through two separate causal routes R1 and R2. An example of this situation is the effect of price changes on the quantity demanded of a commodity: within the relevant causal structure, this causal relation can be decomposed into a substitution effect and an income effect. For so-called inferior goods, i.e., goods that are mainly

[16] The users (clients) of science might be unaware of the underlying common cause or causal structures responsible for the stability of a correlation and still rely on it and use it for prediction. This does not mean that nobody knows about the underlying structures. Correlations can be reliable to a potential user, for example, because a team of experts knows the reasons (the underlying causal structure) that explain why the correlation is stable, and the user simply "trusts" on the results of the research team. This issue is related to the notion of 'epistemic dependence' in social epistemology: the process of analysing data would be too long for any individual to accomplish, thus specialisation occurs as a rational outcome, and the common knowledge goes on developing even if every member of the epistemic communities is knowledgeable only about a piece of it. Everybody's knowledge depends to a certain degree on everybody else's knowledge (see Goldman 2001; Hardwig 1985; 1991). Researchers have to "trust" (to different degrees) that the results of others are well founded (see Andersen and Wagenknecht 2013). Similarly, clients of a science "trust" that scientific results are well grounded, even if they are ignorant of any of the underlying causal structures.

demanded by individuals receiving low incomes, the substitution effect on quantity demanded is inversely proportional to the change in prices, whereas the income effect is directly proportional to the change in prices.

Supposing that the causal relations postulated in consumer theory can be taken to be true, the claim "price changes cause the quantity demanded" can provide reliable intervening power, and yet fail to provide reliable predictive power in the following sense. An intervention on the price level of an inferior good, keeping fixed the influence of prices through one of the two causal routes, either R1 or R2 (e.g., through another intervention), will result in the desired effect on the quantity demanded. Nevertheless, even knowing that the causal claim is true (and the underlying causal structure including R1 and R2), the observation of an increase in the price of an inferior good does not entail any reliable prediction about the way in which it will affect, if at all, the quantity (because the substitution effect and the income effect could cancel each other out, thus resulting in no change whatsoever on the quantity demanded).[17]

A similar situation has been described by Spirtes and Scheines (2003) as an "ambiguous manipulation". Ambiguous manipulations refer to situations in which one can be sure that affecting X will affect Y while at the same time be relatively ignorant about exactly how much or in which direction the effect of an intervention will go. Spirtes and Scheines illustrate this possibility with the following example. Suppose researchers know that the claim 'high cholesterol levels cause heart disease' is true, but in fact there are two types of cholesterol, one that causes heart disease (LDL) and another that prevents it (HDL), and this fact is entirely unknown. In such a case, the interventions (which take the form of specific diets with either high or low cholesterol levels) are actually interventions on the underlying causal factors, but in different unknown proportions. Thus, given the available knowledge, the causal claim has intervening power, since by setting the level of cholesterol in a diet, an effect on heart disease will indeed obtain. Nervertheless, the claim has no predictive power (or at least no definite or reliable predictive power), since the proportion in which the intervention affects the underlying causal factors is entirely unknown, and hence the actual effect of the intervention cannot be unambiguously forecasted.

[17] This particular type of causal structure in which a cause has an influence on its effect through a number of distinct causal paths refers to a distinction between two causal concepts, namely "net" and "component" causes, which will be described in more detail in the following chapter, see section 3.2.

2.2. Practical potential and effectiveness

In this section I start exploring the following idea: To know that 'X causes Y' is true can convey the *potential* to achieve practical goals, and yet not be sufficient to warrant that 'implementing X is or will be an *effective* strategy for Y' in any *concrete* situation. This is a key idea, and in the remaining chapters of this thesis, it will be investigated in more detail, arguing for it from different angles. In the remainder of this chapter, I present a number of conceptual clarifications that will serve as a general framework to better understand the discussion about the practical relevance of causal knowledge to be articulated in the subsequent chapters.

The notions of predictive and intervening power are helpful to understand in virtue of what causal knowledge can *potentially* be employed for practical purposes. Yet, to understand the actual ways in which causal knowledge is (and should be) employed to effectively attain concrete practical ends, the question to investigate is rather about how exactly the practical potential of causal knowledge can obtain in concrete cases. Differently put, the question is this: how can causal knowledge be used in the formulation of strategies and policies such that they wind up being *effective*?

Let us assume that a piece of causal knowledge has intervening power, then the *effective* attainment of a concrete practical goal on the basis of that knowledge requires *not only* an intervention on the putative cause, *but also* that all additional relevant causal elements are taken into account and that they all obtain in the "right" way during the process of policy making. Thus intervening power is part of what gives causal knowledge the *potential* to be used for attaining a desired goal. Yet, for this potential to be *actualised* in an effective way (i.e., to result in the actual attainment of an intended goal), disturbing factors and background conditions should be adequately identified and accommodated. To understand how these additional factors are related to the effective practical use of causal knowledge, let us start by making clear the following notion:

> **Practical potential**: knowledge that X causes Y can be said to have *practical potential* if it can provide either intervening power, predictive power, or a combination of both, to the agents who have that knowledge.

While having this potential is what makes causal knowledge relevant for practical purposes, it is only one of the ingredients that contribute to attaining *effectiveness* (a notion yet to be more precisely characterised later in this section). The role of causal knowledge in the effective achievement of a desired effect is, more often than not, influenced by additional factors or "background conditions". Thus effectiveness depends on the practical potential of causal knowledge and on the state of all background conditions relevant to the concrete situation in which a strategy or policy is to be implemented.

> **Background conditions** refer to any additional factor or condition Z (known or unknown) which can have an effect on the actual attainment of a concrete desired outcome after the implementation of a particular strategy or policy. This can happen by enabling (or constraining) the proper implementation of the particular strategy or by directly (but independently of any intervention on the putative cause) disturbing the occurrence of the desired effect (examples of these factors are: the rules or institutional framework regulating the implementation, the expertise involved, the decision making process, as well as socio-cultural factors, unexpected disturbing social events, and so on and so forth).

That background conditions matter for policy purposes is well recognised by philosophers and scientists. However, a closely related question that is seldom considered in detail is this: how exactly do the background conditions matter for the *effectiveness* of a concrete practical implementation? Causal relations are commonly analysed and understood in the context of a causal structure, that is, a set of additional causal factors Z which have a potential direct or indirect effect on Y, apart from the posited cause X.

The notion of background conditions Z plays an important role in many epistemological accounts of causal inference, since to find or test for genuine causation between X and Y, ideally one would like to keep fixed or under control all additional causal factors (i.e., potential confounders) that can have an influence on Y apart from X. In this way, the causal influence from X to Y can be investigated and computed in isolation. As it was put forward in chapter 1, this method for investigating causal relations—regardless of whether one calls it method of isolation or of abstraction or of controlling for "disturbances" or "potential sources of error"—is ubiquitous in the scientific branch of economics (see Mill 1874 [1843], 3.8; Mäki 1992a; Cartwright 1989, chapter 5; Reiss 2008, chapter 1).

In the process of economic policy making, however, in order to successfully affect or bring about a desired effect in a concrete situation, all the elements in Z that have a bearing on the adequate production of the desired effect need to be brought back into consideration. The effectiveness of a practical strategy or policy in a concrete situation requires a process of de-isolation or concretisation (see, e.g., Svetlova 2013), such that information about all (known) elements of Z that are specifically relevant for the situation at hand is made explicit.

The contrast between the methods for establishing causal knowledge and the procedures for using it for policy purposes will be analysed in more detail at a conceptual level in chapter 5, and at an empirical or evidential level in chapters 6 and 7. For the moment, let me point out that there are two elements in the set of background conditions which can have a significant weight on effectiveness, and which are not usually (explicitly) considered as members of Z in scientific studies on causal inference. These notions require careful attention especially in relation to the effective attainment of practical goals.

> **Degree of proficiency**: the skilfulness, education, training, experience, or any other epistemic feature of the users of the science that can affect the process of deliberation or implementation related to the attainment of a practical goal by employing the relevant available scientific knowledge.

> **Persuasiveness**: the extent to which relevant audiences are convinced about the truth and suitability of a causal claim (or any other piece of knowledge) so as to be used as a basis (or a guide) for practical action.

I shall not go into much detail elaborating on these notions; however it seems worth considering them briefly, if only to understand how they relate to the notion of practical potential and to any piece of causal knowledge.

As can be seen from the description above, *degree of proficiency* is a property of the users of science, rather than of the causal knowledge they use. However, the degree of proficiency of the users contributes (in different ways in different cases) to the effective attainment of intended practical goals. Simply put, whether a certain policy design is the most effective way to achieve a concrete goal or not is independent of whether that policy design is properly or deficiently implemented. It can be said that proper implementation is a precondition for the effectiveness of a policy. The degree of proficiency of the users of scientific knowledge determines the quality of policy implementation, and thus indirectly enables (yet does not secures) that a policy be effective.

The practical potential of causal knowledge and the degree of proficiency of the users of science have a concurrent effect on the effectiveness of a particular strategy or policy, yet they are two different notions. Numerous complaints about the failure of economics in dealing with real world phenomena—and about its alleged lack of practical relevance—could be more properly addressed as failures related to the use (or misuse) of economic science for policy purposes by particular individuals or groups. This is the case, for instance, regarding some accounts on the alleged failure of economics to predict and deal with the recent 2008-2009 financial crisis. Several accounts supporting this accusation concentrate on criticising the methodology and the accuracy of the models used to do economics, rather than on whether particular pieces of economic science were in fact misused or misunderstood by policy makers (see, e.g., Hodgson 2009; Colander 2010; Kirman 2010).

To say that the claim 'X *causes* Y' has practical relevance potential presupposes that there could be somebody who could use the claim to achieve a certain goal or purpose. These individuals are the users and clients of science (see Mäki 2004a): people that use scientific knowledge for different purposes including to offer practical advice to other individuals or to decide about their own strategies for action given particular socioeconomic and epistemic goals. Engineers, economic consultants, governmental policy advisors, medical practitioners, business managers, public administrators, are only some examples. In accordance with this, it should be clear that the effectiveness of a policy depends to a certain extent on the actual abilities and skills of the clients of the science to properly employ the available scientific information for the achievement of their practical goals.

Degree of proficiency is a property of the clients of the science and not of the causal knowledge being employed. And degrees of proficiency can be extremely uneven among individuals. Overall, this is an outstandingly subjective aspect influencing effectiveness, which seems rather difficult to be accurately measured, though it plays a role in almost all real life cases of policy implementation. Yet the main point here is that degree of proficiency has to be explicitly distinguished from the practical virtues of scientific knowledge so that one can separately consider, evaluate, and meaningfully talk about the practical potential of causal claims independently of any further assessment of the qualifications and performance of any user of such causal knowledge.

Persuasiveness is another feature of causal knowledge. The reasons why a causal claim can be (made) more or less convincing are manifold and cannot be exhaustively enlisted. For instance, persuasiveness could be promoted as a direct result of the perceived accuracy and reliability of the claim, but also of considerations of things such as simplicity, tractability, formality, legality, tradition, consensus, authority, and so forth (see, e.g., McCloskey 1983; Mäki 1995; 2004a). Notice that this notion is highly subjective in the sense that what counts as convincing is mainly dependent upon the particular perceptions and predispositions of individuals and their epistemic communities.

Suffice it to say that a causal claim can be persuasive independently of its practical potential. In fact, some causal claims can be very convincing regardless of their being false. Clients of a science are interested in the persuasive potential of scientific knowledge, almost as much as they are interested in its practical potential (see Nelson, et al. 1987; McCloskey 1983). At times, persuasiveness might even be all that policy makers seek in order to push a particular biased agenda (see Milberg 1996; Klamer and Meeham 1999; Stevens 2011). Yet, the points to be emphasised here are 1) that persuasiveness as a feature of scientific knowledge can affect the effectiveness of a practical implementation, and 2) that it should be distinguished from, and not be taken as a surrogate or as a direct indicator of, the practical potential of causal knowledge as it sometimes is.

The notions discussed so far are all related to the practical relevance of scientific claims. Predictive and intervening power largely depend on the posited truth of causal claims, whereas degree of proficiency is mostly a property of the clients of the science, rather than of the causal knowledge that they employ, thus its assessment should be kept separate from any proper appraisal of the practical potential of scientific causal knowledge. The same applies to persuasiveness. While it can be taken as a property of causal claims, it highly depends upon how the claims are perceived (as convincing or not) rather than upon the truthfulness of their causal content.

The standard view suggests that effective strategies follow from true causal knowledge. But effectiveness refers to the extent to which a "strategy" (or policy) contributes to the achievement of a concrete goal under actual circumstances of implementation.[18] Having true causal knowledge is useful for practical purposes, but does not grant effectiveness, for it is only one ingredient of a much more complex configuration.

Using the elements described in this section, the effectiveness of a policy claim can be characterised in function of the practical potential of the causal knowledge employed to base or to infer the claim, its persuasiveness, the degree of proficiency of the agents in charge of the recommended implementation, and the actual state of the remaining of the relevant set of causal background conditions.

[18] This interpretation of the notion of effectiveness is based on the literature on economic evaluation of policy interventions (e.g., Cochrane 1972; Drummond, et al. 1987; Haynes 1999). Nancy Cartwright has also used the term 'effectiveness' with the same meaning in her writings about evidence-based policy (e.g., Cartwright 2009, 2012a, 2012b; Cartwright and Munro 2010; Cartwright and Stegenga 2011).

> **Effectiveness**: degree of success of the actual implementation of a strategy or policy in pursuit of a definite goal under the ordinary circumstances of the case, conditional upon the *practical potential* of the causal knowledge employed, the *degree of proficiency* of the relevant users, the *persuasiveness* of the relevant claims, and all additional *background conditions* involved in the process of implementation.

This is not intended as a precise definition, but simply as a broad characterisation that could serve as a guideline for more detailed and focused investigations of the distinct interrelated aspects that account for the effectiveness of concrete strategies or policies. Effectiveness is a notion that can exclusively be assessed in relation to a given goal. Thus, while the practical potential refers to the prospective predictive or intervening power that true causal knowledge can in principle provide to any agent for a variety of (not yet definitely specified) practical aims, the notion of effectiveness refers to the actual success of a practical implementation in relation to a particular and concrete aim.

2.3. Concluding remarks

The *standard view*—that if 'X *cause*s Y' is true, then manipulating X would be an effective strategy for producing or changing Y in any situation—seems to be an acceptable first approximation of the practical relevance of causal knowledge. Yet it is a highly abstract portrayal of the matter. The relation between a causal claim and practical relevance can take different shapes that involve much more than straightforward intervening power. Similarly, the effectiveness of a strategy or a policy depends on more than the practical potential of a true causal claim; hence true causal claims do not warrant effective strategies (as the standard view can be taken to suggest). A proper approach to the practical relevance of causal knowledge should take into account the distinctions presented above (and probably more). The next step, after recognising that there are several distinct aspects involved in the notion of practical relevance, is to realise that the claim 'X *causes* Y' can have many different meanings as well. The following chapter focuses on how the notion of practical potential relates to distinct forms of causal knowledge.

CHAPTER 3

Causal pluralism and practical relevance

There are several accounts in the philosophical literature proposing that causation is a plural notion and exploring a number of varieties of causal pluralism (e.g., Hitchcock 2003, 2007a, 2007b; Godfrey-Smith 2009; Psillos 2009; De Vreese 2009; Reiss 2011a). According to these accounts, a claim of the form 'X *causes* Y' can have many different meanings. In this chapter, I focus on two forms of causal pluralism, namely: *pluralism about theories of causality* and *pluralism about causal concepts*. The aim is to discuss the difference between the two forms of pluralism and their implications for policy-oriented sciences. Since there are different possible meanings for claims about causal relations, the main conclusion of the chapter is that a proper disambiguation of the meaning of any causal claim is required before using it for practical purposes.

3.1. Pluralism about theories of causality

It is possible to identify at least five main traditional theories of causality available in the philosophical literature. Each of these theories characterises causation in terms of an alternative notion, allegedly more primitive than that of causation, which in turn allows one to define some comprehensive criteria for what counts as causal. The primitive notions employed in the five theories are, respectively: law-like regularities, probabilistic relations, counterfactual relations, physical processes, and potential manipulations. The gist of each of these theories of causation goes as follows:

> **Regularity theory**: X *causes* Y if and only if there is a regular connection between the occurrences of X and the occurrences of Y (see, e.g., Hume 1975a [1739]; 1975b [1777]; Mill 1874 [1843], book 3; Davidson 1967; Mackie 1974).
>
> **Probabilistic theory**: X *causes* Y if and only if the occurrence of X increases the probability of an occurrence of Y (see, e.g., Suppes 1970; Cartwright 1979; Skyrms 1980; Eells 1991).
>
> **Counterfactual theory**: X *causes* Y if and only if both X and Y have occurred and had X not occurred, then Y would have not occurred (see, e.g., Lewis 1973; Swain 1978).
>
> **Process theory**: X *causes* Y if and only if both X and Y have occurred and there is a physical process from X to Y (see, e.g., Salmon 1980; Dowe 2000).

Manipulationist theory: X *causes* Y if and only if bringing about or affecting the occurrence of X brings about or affects the occurrence of Y (see, e.g., Gasking 1955; Menzies and Price 1993; Woodward 1996; 2003).

These theories were (at least originally) intended as accounts of *in virtue of what* a relation between X and Y is considered as causal, and presented in terms of necessary and sufficient conditions. Nevertheless, it is recognised by now that these attempts to capture the 'fundamental nature' of causation with a universal approach are all defective (see, e.g., Hitchcock 2003; Cartwright 2004; Campaner and Galavotti 2007; Reiss 2009a). There is a huge literature that presents and discusses counterexamples designed to challenge either the necessity or the sufficiency criteria for each of the posited universal accounts.[19] As a consequence, the quest for a univocal account of causation has been gradually substituted by a pluralistic view which can be described as follows.

Pluralism about theories of causality is the position that there is no univocal theory providing all necessary and sufficient conditions for causality, but rather each of the general theories might capture one aspect or another (i.e., regular connections, probability raising, counterfactual dependence, processes, or interventions) of the nature and meaning of being causal (see, e.g., Longworth 2009; Psillos 2009).

Accordingly, for a claim "X *causes* Y" to be considered as causal it is sufficient that it fulfils the conditions of at least one of the available general theories of causality. This form of pluralism is entirely compatible with the fact that some relations which one theory of causation takes as causal do not count as causal under some other account. Once the goal of finding the ultimate all-encompassing theory of causation is abandoned and a pluralistic position is adopted, the counterexamples to each of the theories can be regarded as less threatening (see Longworth 2009).

Now let us think of economics for a moment: is it at all clear which theory or theories of causality are endorsed when formulating and testing causal relations in economics? Is it obvious (to economists and to the users of economics) according to which theoretical account the established causal claims in economics should be interpreted? These questions are fundamental since adopting one or another theory of causality to analyse the meaning of particular causal claims can lead to different practical implications.

[19] For instances of discussions about these counterexamples, see the articles included in Sosa and Tooley 1993; and in Collins, et al. 2004.

For instance, if 'X *causes* Y' means that the occurrence of events of type X increases the probability of the occurrence of events of type Y, one can perhaps forecast the *likelihood* of events of type Y after observing an occurrence of X. In contrast, if the meaning of the causal claim is that there is a deterministic physical causal process from X to Y, then one could confidently *produce* Y by triggering X and by ensuring that the causal process is not disrupted.[20] Or one might even be able to *replicate* or *reproduce* the causal process as it would be convenient for attaining a desired practical goal. In the case one also knows that the causal process linking X and Y is invariant to a wide range of interventions, then it might be possible to generate *fine grain variations* of Y (see, e.g., Woodward 2010), or perhaps even to elaborate a *detailed mapping* about how certain precise manipulations of X would generate definite changes in Y.

Again, the story can change a great deal if what one means by 'X *causes* Y' is that had X not occurred, then nor would Y have occurred. Knowing that such a claim is true could be useful to ascribe some *causal responsibility* to factor X, given that one observes that Y obtained.[21] But knowing that an instance of X has been causally responsible for a particular occurrence of event Y provides almost no grounds to infer much about future occurrences of Y. Counterfactual dependence was in fact originally considered mainly appropriate for cases of so-called 'singular causation', cases in which X and Y represent particular occurrences and not general types of events (see Lewis 1973; Sober 1985; and Eells 1991, chapter 6). In this sense, it is not obvious how knowledge of claims about singular causation could be practically useful, say, to generate *reliable forecasting* of future occurrences of Y (but see also the discussion in chapter 5, section 5.1).

The fact that there are several theories that can be used to interpret causation suggests a practical consideration: it seems recommendable that scientists and the users of scientific causal knowledge are clear about the theory (or theories) of causation used to analyse the causal claims that are to be subsequently employed for policy purposes. Still, even if everybody agrees on the theory of causality being used, there are some further distinctions that

[20] Uskali Mäki has suggested an adaptation of a process theory of causality to analyse the notion of the market as depicted in Austrian economics, see Mäki 1992b.

[21] To adjudicate causal responsibility seems to be a significant practical goal in some sciences, such as history, law, archaeology, and the like, in which an event that has already occurred has to be established as either causal or not. For a detailed elaboration on the particular roles of counterfactual causation in law and history, see Hart and Honoré 1985; and Reiss 2009b.

have to be made explicit in order to properly disambiguate the meaning of causal claims. To see this, let us consider the second type of pluralism.

3.2. Pluralism about causal concepts

This form of pluralism—sometimes called "conceptual causal pluralism" (see, e.g., De Vreese 2009)—is different from the previous in that it is not mainly concerned with the variety of theories of *in virtue of what* a posited relation from X to Y is considered causal. Instead, it is concerned with the *different ways* in which the causal influence from X to Y can obtain relative to the causal structure in which it occurs. The characterisations and labels for the distinct ways in which genuine causation can obtain are commonly referred to as "causal concepts" (see, e.g., Hitchcock 2003, 2007b; Godfrey-Smith 2009; Reiss 2011a).

To see the difference more clearly, suppose that there is a genuine causal relation between X and Y within a causal structure that includes some known additional factors $Z = \{Z_1, Z_2, \ldots, Z_n\}$, then the causal influence from X to Y can, for instance, be defined as either a *net effect* or a *component effect*. Following one version of this distinction (see Hitchcock's 2001b), two causal concepts can be characterised as follows:[22]

> Net effect: *X* has a net effect on *Y* if and only if *Y* varies as *X* is varied while holding fixed other appropriate factors, including common causes of *X* and *Y*, but excluding factors intermediate between *X* and *Y* (Hitchcock 2001b, p. 372).[23]

> Component effect: *X* has a component effect on *Y* along a particular causal route if and only if *Y* varies as *X* is varied while holding fixed other appropriate factors, including factors that are intermediate between *X* and *Y* along other routes (Hitchcock 2001b, p. 374).[24]

Alternatively, the causal influence from X to Y can be characterised in terms of its sufficiency and necessity conditions. Accordingly, X can be either *sufficient*, but not necessary to produce or affect Y (since perhaps Y can also be

[22] Hitchcock's treatment of these causal concepts is explicitly meant to be "theory-neutral" with respect to any existing theory of causality (see Hitchcock 2001b, 369). To him the distinction between the two concepts is meant to hold "without presupposing any one theoretical perspective [about causation]" (p. 369).
[23] The notion of a 'net effect' is also sometimes called 'total effect' (Pearl 2000, pp. 151-152, 164; Pearl 2001) or 'total cause' (Woodward 2003, pp. 50-51).
[24] This causal concept, which Hitchcock (2001b) calls 'component effect along a causal route', is essentially the same that Woodward (2003, pp. 50, 57) calls 'contributing cause', and fairly similar to what Pearl (2001) defines as 'path-specific effect'.

caused by some Z_i alone), or *necessary*, but not sufficient for Y (since perhaps some Z_i is always required to interact with X in order to cause Y). Or as John Mackie (1974, chapter 3) famously pointed out, X can be an *INUS* condition (an insufficient but necessary part of a set of conditions that are together unnecessary and sufficient) for Y.

Then again, in some other cases, the causal influence from X to Y consists of preventing or interrupting the occurrence of Y, rather than bringing it about. In such cases, X is a *preventative* of Y. Following one version of this concept:

> Preventative: X is a preventative of Y if and only if X has occurred and Y has not, and there is an interaction between X and a causal process (generated by some Z_i) such that if X had not occurred, the causal process would have resulted in Y (see Dowe 2001, 221).

Many additional causal concepts can be characterised using alternative criteria (see, e.g., Pearl 2000). The examples here are meant to be illustrative, not exhaustive. Even so, the second type of causal pluralism can be described as follows.

> ***Pluralism about causal concepts*** is the position that there are different causal concepts, each one corresponding to a distinct type of causal influence occurring in a causal structure. These concepts are all causal, and none of them is privileged as the main or the most basic concept of causation (e.g., Hitchcock 2001b; Hall 2004; Cartwright 2004; Reiss 2009a, and 2011a).

Accordingly, given that a certain claim is deemed to be genuinely causal (in line with one, some, or all general theories of causality), it can have different meanings depending on the causal concept that properly captures the type of causal influence singled out by the claim. Thus "X *causes* Y" means different things depending on whether:

X is a *net cause* of Y, or
X is a *contributing cause* of Y, or
X is a *sufficient cause* for Y, or
X is a *necessary cause* for Y, or
X is a *preventative* for Y, or
X is defined as any other causal concept in relation to Y.

Notice that, in principle, the different causal concepts appearing in the propositions above could all be characterised by any one of the general theories of causality (see Hitchcock 2001b). Hence, conceptual causal pluralism does not entail pluralism about theories of causality. Analogously, it is possible to hold a monistic position about causal concepts (i.e., to argue that there is only one causal concept) and also accept that different theories of causality capture different dimensions of what is to be causal (see, e.g., Russo and Williamson 2007; Williamson 2008; Casini 2012). Hence, pluralism about general theories of causality does not entail conceptual pluralism.[25]

The implications for the practical relevance of causal claims should be obvious. If different interpretations of the claim "X *causes* Y" correspond to different causal concepts, and thus refer to different ways in which the causal influence from X to Y obtains, then the same causal claim can have different practical potential and be useful for policy in different ways depending on which interpretation is taken as its actual meaning. Different ways of 'causing', denoted by distinct causal concepts, need not be practically relevant in exactly the same way for the attainment of a particular goal. And therefore—at least for policy purposes—the specifics of the causal concepts employed in scientific claims should be made explicit.

For instance, the practical relevance of a causal claim could differ, on the one hand, when it means that X is a *preventative* of Y, and on the other, when it means that X is a *contributing cause* of Y. For if X is a preventative of Y, then one knows that a causal interaction between X and a causal process that results in Y can preclude the occurrence of Y (see Dowe 2001). In such a case, X is sufficient to prevent Y. Whereas if X is a contributing cause of Y, then all one knows is that X has a component causal influence on Y along one particular causal path (see Hitchcock 2001b, 374; Woodward 2003, 57). Given that there could be various paths going from a cause to an effect, component causal influences need not be sufficient for producing their posited effect. From this comparison, it seems that preventatives can have a higher practical power than contributing causes.

[25] Some of the new pluralistic approaches (both about causal theories and about causal concepts) have become more epistemologically rather than metaphysically motivated, and hence have moved from questions about what is the nature of causation to questions about what are the most useful ways of investigating and learning about causal relations. Along the same line, causal accounts such as Woodward's or Pearl's—which can be taken as pluralistic in the two forms I have described in this chapter—are not inquiries in search for the best theory about the nature of causation, but rather use elements of several theories of causality (probabilities, counterfactuals, interventions, and so forth) in order to investigate and characterise different causal concepts, which then are used to illuminate various issues of causal inference and causal explanation (see Woodward 1996; 2003; and Pearl 2000).

As another example, knowing that X is a *net cause* of Y, i.e., that the causal influence from X to Y includes all relevant causal paths to Y, and thus knowing that that X is sufficient for affecting Y in a certain causal structure (see Hitchcock 2001b, 369-373) offers a different practical power from knowing that X is an INUS condition for Y (see Mackie 1974, chapter 3). If the goal is to produce, or to predict as accurately as possible an occurrence of Y, then knowing that X is a net cause of Y confers a more reliable practical power than knowing that X is an INUS condition, which in fact—without having also the appropriate information about the additional causal factors that together with X are at least sufficient for Y—would only offer a limited and somewhat unreliable practical power.

These illustrations are meant to give a broad impression of the different ways in which scientific causal knowledge can be exploited for practical purposes depending on its different causal interpretations. Different causal concepts need not have the same practical relevance, and hence conceptual causal pluralism has direct consequences on how one would interpret and be entitled to use causal knowledge to design or implement policy recommendations. Notice that this is a crucial feature of the practical relevance of causation about which the *standard view* remains completely silent.

If the ideas presented in this section are correct, then clarifying or disambiguating the meaning of causal claims is of utmost importance before using them as the basis for any recommendation or implementation of a policy. It is still to be seen whether this step is completed or bypassed in practice, i.e., whether the causal claims that are actually used to guide policy have an accurate and unambiguous meaning (this will be done in the following chapter). For the moment and to conclude this chapter, I will offer a brief and simple example of causal claims that are well established in economics, and which are meant to be useful to produce practical results, yet they are implicitly open to be interpreted according to distinct causal concepts.

3.3. Concluding remarks

In this chapter I have offered an elaboration on causal pluralism and offered a conceptual approximation of how distinct causal concepts can be related to distinct forms of practical potential. Different meanings of a causal claim can vary the relevance of the claim for the production of practical goals in actual situations. Furthermore, given the various distinct possible meanings for a single causal claim of the form 'X causes Y', it is recommendable that a proper disambiguation of the precise meaning of causal knowledge is carried out before it is used for policy purposes. In the following chapter, I will illustrate with a case study the semantic complexity of the type of causal claims that are often used to support concrete policy recommendations.

CHAPTER 4

ON THE MEANING OF CAUSAL GENERALISATIONS: THE OECD RESEARCH ON UNEMPLOYMENT POLICY[26]

In the previous chapter I argued that a causal claim 'X causes Y' can have different meanings depending on different conceptual characterisations of the causal relation (section 3.2), and thus I suggested that a careful semantic disambiguation should take place before any piece of causal knowledge is used for policy purposes. In the present chapter, the complexities related to analysing the meaning of causal generalisations in policy-oriented social sciences are illustrated in the context of a concrete case of economic research explicitly intended to guide public policy, namely: the OECD research on the causes of unemployment.

In addition to the different interpretations that can be given to causal claims (as a consequence of conceptual causal pluralism), the analysis of this particular example brings into the open two additional potential sources of semantic complexity: ambiguities related to the meaning of the causal relata; and perhaps more significantly, ambiguities related to the proper specification of the population of which causal generalisations are meant to be true. The latter—as will be shown in the remaining chapters of this thesis—turns out to be a particularly important issue in policy-oriented social sciences.

In economics, as in other policy-oriented social sciences, causal generalisations are the main kind of causal claims investigated, established, and used in order to support practical recommendations. The research on the institutional determinants of unemployment done by the Organisation for Economic Co-operation and Development (OECD) in the 1990s provides a concrete illustration of this practice. The initial motivation for this research came from the persistence of high unemployment in most OECD member countries throughout the 1980s and into the early 1990s. Then the ministers of these countries required the OECD secretary-general "to initiate a comprehensive research effort on the reasons for and the remedies to the disappointing progress in reducing unemployment" (OECD 1994a, 1).

The major outcome of this effort was the 1994 *OECD jobs study*, which was presented in two parts: *a scientific report*, subtitled "evidence and explanations" (OECD 1994b) and explicitly put forward as the evidential base for the second

[26] This chapter is based on an article co-authored with François Claveau, only with a number of modifications in order to make explicit the connections with the other chapters of this monograph. For the published article, see Claveau and Mireles-Flores 2014.

part, *a policy-oriented report* that was subtitled "facts, analysis, strategies" (OECD 1994c). The transition between the two parts of this study was done mainly through the formulation of causal generalisations.

From a broad perspective, the process whereby scientific causal knowledge was used to come up with effective remedies to unemployment can be thought of as comprising the following three stages: first, a scientific evidential base was gathered and investigated; second, causal generalisations about unemployment were established on account of the evidential base; and third, policy recommendations were proposed on the basis of these scientific causal claims.[27] In this chapter, the focus will be exclusively on the analysis of causal generalisations in relation to their use for policy purposes at the third stage.

Some examples of causal generalisations presented in the OECD research on unemployment are the following:

1. The lack of labour market flexibility of the OECD economies is the principal cause of high and persistent unemployment (OECD 1994b, vii).

2. Government-imposed barriers to wage flexibility cause higher unemployment (OECD 1994b, part 2, chapter 5).

3. Higher employment protection causes higher unemployment (OECD 1994b, part 2, section 6.3).

4. Short-time work schemes help preserve permanent jobs (OECD 2010, 68).[28]

5. More generous unemployment benefits cause higher unemployment (OECD 1994b, part 2, chapter 8).

It is from causal generalisations such as these that policy recommendations are inferred in the OECD study. Yet—as it was argued in the previous chapter—the first step in order to use these causal claims for policy purposes is to get their meaning right. So the main question in this chapter is the following: *what is the meaning of the causal generalisations presented by the OECD?*

[27] To be fair to the OECD, one can find here and there statements in the reports that make the narrative more complex. Nonetheless, what seems beyond doubt is that, after the publication of the report, the causal generalisations and the associated policy recommendations presented by the OECD had great *persuasive power*, and became established recipes for the relevant expert community with all caveats stripped and until very recently entirely forgotten. For discussions of the impacts of the OECD study on the expert academic community, see, Freeman 2005, 131-132; Blanchard 2006, 51-52; Boeri and van Ours 2008, 1-2. For a more recent and revised perspective on unemployment by the OECD, see OECD 2006.

[28] Short-time work schemes are public schemes inciting employers to temporarily reduce the number of working hours of their employees instead of laying them off.

There is a long tradition in philosophy devoted to the study of meaning, namely: semantics. A standard way of investigating the meaning of linguistic utterances is to base the analysis on the notion of representation (see Speaks 2011; Peregrin, 2012, 3). According to this view, singular terms are meaningful in so far as they represent, or stand for, something. Nouns stand for objects, predicates stand for properties and relations, and the meaning of compound statements is entirely dependent upon the meaning of its constituents. Knowing the meaning of a compound statement would then amount to know its truth conditions in terms of some extensional relations among its components. In other words, the meaning of a sentence is given by "what the world would have to be like for [the sentence] to be true" (Heim and Kratzer 1998, 1)

This semantic approach can be called "referentialist", since it bases a theory of meaning upon a theory of reference, or in Robert Brandom words "it moves from a story about what is represented to one about what is expressed" (Brandom, 2007, 651).[29] For example, the meaning of a sentence like 'Snow is white' would depend on whether all its component terms ('snow', 'is', and 'white') stand for or "refer to" something, and on whether the posited relation among these elements actually holds. To know the meaning of 'Snow is white' would then amount to know that the sentence is true if and only if there is a set of objects referred to by the word 'snow', such that all the elements of this set have the property referred to by the word 'white'.

In line with the referentialist semantic tradition, the analysis of the meaning of scientific causal generalisations has been posed anew among recent philosophical accounts of causation. Accordingly, the meaning of sentences of the form 'X causes Y' is typically identified, on the one hand, by specifying the referents of the causal relata 'X' and 'Y' and, on the other, by providing an analysis of what the verb 'to cause' stands for.[30] The work of James Woodward is emblematic of this approach. He explicitly puts forward that a goal of his book *Making things happen* (2003) is to provide an account for "capturing or clarifying [the] 'content' or 'meaning'" of causal claims (Woodward 2003, 7). The meaning of causal generalisations according to his account can then be unpacked by properly characterising their truth conditions in terms of invariance under possible changes and interventions. Different causal concepts

[29] This approach is also sometimes called "truth-conditional" or "representationalist" semantics (see, e.g., Peregrin 2012).
[30] There are some exceptions to the referentialist approach to the meaning of causal claims, see, e.g., Williamson 2005, chapter 9; Spohn 2006; Beebee 2007; Reiss 2011a, and 2012.

refer to different sets of potential interventions under which a causal claim is taken to be true (see Woodward 1996; 2003, chapter 6).

A distinctive characteristic of the recent philosophical literature on causal knowledge is the explicit aim at being relevant to social scientists by helping them to get the meaning of causal claims right, and to contribute in this way to improve their practice. Again, in Woodward's words:

> [M]y project has a significant *revisionary* or *normative* component: it makes *recommendations* about what one ought to mean by various causal and explanatory claims, rather than just attempting to describe how we use those claims. It recognizes that causal and explanatory claims sometimes are confused, unclear, and ambiguous and suggests how these limitations might be addressed (Woodward 2003, 7, emphasis in the original).

Thus, whenever a causal claim is "confused, unclear, and ambiguous", a philosophical account of causation such as Woodward's aims at suggesting "what one ought to mean" by such a claim. Keeping these *revisionary* aspirations in mind, the example analysed in this chapter is intended as a contribution in two respects: First, the analysis investigates whether the causal generalisations proposed by the OECD are precise enough to serve as reliable guides for unemployment policy, or whether they are "confused, unclear, and ambiguous" instead. And secondly, it illustrates the application of a referentialist analysis to the meaning of causal claims as they occur in actual scientific practice, and thus the example sheds light on virtues and limitations of this type of semantic approach.

To guide the analysis, the following schematic form of a causal generalisation can be used to identify the elements in need of clarification:

$$\text{(For } P\text{)}, X \hookrightarrow Y \qquad \text{(schema CC)}$$

In this schema, the symbol '$\hookrightarrow$' is a connective that refers to a generic causal relation where the causal influence goes from X to Y (in principle signifying any conceivable causal concept and analysable by any theory of causality), 'X' and 'Y' stand for the causal relata,[31] and the clause 'For P' specifies the relevant population for which the causal claim is meant to be true, where P is composed by individual units u_i, such that $P = \{u_1, u_2, \ldots, u_n\}$.

[31] Following a strong trend in the philosophy of causation (e.g., Spirtes, et al. 1993; Pearl 2000, Hoover 2001; Hitchcock 2001a, Woodward 2003; Hausman 2005), and in conformity with general usage in economics, upper-case italics (X and Y) are variables, and lower-case italics (x and y) represent specific values of these variables.

To disambiguate the meaning of the causal generalisations presented by the OECD, the following four questions have to be answered:

1. What do the causal relata *X* and *Y* refer to?
2. What does the causal relation '$\hookrightarrow$' stand for?
3. What does the clause 'For *P*' refer to? (i.e., what is the relevant population?)
4. Which unit-level causal claims are entailed from the truth of the causal generalisations?

Each of these questions is discussed respectively in the subsequent four sections of the chapter.

4.1. The meanings of the causal relata

Consider the unemployment rate (*U*), which is the variable standing for the posited effect common to all the OECD causal generalisations presented above. What is the meaning of variable *U*? What does a change in this variable amount to in a real economy? The unemployment rate is defined as a ratio between the number of participants in an economy having the status of 'active and jobless' to the number of participants being 'active' but in a current job. To calculate this number, a subset from all individuals in an economy, including only the ones that are 'active' is identified. The number of individuals in this subset constitutes the denominator of the ratio. Then a subset of the 'active' is taken apart by considering the individuals who are officially 'jobless', and the resulting number constitutes the numerator of the ratio.

Unfortunately, this relatively simple characterisation of *U* as a ratio does not seem to have a clear-cut referent. The ambiguity comes from the specification of the two relevant categories—'active' and 'jobless'—since, as most specialist would recognise, there are many different ways of defining these categories, and there is no unambiguous method for measuring them. Precisely for this reasons, the International Labour Organization (ILO) has made an effort to provide detailed guidelines on how to define and measure these categories in a uniform way (see ILO 1982, 2-5). These guidelines have actually helped pin down a more definite meaning for the concept 'unemployment rate', however some semantic ambiguities remain, since the guidelines still leave some margins for subjective (and contextual) interpretation. As the OECD recognises, notions like 'active job-search' are "in some countries interpreted rather widely" (OECD 1994a, 186).[32] Ultimately, the particular decisions that are made about the meaning of the 'unemployment rate' will affect the meaning of the causal generalisations in which it occurs. In as much as the referent of U is left ambiguous, the meaning of these causal generalisations cannot be entirely clarified.

The semantic complexity of these generalisations greatly increases when one considers the meaning of some of the variables standing for the posited causes such as: the 'adjustment potential of an economy', the 'generosity of unemployment benefits', or the 'short-time work schemes', which can be labelled as variables A, B, and S, respectively. To elucidate a definite meaning for these variables turns out to be much more challenging than in the case of U, since these are composite variables and thus their values are *multidimensional*.

To understand this, consider a distinction between two types of variables: unidimensional and multidimensional.

> A ***unidimensional variable*** is an ordinal variable V for which any two values v, $v' \in D_v$, are either $v>v'$, $v<v'$, or $v=v'$.
>
> A ***multidimensional variable*** is a variable composed by a set of unidimensional variables, i.e., $M=(V_1, V_2,\ldots, V_n)$, such that any value of variable M is in turn a compound of the values taken by the component variables, i.e., $m=(v_1, v_2,\ldots, v_n)$.

When a variable is multidimensional and takes a particular value, it takes in fact a specific value for each of its dimensions, thus for every two realizations m and m', there are two realized vectors $m=(v_1, v_2,\ldots, v_n)$ and $m'=(v_1', v_2',\ldots, v_n')$. Hence, to make sense of any change in the value of this type of variables, one has to disambiguate the meanings of assertions about changes in the values of these vectors.

For instance, a first attempt to capture the meaning of variable B (the generosity of unemployment benefits) would have to distinguish between at least three dimensions: 'the level of benefits' B_l, 'the duration of entitlement' B_d, and 'the eligibility conditions' B_e (see Nickell, et al. 2005, 4; and Boeri and van Ours 2008, sec.11.1). However, specifying B as a tridimensional vector would still be an oversimplification, since the first two dimensions, B_l and B_d, are in

[32] The OECD continues: "[G]reater standardization, for example with a consistently strict interpretation of the notion of 'step of active job-search', could make a significant difference to the level of unemployment reported in labour force surveys for some countries" (OECD 1994a, 187).

fact multidimensional as well. The following comment by an OECD economist captures the difficulty involved in trying to clarify the meaning of 'the level of benefits' B_l, which is generally expressed in terms of the 'replacement rate' (i.e., the ratio of unemployment benefits to previous employment earnings):

> There is no such thing as *the* replacement rate in any OECD country, rather there are a myriad of replacement rates corresponding to the specific personal and family characteristics of the unemployed, their previous history of work and unemployment, and the different structures and entitlements of unemployment insurance (UI) and social assistance (SA) systems in OECD countries and the ways in which these systems interact with tax systems. Once one tries to grapple with these complexities in order to compute replacement rates for the purpose of international comparisons, the task becomes a daunting one (Martin 1996, 100).

It is thus obvious that the generalisation 'the generosity of benefits (*B*) increases unemployment (*U*)' is susceptible to a variety of interpretations.[33] To be sure, depending on what a change in variable *B* is taken to be, there could be cases when a change in the unemployment benefit system could be interpreted unambiguously as an increase in generosity, say, because all the component dimensions are changed in the same direction. Nevertheless, it would not be crystal clear how to measure the more complicated cases in which all the composite dimensions would vary in different directions.

In practice, instead of analysing in detail each one of the different dimensions of a multidimensional variable, authorities rely on transformations of the multidimensional variable into a unidimensional scale. For instance, in relation to variable *B*, a "summary measure of benefit entitlements" has indeed been constructed for the purposes of the *OECD Jobs Study*. The explicit goal of this measure was "to capture the degree of 'generosity' of a country's benefit system" (OECD, 1994a, part II, 172). Nonetheless, even if the construction of such a measure can be taken as an impressive achievement—since for each country, it averages the replacement rates across 18 distinct personal situations—the OECD still is of the opinion that it is only "a very approximate indicator" of actual generosity, as there are many instances where the measure would actually fail to register a change in generosity while intuitions would go

[33] The causal variables in other generalisations presented in the OECD research have the same multidimensional character, for instance, if one considers the purported cause 'short-time work schemes', variable *S*, appearing in the third causal generalisation among the examples above, one founds that recent discussions about short-time work schemes actually decompose it into 14 dimensions (which are then regrouped into four main families of features; see OECD 2010, Annex 1.A1).

in the opposite direction (OECD, 1994a, part II, 173-176). A definite referent for *B* thus remains somewhat elusive.

To sum up, many implicit methodological decisions have to be made about the specification of the relevant categories (e.g., being 'active' or 'jobless'), and about the multidimensionality of some concepts involved, in order to get a somewhat clear notion of what is meant by the causal relata in these policy-oriented causal generalisations.

4.2. The meanings of the causal relation

As was mentioned before in chapter 3, the problems with each of the main univocal theories of causality have led philosophers to develop pluralistic accounts that desist from reducing causality to a single primitive notion, and rather propose to characterise causal relations in a clear and tractable manner. In economics, it is seldom explicit which theory of causality one is supposed to hold to understand the causal relations presented and discussed. However, some of the recent work on causality and causal inference in special sciences suggests that certain non-reductive versions of the manipulationist approach are indeed suitable for analysing economic causal claims (see, e.g., Hausman 1998; Pearl 2000, chapter 5; and in particular Hoover 2001). Whether this is the "most adequate" approach to characterising causality in economics—given the existing *plurality of theories of causality*—is an open question out of the scope of this chapter. In this section the focus is on the consequences of *pluralism about causal concepts* (as described in chapter 3, section 3.2) for the meaning of the OECD causal generalisations.

To begin, let us consider the potential meanings of the causal relation '$\hookrightarrow$' in a claim as represented by schema CC, but with only one single unit as the relevant population referred to in clause 'For *P*'. Notice that, in the context of the OECD research, the units in the population refer to individual countries. To restrict the analysis to "single-unit causal claims" is useful for at least two reasons: first, it helps us focus on what happens when distinct causal concepts are singled out by the causal relation '$\hookrightarrow$', keeping this issue separated from other sources of semantic ambiguities (e.g., different specifications of the population of application for a causal claim; see section 4.5 below). And secondly, the clarification of the meaning of claims with multi-unit populations can be subsequently approached by asking which single-unit causal claims are actually entailed by a causal generalisation (also discussed in section 4.5 below).

The issue of which single-unit claims are warranted by the truth of causal generalisations is especially significant in policy-oriented sciences in which scientific causal knowledge is meant to be useful for the production of effects in specific target subpopulations or individual units, and not necessarily in the population as a whole. For instance, policy makers aiming at reducing unemployment can be thought of as being interested not necessarily in a reduction of the general incidence of unemployment for the totality of OECD countries, but mostly in how *one particular unit* (i.e., their own specific country) can reliably reduce *its* current situation of unemployment.

There are at least two distinctions among causal concepts (introduced in chapter 3, section 3.2) that are relevant to the meaning of '↪' in the case of the OECD generalisations: (1) *net* versus *component* causal effects, and (2) *necessity* versus *sufficiency*.

4.2.1. Net versus component causal effects

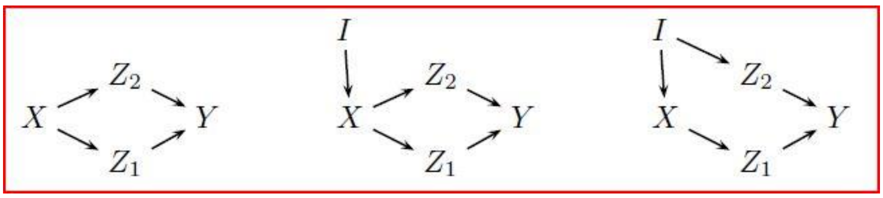
This distinction comes from the fact that in some cases X can affect Y through multiple causal paths in a given causal structure. Then a claim about a net effect means that X affects Y taking into account all the existing causal paths, while a claim about a component effect means that X affects Y only along a particular causal path (see Hitchcock 2001b, 372, 374).

Consider a causal structure for some particular unit, which can be represented with the following equations:

$$X = E_X$$
$$Z_1 = \alpha_1 X + E_1$$
$$Z_2 = \alpha_2 X + E_2$$
$$Y = \beta_1 Z_1 + \beta_2 Z_2 + E_Y$$

In these equations, X and Y are variables standing for the causal relata, Z_1 and Z_2 represent causal factors that are intermediate between X and Y along two distinct causal paths, and the Es represent all other (uncorrelated) relevant causal factors. The causal graph for this system is shown in Figure 1, Graph 1.

Figure 1



Graph 1 **Graph 2** **Graph 3**

Accordingly, the *net* effect of X on Y (denoted here as '$Net\Delta y$') can be interpreted as the change in Y due to an intervention I that varies the value of X from x_0 to a new value x_1 (see Figure 1, Graph 2). Thus the *net* causal effect of the change in X ($\Delta x = x_1 - x_0$) will be $Net\Delta y = (\beta_1\alpha_1 + \beta_2\alpha_2)\Delta x$.

The interpretation of the *component* causal effect of Δx on Y along one causal route (here denoted with the label '$Comp\Delta y$') can be similarly obtained, yet by introducing a more complex intervention. To consider only a causal effect along the causal path passing through Z_1 (as shown in Figure 1, Graph 3), the required intervention involves changing the value of X from x_0 to x_1, but also breaking the influence of X on Z_2 such that Z_2 remains fixed with the value z_0—i.e., the value it would normally have when X is set to x_0, even though the value of X is set to x_1. The component effect along the causal path of Z_1 will then be $Comp\Delta y = \beta_1\alpha_1\Delta x$.

Now consider the OECD causal generalisation: 'generosity of unemployment benefits (B) *causes* unemployment (U)'. As Daniel Hausman suggests, the 'causal role' of a causal relation refers to the causal influence of a cause on its effect being either *positive* or *negative* (Hausman 2010, 48). Accordingly, the causal role of B in relation to U is commonly believed to be positive due to evidence of mechanisms operating in the labour market such as the 'job search effect', i.e., that as generosity of benefits increases, jobless individuals covered by an unemployment benefit program tend to search for jobs less intensively (see Boeri and van Ours 2008, 11.2.1). Exclusively taking into account this component effect, an increase in B will be expected to cause an increase in the unemployment rate U. The corresponding policy recommendation would be to reduce and keep under strict control the generosity of unemployment benefits "at levels that do not discourage job search excessively" (OECD 2006, 21), which is indeed what the OECD study recommends.

However, this would be oversimplifying the matter, since some economists argue that there is also a countervailing 'entitlement effect' that negatively relates the generosity of unemployment benefits to the unemployment rate. As the story goes, if B increases, then the jobless individuals currently ineligible for unemployment benefits would have a stronger incentive to get employed soon, because they will be entitled to higher unemployment benefits in case they lose their jobs in the future (see Boeri and van Ours 2008, 11.2.2). Thus, exclusively taking into account this component effect (i.e., the entitlement effect), an increase in B will be expected to cause a decrease in the unemployment rate U, and more generous unemployment benefits would then supposedly be an effective policy recommendation to reduce unemployment.

As a matter of fact, the 'job search effect' is believed to dominate the 'entitlement effect', thus according to the extant literature, the net causal effect from *B* to *U* is normally positive. Yet, even when the net effect is well known in general, making explicit all the potential component effects can still be quite useful for policy purposes. For instance, before implementing any concrete change in the generosity of a system of unemployment benefits, it may be wise that policy makers first evaluate whether the 'entitlement effect' could possibly counterbalance the other component effects with an opposite causal role in the particular country in which the policy will be implemented.

4.2.2. Necessity versus sufficiency

With respect to the notions of necessity and sufficiency, the single-unit causal claim '(For u_i), $X \hookrightarrow Y$' can mean four different things: First, that for unit u_i, bringing about a change in *X* is *necessary* and *sufficient* to induce a change in *Y*. Second, that for u_i, a change in *X* is *necessary* (but not sufficient) for a change in *Y*. Third, that for u_i, a change in *X* is *sufficient* (but not necessary) for a change in *Y*. And fourth, that for u_i, a change in *X* is an *insufficient* but *necessary* element of an *unnecessary* but *sufficient* set of causal factors, i.e., an *INUS* condition, for a change in *Y* (see Mackie 1974, chapter 3). This fourfold distinction can be illustrated employing some simple equations representing each of these types of causal connections.

First, an instance in which a change in *X* is *necessary* and *sufficient* for changing *Y* can be represented by:

$Y=\alpha X$ (with $\alpha\neq 0$)

Alternatively, an example of a causal system in which a change in *X* is *necessary* (but not sufficient) for a change in *Y* can be given by:

$Y=Z*X$ (where *Z* is a variable for which $D_Z = \{0, 1\}$)

Notice that, in this latter expression, the multiplication of the variables indicates that it is a causal interaction between *X* and *Z* what is required to affect *Y*, rather than a change in *X* (or in *Z*) alone (see Woodward 2003, 44-45). Whereas a case in which a change in *X* is *sufficient* (but not necessary) for a change in *Y* is given by:

$Y=\alpha X+E$ (where *E* stands for other relevant factors for *Y*)

In contrast to the previous expression, the linear addition of variables implies that there are no interactions among the causal factors, i.e., a change in any independent variable (in this case either *X* or *E*) is sufficient for a change in *Y*. Finally, by combining some components of the previous equations, it is possible to get an example in which a change in *X* is an *INUS* condition for a change in *Y*:

$$Y=Z*X+E$$

In this last case, a change in *X* is a necessary element of a set of conditions (but not sufficient, since *Z* also would have to take the value of 1) which are jointly sufficient to generate a change in *Y*, yet also jointly unnecessary (since a change in *Y* can alternatively be the consequence of a change in *E*).

These distinctions are mainly intended to clarify two points. First, the four causal notions depicted above refer to four different ways in which a change in *X* can bring about a change in *Y* for a given unit u_i. The main difference among these concepts is that each one imposes different requirements on the changes in *X* in relation to a causal structure, so that a change in *Y* actually obtains. Thus, for policy purposes it would seem indispensable that the different requirements of sufficiency or necessity for any causal claim are made explicit and that the particular causal structure of the actual system in which a policy is going to be implemented is properly checked to conform to such requirements. In other words, knowing that 'X *causes* Y' is true is not enough for policy purposes without also knowing the specifics of the meaning of '↪' in terms of its necessity and sufficiency requirements in relation to the causal structure of application.

Secondly, when a causal relation is presented as an equation—or alternatively when some structural equation is given a causal interpretation—the functional specifications of the equation presuppose the choice of a particular causal concept in terms of sufficiency and necessity requirements. This point is especially relevant in sciences such as economics in which most of the empirical research conducted to build the 'evidential base' for causal generalisations employs some version of structural equations analysis.[34] In general (yet not always), the regression literature in economics employs linear

[34] Some methodological accounts on the use of regression analysis as the primary tool in empirical economics are: Pearl 2000, chapter 5; Hoover 2001, chapter 7; Morgan and Winship 2007, chapter 5; Angrist and Pischke 2009.

specifications to establish constant (and homogeneous) causal effects.[35] A typical regression equation looks like:

$$\boxed{Y = \alpha X + \sum_i \delta_i Z_i + U_Y}$$

From this linear functional form, it is clear that empirical studies which employ this basic type of equations to estimate causal effects assume that a change in X is *sufficient* for a change in Y, regardless of any change in the values of the other variables Z_i.[36]

Issues of necessity and sufficiency could lead to some semantic ambiguity in relation to policy-oriented scientific claims. These can be exposed by probing the correct interpretation of the causal claim 'generosity of unemployment benefits (*B*) *causes* unemployment (*U*)'. Given that the OECD research on unemployment recognises that distinct causes can affect *U* through paths that are independent from *B*, then any interpretation of *B* as a *necessary* cause can safely be ruled out, which leaves us with two alternative interpretations: *B* is either a *sufficient* cause or an *INUS condition* for a change in *U*.

The OECD research makes extensive use of regression methods based on linear equations such as the one described above in order to study the labour market institutions. Hence, there is some indication that the causal claim can be interpreted as 'a change in *B* is *sufficient* to bring about a change in *U*'. Nevertheless, closer inspection of the narrative in the OECD research reveals the recognition that there are causal interactions among distinct factors that affect unemployment, such as interactions between *B* and changes in the GDP, for each of the member countries. Therefore, it might be more appropriate to interpret *B*, the GDP, and any other relevant causal factor as *INUS* conditions. And this seems to be the OECD position in some parts of the reports:

---

[35] This sentence in fact refers to two assumptions common to regression analysis, 'linearity' and 'homogeneity', yet the focus in this section is only on the consequences of the former. The consequences of the latter on establishing causal effects in heterogeneous populations for policy purposes are extremely significant, since (as it will be further discussed in section 4.5 below) in heterogeneous populations it is not straightforward whether generalisations that are true for the population as a whole would also be true for specific individual units of the population. This issue alone is the motivation for most of the discussion in the following chapter. For some of the few existing philosophical discussions of this topic, see Hitchcock 2001a; Steel 2008; Hausman 2010.

[36] The regression equation could let go this assumption of sufficiency by allowing interaction terms like X^*Z_i, yet this would again imply a quite specific form of interaction between the regressors.

> The hypothesis here is not that unemployment benefits (or other structural factors) may cause movements in unemployment independently of movements in GDP, but rather that they may contribute, for example, to the unexpected depth or prolonged nature of a recession (OECD 1994b, 171).

Another reason in favour of reading the causal claim as referring to an *INUS* condition is the recognition by the OECD that "institutional factors influence whether or with what lag new benefit entitlements affect unemployment" (OECD 1994b, 211). Thus, according to the OECD, institutional conditions are among the additional structural factors that usually interact with *B* in order to affect the level of unemployment. Still, if this is the correct reading, then the OECD oscillates between a *sufficiency* interpretation (implicit in the empirical research methods employed) and an *INUS* interpretation (semi-explicit in the rhetoric of the reports) of some of the proposed causal generalisations.

4.3. Specifying the relevant population

The truth value of a causal generalisation can vary depending on particular interpretations about the right population of application denoted by the clause "For *P*". This section explores the consequences of this additional source of semantic ambiguity for causal generalisations. The first thing to notice is that the contextual qualification 'For *P*' of causal generalisations is seldom explicitly stated. One would usually say that 'smoking causes lung cancer' without specifying what are exactly the units of analysis and the population of application, such as 'for humans'. People typically accept as true claims like 'seatbelts save lives' without any specification of whether this claim is meant to be true 'for all humans', 'for all accidents', 'for humans of a certain type in accidents of a certain type', and so on (see Hausman 2010). This type of vagueness is ubiquitous in causal generalisations in the social sciences. Yet many causal generalisations are only true for specific populations or subpopulations or they are true about a certain property of the population but hold no clear truth preserving inference relation with the units.

There are two paradigmatic ways by which the relevant units of *P* can be specified. First, the specification can be *extensional*—e.g., explicitly referring to some particular individuals or to a concrete group or groups of them. The population is then composed of the specific units actually designated. The second option is an *intensional* specification. In this case, the units of the population are systems meeting certain conditions or having certain causal properties—e.g., humans living in Europe, workers between 20 and 40 years of age, modern market economies, and so forth. While the elements of an extensionally-specified population are fixed once they have been designated, the elements of an intensionally-specified population could be changing continuously.

Consider again the causal generalisation ‘more generous unemployment benefits cause higher unemployment’. It is plausible to say that the implicit population for such a claim is meant to be extensionally specified when stated in the OECD reports. The OECD’s primary role is to advice its members on policy issues, and it had 24 member countries in the early 1990s (when the *OECD Jobs Study* was commissioned), hence the list of the 24 members could be taken as the implicit population for the causal claim.

One could argue just as well that the implicit population for the claim should be viewed as intensional, for instance: countries that fulfil a set of specific conditions and possess a set of causal characteristics which are required to be a member of the OECD. The population so specified could vary in time as the actual membership of the OECD changes. In fact, the OECD membership has substantially changed since the early 1990s from 24 to 34 countries. With the intensional specification, the new members will thus be included as units of P as soon as they have joined the OECD assuming that they in fact have the required causal properties. Yet, new countries could be accepted as members of the OECD without having the relevant properties for which the causal generalisation is meant to be true. In principle, any country having the required causal properties would be part of the relevant population for which the generalisation is true, regardless of it being or not a member of the OECD.

A policy relevant question would then be: are the causal generalisations about unemployment, which were established during the years prior to 1994 (and from a concrete evidential base), true for all new members of the OECD? Consider, for example, Mexico, which became an OECD member in 1994. Mexico is in many respects quite unlike all the other countries that were already members: its “official” unemployment rate during the early 1990s was below 3%, the proportion of its informal economy was (and still is) estimated to be far larger than that in all other OECD member countries, and it did not then (and still does not) have an unemployment benefits system.[37] These facts about Mexico (particularly the last one) should make one doubt that the causal generalisation ‘more generous unemployment benefits cause higher

[37]See Fleck and Sorrentino 1994; and references therein for better support for the distinctive Mexican causal characteristics.

unemployment' is really meant to apply to it, even if the claim would be true for all the other members of the OECD.

For a precise intensional specification of *P*, ideally the complete conjunction of all relevant causal properties would have to be stated. Furthermore—and as it will be argued in more detail in the following chapter—the methodological assumptions under which the OECD causal generalisations were established must be made explicit and tested at the unit level. For example, it should be checked how a target unit like Mexico stands in relation to the causal structure of the units that were taken into account when the generalisations were originally established.

The most evident problem with getting a precise intensional characterisation of *P* is that in practice it is quite challenging (if not impossible) to get an exhaustive description of all the relevant causal properties required for a unit to be included. Since an intensionally-specified population has an unspecified extension, we always face the danger that a new country would fit the criteria for being a member of *P* (according to certain known relevant causal properties), but that in fact the generalisation would be false for it, because of some other unknown causal properties specifically related to this new country.

4.4. Causal generalisations in terms of single-unit causal claims

The possible choices related to specifying the population for which a causal claim is meant to be true are particularly significant for generalisations that are intended for policy use. Even if a causal generalisation can be reliably taken to be true in *general*, a user of the claim might want to know whether the claim is also true for the *particular* target unit or subpopulation of interest in which an actual policy is intended to be effective. It is often not straightforward which causal claims about individuals (i.e., single-unit causal claims) are in fact entailed by a causal generalisation. Among other things, this depends on whether the population *P* for which the causal generalisation is meant to be true is homogeneous or heterogeneous.

Causally homogeneous population: In this case, each single-unit causal claim is true for all members of *P* and they all have exactly the same causal meaning, i.e., the causal relata '*X*' and '*Y*' are unambiguously defined and measured uniformly in all the units, and the causal relation '$\hookrightarrow$' refers to the same causal concept and has the same (either positive or negative) causal role in all of the single-unit claims. The meaning of a causal generalisation for this kind of population can be understood as an inference from the conjunction of all the single-unit claims. For example, consider the causal generalisation 'for the population including all samples of element bismuth, heat of 271°C causes the sample to melt'.[38] The homogeneity among the members of this population (in relation to melting) warrants that the truth of any single-unit claim about the melting temperature of any sample of bismuth can be reliably inferred from the truth of the generalisation. If a population *P* is homogeneous in this sense, then the generalisation '(For *P*), $X \hookrightarrow Y$' is true if and only if each of the single-unit claims '(For u_i), $X \hookrightarrow Y$' are also true.

Causally heterogeneous population: In this type of situation either: 1) some single-unit causal claims are true for some members of *P*, but false for some others, or 2) all single-unit claims might be true for all members of *P* in the sense that there is a *causal influence* from X to Y, but the *causal role* (either positive or negative) is different among them (see Hausman 2010). A causal generalisation for this kind of populations is the result of an inference from a subset of single-unit causal claims to a statement about the whole population *P*. More precisely, a causal generalisation refers to a certain aspect of the whole *P*, based upon a subset of single-unit causal claims about some members of *P*. For this type of heterogeneous population, however, the truth of a generalisation does not entail the truth of any particular single-unit causal claim.

In the OECD unemployment reports, just as in most policy-oriented economics research, the evidential base from which the causal generalisations are inferred consists to a great extent of empirical studies using regression analysis. The logic behind this empirical method can be captured by the 'counterfactual model for data analysis' (Morgan and Winship 2007), also known among statisticians and economists as the 'potential outcomes framework' (see Rubin 1990, 2005; Holland 1986; Heckman 2000, 2005).

Broadly put, given that there are some well-defined causal states *X* for the members of a population *P*, single-unit causal claims could be characterised as follows: For each unit u_i, there are two potential values *Y*: y_1 for the outcome of the "treatment" state x_1, and y_0 for the outcome of the "control" state x_0. Only one of these outcomes is observed for each u_i, while the other is hypothetical, yet inferred from the available data (see Morgan and Winship 2007, chapter 2). Then the causal effect on *Y* for a single unit is defined as the arithmetic difference between the two potential values (y_1- y_0). In this account, a single-unit

[38] This example is an adaptation of a case made by John Norton (2003, 649) on "material induction".

claim '(For u_i), $X \hookrightarrow Y$' means that for unit u_i, if X varies from x_0 to x_1, then Y varies in a magnitude equal to $y_1 - y_0$.[39]

Once all causal responses for each unit are obtained (from a data set about P), the population-level causal effect is established by making an inference from a distribution of all single-unit causal effects (plus a set of methodological assumptions). The meaning of the causal effect for the whole population thus varies depending on which property of the distribution is singled out by the estimation procedure (usually some average measure, e.g., mean, median, weighted mean, or the like). Thus the meaning of a causal generalisation established in this way can be stated as follows: 'For the whole P, if X varies from x_0 to x_1, then Y varies on average the magnitude estimated as the causal effect ($y_1 - y_0$). If a population is causally heterogeneous in the sense described above, then it is not obvious how one can infer anything definitive about the truth of a particular single-unit claim from the truth of a generalisation about an estimated average causal effect.

For instance, the OECD generalisation "more generous unemployment benefits (B) cause an increase in unemployment (U)" can mean different things depending on which particular property from the set of single-country causal claims is picked out by the generalisation. For a population of countries $P_c = \{c_1, c_2, \ldots, c_n\}$, let us denote for each country in the population a single-unit causal effect by $\Delta e_i > 0$. Then the following are three possible (non-exhaustive) interpretations of the generalisation about a causal effect of B on U.

1. If the generalisation assumes causal homogeneity among all members of P_c, then its meaning is: "For all members of P_c, $\Delta e_i > 0$".

2. If the generalisation singles out the median, then its meaning is: "For P_c, there is a number m of countries, for which $\Delta e_i > 0$, such that $m > n/2$".

3. If the generalisation singles out the mean, then its meaning is: "For P_c, $1/n \, \Sigma_i \, \Delta e_i > 0$".

The mean interpretation of the claim is commonly called the 'average causal effect', and it is by far the most usual interpretation in the econometrics literature. But notice that the connection between average-effect causal generalisations and single-unit causal claims is highly elusive. The generalisation understood this way does not entail at all that if one draws one country from the population, the causal effect of B on U for this country will be

[39] The potential outcomes framework and its methodological assumptions are described in more detail in the following chapter 5, see section 5.5.

positive as it is indicated by the average. Moreover, the generalisation does not even entitle one to infer that the causal effect would be positive for *the majority* of countries (which is an inference that would be entailed by the median interpretation), since the positive average effect might have been pulled up by a few very large positive single-unit effects in a small minority of countries (while the majority of the total of countries exhibits null or negative single-unit effects).

Econometricians know about these complications related to heterogeneity and in fact have a good amount of sophisticated techniques to deal with them (see, e.g., Angrist and Pischke 2009). Yet, the research results offered to the users of the science, such as the results summarised in the OECD reports, refer to *average causal effects for a population P as a whole*. One can only wonder whether policy makers, when dealing with their own particular unit-level targets, would get the correct interpretation of these causal claims, with all the relevant contextual and technical qualms concerning their causal interpretation.

The point to be taken from this section is that the OECD causal generalisations have entirely different meanings depending on how 'For *P*' is specified, and depending on which properties of the distribution of single-unit causal claims the generalisations refer to. The example also suggests that the more causally heterogeneous a population is, the less reliable the inferences about single-units are likely to be when made only upon the basis of a causal generalisation. The specific assumptions about the truth conditions for a causal generalisation and about the heterogeneity of the relevant population *determine* which inferences at the single-unit level are allowed. This idea—which has crucial implications for the use of causal claims for policy purposes—is explored in much more detail in the following chapter.

4.5. Concluding remarks

The practical value of causal knowledge goes far beyond the intuition that the manipulation of causes is an effective strategy to bring about desired changes in the effect. Disambiguating the meaning of scientific claims that are intended as the basis for economic policy is a priority in the process of policy making. The OECD generalisations and policy recommendations on unemployment are a good example to illustrate these ideas.

The semantic analysis developed in this chapter has shown that the causal generalisations presented by the OECD can have different meanings depending on: a) what are the precise referents of the causal relata, b) which causal concept the causal claims refer to, and c) how the population of application is specified. Moreover, d) the causal effect that the proposed generalisations single out for the whole population of member countries of the OECD do not entail any clear-cut or reliable inference about single-unit causal effects.

The semantic complexity of these causal generalisations leaves the door wide open to subjective (and perhaps arbitrary) decisions made by the researchers and policy makers in order to grasp their own interpretation of these generalisations. More communication and collaboration between economists, philosophers, and policy makers can help make causal notions explicit and clear so that scientific causal knowledge can have a more useful and effective role in the solution of practical socio-economic problems.

PART III
EVIDENCE AND ECONOMIC POLICY MAKING

Causal generalisations are claims of the form 'X causes Y' which are meant to be true for a whole population P given a number of conceptual, methodological, and inferential assumptions. Yet, they are typically presented to non-academic audiences as scientific results without any definite provisos about how they are meant to apply to individual members of a population, or about specific requirements of the background context of application. Policy makers are usually interested in formulating effective policies for particular target units (or sub-groups) of larger populations, e.g., a community, a city, a country, and the like; however—as it is suggested in the example discussed in chapter 4—it is not always clear how to infer *unit-level causal claims* of the form: 'for unit u, X is an effective strategy for Y' from a *causal generalisation* of the form: 'in general for population P, X causes Y'.

Apart from the normal scientific knowledge, additional information is required to warrant reliable inferences from causal generalisations to particular instances of policy implementation. As suggested in chapter 2, the concept of effectiveness depends on knowledge about a number of different elements, in addition of true causal generalisations established in isolation (see chapter 2, section 2.2). There I tentatively characterised *effectiveness* as depending upon the following notions:

> **Effectiveness**: degree of success of the actual implementation of a strategy or policy in pursuit of a definite goal under the ordinary circumstances of the case, conditional upon the *practical potential* of the causal knowledge employed, the *degree of proficiency* of the relevant users, the *persuasiveness* of the relevant claims, and all additional *background conditions* involved in the process of implementation.

In the following chapters I will analyse the role of evidence in supporting the *effectiveness* of policy recommendations. My main aim is to show that there is an important contrast between the logic and the methods of assessing the evidential support, on the one hand, of causal generalisations, and of policy recommendations, on the other. As explained in previous chapters, the *practical potential* of a causal claim depends to a great extent on which causal connotation is referred to by the claim. Different notions of causality require different assumptions and appraisal criteria which in turn determine the specific meaning of a piece of scientific causal knowledge. And different meanings of causality lead to different implications for practical purposes (see chapter 3).

The evidence required for testing scientific causal generalisations is obtained employing methods of causal inference that allow researchers to control for all disturbing factors (the *background conditions*) so that the posited causal relation can be assessed in isolation. Whereas the evidence required for evaluating the effectiveness of a policy recommendation has to be obtained by taking into consideration *all dimensions of the goal at hand*; by using *all the relevant economic knowledge available that relate to these dimensions*; and by using *whatever empirical work that could shed light on the goal at hand and on all its relevant dimensions* (see chapter 1, section 1.5; and contrast with Colander 1994, 41-45).

There is no absolute or overall procedure to achieving policy effectiveness, so the only way of offering some substantive insight on it is by analysing in some detail: a) the typical evidential requirements and testing methods employed to establish causal generalisations, and b) what I put forward as the distinct evidential requirements and methodology of evaluating policy hypotheses. In the following chapters, I shall offer answers to the questions: How does evidence for generalisations relate to policy recommendations? What does it count as acceptable evidence for policy goals? Where can one search for and attain such evidence? Is there any relevant information for policy effectiveness in the existing scientific evidential base of causal generalisations?

In **chapter 5**, I discuss in detail an aspect of the practical relevance of causal knowledge that was prompted by the case study analysed in chapter 4, namely: how is it possible to infer effective strategies for particular country-cases from causal generalisations such as those presented by OECD? I start by exploring the existing philosophical literature on the meaning of causal generalisations, and on the relation between general- and unit-level causation (all this within a probabilistic framework). I argue that most proposals fail to offer a clear and useful account of the inferential relations between causal claims at the general and at the unit level. A key part of this chapter consists in showing how different versions of a causal theory—the probabilistic account in this case—define their causal criteria by making different assumptions about inferential relations between singular and general cases. Defining different criteria (e.g., unanimity, Pareto, or fair-sample conditions) for what counts as causal determines, then, the potential inferences that can be made from the resulting causal claims. I then show how the potential-outcomes framework commonly used to establish causal generalisations in economics and other social sciences relies on a number of conceptual and methodological assumptions (very similar to those used, e.g., in the fair-sample account), which at the same time enable and constrain the inferences that can be made from general results to the unit level. This illuminates a number of facets related to the problem of inferring unit-level causation from *average causal effects* in the presence of causal heterogeneity (as suggested in chapter 4, section 4.4). A way to deal with this issue is by making explicit all the relevant underlying methodological assumptions that are used by researchers to establishing the truth of causal generalisations. Clear knowledge of such assumptions by policy makers can provide useful information about which unit-level and contextual aspects should be checked and further evaluated to help making unit-level inferences on the basis of causal generalisations more reliable.

In **chapter 6**, I suggest that causal generalisations and their practical relevance are inferentially connected to certain features of the *evidential base* employed to established them, and then elaborate on such features. An evidential base is the set of all the evidence that supports and allows an epistemic agent or community to believe that a piece of scientific knowledge is true. In economics, the evidential base used to support causal knowledge commonly includes a set of theoretical results and empirical studies. In this chapter, I focus on understanding the empirical methodology employed to test causal generalisations and on how to use that understanding to support effective policy making. As I have argued in the previous chapters, knowing that a causal generalisation is true is not sufficient for securing the effectiveness of policy strategies; therefore, *good evidence for the truth of a causal generalisation is not necessarily good evidence for believing in the effectiveness of a related policy recommendation*. The first part of this chapter is devoted to critically review the existing philosophical literature on the so-called evidence-based movement and Nancy Cartwright's "evidence for use" approach. I argue that the methodological worries related to the evidence-based movement and of Cartwright's account are not useful enough to properly understand the evidential requirements relevant to policy purposes in economics. The literature has focused on the methodology of randomised controlled trials (RCTs),[40] which are an allegedly important but not representative part of the evidential methods used in most of applied economics. The gold standard of evidence in economics is not RCTs, but a collection of well-established

[40] In relation to the use of randomised control trials in evidence-based medicine and health policy, see Cartwright 2012b. I will comment further on most of her work on the "evidence for use" approach in section 6.1.

econometric techniques. While there are some similarities in the epistemic aims and in the logic of these evidential techniques (for instance, both RCTs and econometric methods aim at testing causal relations by controlling as best as possible for all potential disturbing and contextual factors), the use of econometrics to establish causal claims has some specific methodological consequences for economic policy making. In the second part of the chapter, I analyse and discuss conceptually the basic problems of policy inference in relation to economics. Following Christopher Hitchcock (2001a), I identify different types of causal claims so as to characterise more precisely both: causal generalisations about large populations and policy hypotheses about particular units. I thus propose an account of evidence to support policy hypotheses that is different, and yet grounded upon, the typical ways in which empirical economists obtain evidence for causal generalisations. In my proposal, the distinction between scientific economics and the "art" of economic policy making is made explicit at the evidential level.

In **chapter 7**, I illustrate what I have said in the previous chapters by analysing in detail the empirical literature in international trade economics that studies the economic benefits of free trade. A huge amount of empirical research has been produced with the aim of evaluating the hypothesis that: (more) trade liberalisation causes (more) economic improvements (in terms of growth, investment, employment, and the like). I explore these studies so as to characterise which type of evidence and which evidential methods are employed to support their results. Three methodological issues emerge: a) the lack of definite referents for the relevant causal relata, and specially for the posited causal variable "trade liberalisation"; b) the use of econometric methods as a "gold standard" of scientific evidence, which already come with a built-in conception of causality, and which in turn restrict the kind of inferences that can be reliably made about the effectiveness of policy implementations; and c) a tendency to seek for generality over specificity in the results of the research. I call the latter the "saving-the-generalisation attitude" among empirical economists. The pursuit of generality has been originally intended to enhance the confidence on the truth of purported policy-oriented causal generalisations, but it now seems to be backfiring by leading to scientific results in the form of broad causal generalisations with no clear ways to be applied in specific real cases. Even if the existing empirical evidence could be taken as acceptable to support that there is "some" causal relation between trade liberalisation and economic benefits (usually in the form of an *average causal effect* estimated using well-established econometric techniques), I argue that the existing evidence is not conclusive at all (and probably not even appropriate) to support concrete policy recommendations for individual countries and their specific socio-economic conditions. In the final part of the chapter, I elaborate on my constructive proposal. I claim that there is a lot of information valuable from a policy-making perspective that is missing in the reporting of empirical research, I discuss this information in relation to the case at hand, and suggest that a stronger focus on this information (in accordance with the methodology of economic policy-making, as understood in this dissertation) will improve the reliability of economic science as a basis for attaining effective economic policy.

CHAPTER 5

INFERRING POLICIES FROM CAUSAL GENERALISATIONS

At the end of the previous chapter (section 4.4), it was suggested that the types of inferences about unit-level causal effects which can be made from causal generalisations vary depending on the homogeneity or heterogeneity of the relevant population to which the claim refers. When units in a population are heterogeneous in aspects that are causally relevant to the influence of a causal effect, it is not clear how one can infer anything reliable about the posited effect for particular units, exclusively on the basis of a causal generalisation. In this chapter I investigate whether there can be a methodical way of understanding which inferences about individual units of a population are allowed, and which ones are not, from the type of causal claims commonly employed in economic policy making.

The OECD research results were presented in the form of *causal generalisations*, such as "the lack of labour market flexibility causes unemployment" or "unemployment benefits cause unemployment". Causal generalisations are probabilistic causal claims that are meant to be true for a wide assortment of units and causal backgrounds (see Hitchcock 2001a). In the OECD case, policy recommendations to control unemployment levels were inferred from causal generalisations and offered as effective strategies to be implemented in individual OECD member countries (see OECD 1994c).

Similarly, in the research on the economic benefits of free trade (a case to be analysed in more detail in chapter 7) the theoretical and empirical results are presented in the form of causal generalisations such as "trade liberalisation causes investment", "trade liberalisation causes economic growth", or "trade liberalisation reduces income inequality". These generalisations are then employed in policy deliberations to justify prescriptions for specific countries such as "for Mexico, increasing trade liberalisation will increase capital investment, cause economic growth, and reduce income inequality"—which were among the reasons considered by Mexican economic authorities in favour of joining the North American Free Trade Agreement (NAFTA) back in 1994.

The fact that these causal claims about unemployment and about the benefits of free trade are probabilistic causal generalisations makes the main question of this chapter a genuine policy-related worry: *even if one knows that a causal generalisation is true, will this warrant its effectiveness guiding policy for any particular individual-unit case?* Or, to put it in terms of one of the examples above: even if one knows that the claim “trade liberalisation causes economic growth” is true, will that be enough to be certain that more trade liberalisation will cause growth in the context of any given individual community or country?

In practice, causal generalisations are hardly ever universal or exceptionless. On the contrary, as a consequence of the typical heterogeneity of the populations for which they are meant to be true and the pervasive unevenness of background conditions, they are often “irregular” (see Hausman 2010). Causal generalisations are taken as true only after making a great deal of simplifying and idealising assumptions regarding the populations of which they are about and the influence of background causal factors. Thus these claims wind up being true only in idealised settings in which all relevant disturbing factors are assumed to be fixed or absent or controlled for (see Mill 1844 [1830], 97-98; Cartwright 1989, chapter 5; Hausman 1992, chapter 8; Mäki 1992a, 1994; Reiss 2008, chapter 1).

The gist of this chapter is the following: causality accounts make a number of assumptions at the conceptual level in order to define the truth conditions for probabilistic causal generalisations. These assumptions refer to different specific configurations of the causal properties of the units in a population, and to the causal significance of background factors. Deciding on using one set of assumptions rather than another, and setting up the truth conditions in one way or another, determines—as I will show—different and particular conceptions of causality. That is to say, different methods of causal inference (i.e., methods to test for causality) have distinct causal concepts already built into them before any testing is done. When a causal generalisation is thus assessed and accepted as causal, the specific assumptions used in order to define and test its causal status strongly determine the inferences that can be made about how the posited causal influence actually would obtain in any particular unit or member of the relevant population.

The question about how to infer from established causal claims at the general-population level claims at the unit level is closely related to a relatively old philosophical question: how does general causation relate to singular causation? An answer to the latter question could be useful to answer the former, so I begin this chapter by looking at some existing attempts in philosophy to tackle it. Unfortunately, there seems to be no unambiguous conceptual framework accounting for how exactly causation at the so-called general level relates or connects to causation in singular cases. Still, the analysis presented in this chapter helps to identify more clearly a number of features in causal inference that are actually problematic (or restrictive) for practical and policy purposes.

In sections 5.1 and 5.2, I analyse how general and singular causation relate to each other within the framework of probabilistic theories of causality, and show how the assumptions required to set up truth conditions for general causation affect the inferences that can be made about singular cases. In section 5.3, I consider an improved version of the probabilistic account, proposed by Christopher Hitchcock (1993, 1995, 2001a) and argue that while it solves some problems of previous accounts it still fails to offer an appropriate answer to the main question of this chapter. In section 5.4, I claim that existing accounts place a strong emphasis on saving the truth of causal generalisations at the expense of saying almost nothing about how to infer the truth of claims of singular cases of causation. In section 5.5, I illustrate the type of assumptions required to establish the truth of causal generalisations in the methodology of causal inference commonly used in empirical social scientific practice, and in which sense such assumptions can be restrictive for policy purposes.

5.1. Two notions of causation

Traditional theories of causality separate the study of causation into the analysis of two distinct notions.

General causation (sometimes also labelled "type-level" or "property-level" causation) refers to relations reported by sentences like:

Smoking causes lung cancer.

Singular causation (sometimes also labelled "token-level", "actual", or "event" causation) refers to relations reported by sentences like:

Pedro's smoking caused his lung cancer.

Proponents of the distinct theories of causality have been inclined to separate the analysis of *general* and *singular* causation on the presupposition that they are two fundamentally distinct notions. In line with this, most accounts have focused on explicating either one or the other notion, leaving the inferential relations between general and singular causal claims relatively unexplored. For instance, David Lewis's (1973) counterfactual theory or Wesley Salmon's (1980) process theory were explicitly intended to characterise cases of token-level (singular) causation, and in turn were essentially not concerned with the relation between singular cases and type-level causation. Alternatively, probabilistic accounts of causality in their origins were explicitly devoted to the analysis of type-level (general) causation (e.g., Suppes 1970; Cartwright 1979; Skyrms 1980).

Moreover, there is no clear agreement among the few accounts that take a more tangible stance on how the two notions relate to each other. For example, proponents of the regularity account have typically interpreted cases of singular causation as instantiations of true general causal regularities, so that the truth of claims about individual cases depends on (or follows from) the truth of general causal laws (e.g., Davidson 1967). The opposite view has also been put forward, namely that claims about singular causation are true independently of there being true causal regularities, and that the truth of general claims depends on whether they are proper descriptions of patterns of cases of singular causation (e.g., Ducasse 1926; Dowe 2000; Glennan 2011). Still another view is that there is no special priority between singular and general causal claims, since they refer to two related but essentially different concepts of causation (e.g., Sober 1985; Eells 1991; Hitchcock 1995).

Elliott Sober (1985) and Ellery Eells (1991) have explicitly argued for a fundamental distinction between singular and general causal relations based on two distinct concepts of causation, each of which is said to require a different theory of causality in order to be properly characterised. According to Sober's (1985) argument, cases of general causation—which he calls cases of "property causation"—refer to relations between "properties of a population" and can be suitably analysed by means of probabilistic theories of causality (Sober 1985, 420). So claims like "smoking causes lung cancer" can be interpreted as "smoking increases the probability of developing lung cancer". However, the same probabilistic account—Sober argues—delivers ambiguous or counterintuitive results when employed to characterise cases of "token causation", which are causal relations between token events (Sober 1985, 407). To illustrate Sober's point, consider the following two famous examples.

> **Counterintuitive example 1**: A golf ball rolls towards the green with a high probability to drop in the hole. A squirrel passes by and kicks the ball diverting its trajectory and thereby reducing its probability of dropping in. However, the ball goes on rolling and drops in. Did the squirrel's kick cause the ball to drop in? Intuition says yes, but probabilistic theory of causality says no, for the kick reduced the probability of the ball dropping in.[41]

[41] The original version of this example is attributed to Deborah Rosen by Patrick Suppes (1970, 41).

Counterintuitive example 2: Sherlock Holmes walks into a valley as he is observed by Moriarty from the top of a cliff. Moriarty has a huge stone ready to be pushed down the cliff so that it will follow a specific trajectory with a high probability of killing Holmes. Watson being also at the top of the cliff realises that by pushing the stone at random, the probability of killing Holmes will significantly decrease, so he rushes to the stone and pushes it. The stone rolls down and kills Holmes nonetheless. Did Watson's push cause the stone to kill Holmes? Intuition says yes, but probabilistic theory of causality says no, for Watson's random push reduced the probability of killing Holmes.[42]

After considering variants of these examples, Sober concludes that there must be two distinct concepts of causation, "the first concerns the causal significance of *properties*, while the second addresses the causal significance of *token events*" (Sober 1985, 407). The probabilistic theory of causality which otherwise adequately characterises property causal relations seems to fail to account for cases of token causation, such as those presented in the examples, and hence some alternative theory seems to be required in order to analyse cases of token causation. Sober suggests that Wesley Salmon's (1980) causal process theory is a suitable candidate for dealing with cases of token causation (Sober 1985, 408), such that, in the examples above, it can then be true that "X type events *reduce* the probability of Y type events" at the property level, and simultaneously that "token event X initiates a causal process that results in the occurrence of token event Y".[43]

This type of argument has been used in the literature to make two basic points: a) singular and general causality are two distinct and independent causal concepts; and b) traditional probabilistic theories of causality are inadequate to characterise cases of singular causation and thus the two causal notions require different theories for their analysis. I will come back to these in section 5.3, after considering more closely how probabilistic accounts ended up focusing almost exclusively on saving the truth of type-level causal claims.

5.2. Type- and token-level causation in the probabilistic approach

As a result of the alleged complications related to the characterisation of singular causation, probabilistic accounts focused on the analysis of claims about general causation and on how to determine the truth conditions for such

[42] The original version of this example is by Irving J. Good (1961-1962).
[43] Ellery Eells arrives to a similar conclusion, namely that "token level" (singular) causation and "type level" (general) causation are two independent concepts that require different theoretical analyses, however he proposes a variant of the probabilistic theory of causality to analyse the cases of token-level causation (see Eells 1991, chapter 6).

claims. As it was mentioned before, this was done with no commitments about the approach being also suitable for analysing cases of singular causation (see, e.g., Suppes 1970; Cartwright 1979; Skyrms 1980). However, these proposals included at least implicitly a stance about the relation between type and token causal relations, since their strategy for defining necessary and sufficient conditions for the truth of causal claims consisted essentially in setting specific constraints on unit-level causal background conditions.

The basic schema of this strategy is as follows: For the claim 'X causes Y' to be true for population P, X must increase the probability of Y in "enough" individual units of P that share the same specific relevant causal background. Here 'the relevant causal background' refers to the set of all factors apart from X that can have a causal influence on Y, and which have to be held fixed (or controlled for) to assess the truth of the causal claim for P. Yet it is the specification of *how many individual unit cases* are "enough" where the various proposals significantly differ from each other.

The specification of the particular constraints imposed on the units at the token level amounts to taking a stance on the inferential relation between type-level and token-level. Different assumptions about how many unit cases are "enough" for the truth of general causal claims entail different inferential relations between the causal generalisations and claims about single-units. To clarify this point, let us consider three of the typical proposals of truth conditions for causal generalisations: contextual unanimity, Pareto dominance, and the fair-sample condition.

A traditional phrasing of the probabilistic theory of causality is that, for population P, event X causes event Y if and only if "the appearance of the first event is followed with a high probability by the appearance of the second, and there is no third event that we can use to factor out the probability relationship between the first and the second events" (Suppes 1970, 10). This can be expressed in terms of conditional probabilities:

"For P, X causes Y" is true if and only if $p(Y \mid X, Z) > p(Y \mid Z)$

Where X and Y stand for repeatable events, p(.) stands for a function of objective probability, and Z represents the complete set of additional causal factors that could affect Y in P (apart from X) and which have to be held fixed in order to make sure that "there is no third event" acting as a confounder. Population P is composed by individual units u_i, such that $P = \{u_1, u_2, \ldots, u_n\}$, so Z can be interpreted as a partition $Z = \{z_1, z_2, \ldots, z_n\}$ where each of the z_i

corresponds to a state description of the specific causal background (relative to Y) for each unit u_i.[44] The different proposals for the truth conditions for causal generalisations can then be understood as different proposals of just in how many unit-specific causal background contexts z_i the occurrence of X has to increase the probability of Y, for a causal generalisation to be true. Then the three proposals mentioned before can be described as follows.

Contextual unanimity condition: the claim "For P, X causes Y" is true if and only if $p(Y \mid X, Z) > p(Y \mid Z)$ in every z_i relative to population P (Cartwright 1979; Humphreys 1989; Eells 1991). If the population is causally homogeneous with respect to Y, then the same z_i characterises the causal background contexts for every unit of P, yet the condition can also apply to well-defined homogeneous subpopulations (see Eells 1991). In this account, the causal generalisation "For P, X causes Y" is true if and only if each single-unit claim "for u_i, X increases the probability of Y" is true for every unit member of P.

Pareto dominance condition: the claim "For P, X causes Y" is true if and only if $p(Y \mid X, Z) > p(Y \mid Z)$ in at least one z_i, and $p(Y \mid X, Z) \geq p(Y \mid Z)$ in all the remaining causal background contexts that are different to z_i (Skyrms 1980). In this account, the causal generalisation "For P, X causes Y" is true if and only if there is at least one member of Z for which the single-unit claim "for u_i, X increases the probability of Y" is true, and there are no members of Z for which single-unit claims of the form "for u_i, X decreases the probability of Y" are true.

Fair-sample condition: the claim "For P, X causes Y" is true if and only if $p(Y \mid X, Z) > p(Y \mid Z)$ in a "fair sample" of population P (Dupré 1984), where 'a fair sample' refers to a subpopulation of P selected with no bias with respect to the relevant causal factors in Z (p. 173). Ideally this requirement warrants an even-handed distribution of all the specific relevant causal background contexts z_i among the members of the sample. This account is typically interpreted in terms of a randomised controlled experimental design, since an ideal randomization would generate precisely a fair sample in the required sense (see Eells 1987, 111; Hitchcock 2001a, 224-225; Hausman 2010, 55-56). Then in this account, the causal generalisation "For P, X causes Y" is true if and only if the addition of all $p(Y \mid X, z_i)$ for all u_i in the fair sample for which X occurred is larger than the addition of all $p(Y \mid z_i)$ for all u_i in the fair sample for which X did not occur.

[44] This unorthodox way of defining Z as a partition corresponding to the units of P is only intended to keep the presentation expressed in terms of units rather than in terms of background conditions as much as possible. The essence of the different truth conditions for each of the proposals remains unaffected, which is what mostly matters for my purposes.

Which of these accounts is used to define the truth conditions for a causal generalisation determines and constrains the possible inferences one can make from the generalisation to individual cases. Consider the common example of "smoking causes lung cancer". What do the three proposals above really tell about the truth of single-unit causal claims? According to the contextual unanimity proposal, "smoking causes lung cancer" is true for P if and only if it is true that smoking causes lung cancer for each z_i for all the members of that population. Thus in principle, contextual unanimity could warrant that the truth of a single-unit claim for any particular u_i in P actually follows from the truth of the causal generalisation. Unfortunately, this would be the case exclusively under the assumption that P is a causally homogeneous population. For practical purposes, if the units of the relevant target population are not causally homogenous in relation to the variable of interest, then the truth of single-unit causal claims cannot be straightforwardly inferred from the truth of a causal generalisation. The contextual unanimity condition might work well as an ideal standard for the truth of abstract causal generalisations, nevertheless in practice many generalisations that are of interest for policy purposes are meant (and expected) to be true about heterogeneous populations precisely in order to be useful (see Hausman 2010).

According to the Pareto dominance condition, the generalisation is true if there are at least *some* u_i for which smoking causes lung cancer and there are no units in P for which smoking prevents it. In this case, the truth of a single-unit causal claim cannot be reliably inferred from the truth of the generalisation, since all one knows is that the causal claim is true for *some* units in P, but not necessary for all. The relevant population need not be causally homogenous for the causal generalisation to be true, and thus the claim might or might not be true for any particular u_i depending on its particular set of causal background conditions. Knowing that a causal generalisation is true—according to Pareto dominance—is simply not enough for inferring the truth of any particular single-unit causal claim.

Finally, according to the fair-sample condition, "smoking causes lung cancer" is true if and only if in an ideal randomised experiment, the frequency of lung cancer is larger among the units that smoke than among the units that do not. In this case, it is far more ambiguous than in the previous cases how exactly the truth of a single-unit causal claim is to be inferred from the truth of a generalisation. The causal influence referred to by the generalisation is an average measure of the experimental responses observed in the fair sample. Since the sample is "fair" by definition, the result can be generalised for the whole population as an average effect. Yet it is not clear how from the average probability of developing lung cancer as an effect of smoking—which has been estimated from a sample of the whole P—the probability of developing lung cancer as an effect of smoking can be inferred for any specific unit u_i.[45]

The sketches presented here should suffice to illustrate how the philosophical discussion about how general and singular causation relate to each other within the probabilistic theory of causality has focused on what the necessary and sufficient conditions are (in terms of single-unit cases) such that causal generalisations of the form "For P, X causes Y" can be true. However, the posited accounts do not tell much about how to infer the truth conditions for particular single-unit claims of the form "For u_i, X causes Y" given the truth of causal generalisations. Before moving on to analyse in more detail the implications of these theories being focussed on saving the truth of causal generalisations, I will consider in the following section a final and fairly neat philosophical attempt to connect general and singular causation.

5.3. Connecting singular and general causation: Hitchcock's attempt

Conceivably as a consequence of Sober and Eells's influence, most accounts of probabilistic causation (as those discussed in the previous section) have been explicitly devoted to analyse type-level causation, with no particular interest in, or strong commitment to, analysing token-level causation. But is it really the case that general and singular causation are two independent notions as Sober and Eells (and the examples in section 5.1) suggest?

Christopher Hitchcock's (1993, 1995, 2001a) revision of the probabilistic account of causality is an ambitious attempt to dissolve the strong dichotomy between singular and general causation, such that both notions can be characterised with no need of two independent theories. In this section I show how Hitchcock's proposal actually succeeds in conceptually connecting type-level causal claims with token-level claims within the same probabilistic theoretical approach. Nevertheless, the success of this approach depends—as I will show—on the inclusion of several strong assumptions about probabilities and about the causal backgrounds of individuals, which make the account rather stylized and unlikely to be applicable to concrete cases in the social realm. Be that as it may, Hitchcock's proposal is a clear illustration of how assumptions at the conceptual level in the characterisation of causal relations

[45] In fact, under the fair sample account, the more heterogeneous the population is, the less reliable any direct inference about unit-level effects will be. See discussion in section 5.5 on the assumptions of the potential outcomes framework.

determine (or restrict) the inferential relations between type-level generalisations and unit-level causal claims about concrete cases.

According to Hitchcock, even if one accepts that singular and general causal relations refer to distinct *causal concepts*, one can still challenge the idea that they need to be analysed by different *theories of causality*. He argues that the probabilistic theory (with some additions) can handle properly both notions of causation.

Hitchcock (1993; 1995) presents a refinement to the probabilistic theory of causality in which he suggests that causal claims should be understood as reporting comparisons between different values of conditional probability functions. Thus, they should be interpreted as reporting that the value for Y that a conditional probability function yields when X takes a certain value is larger than the value it yields when X takes some other specific value. Causes increase (or decrease) the probability of their effects *always in contrast to alternative causes*, instead of exclusively in contrast to the absence of the putative cause (Hitchcock 1995, 261). Hence, in this account the inequalities that describe probability increases can be rewritten as:

$$p(Y \mid X=x_1, Z) > p(Y \mid X=x_2, Z)$$

The apparent problems in the examples presented in section 5.1 vanish as soon as the contrast values for X are properly set. In those examples, instead of one sole inequality describing a type of event X decreasing the probability of its effect Y (an effect which actually ends up occurring), there are *two* inequalities to be considered: the first describes one type of event x_1 decreasing the probability of the event Y, *in contrast to* another kind of event x_2; while the second inequality describes the same event x_1 increasing the probability of Y, *in contrast to* an alternative event x_3.

$$p(Y \mid X=x_1, Z) < p(Y \mid X=x_2, Z)$$

$$p(Y \mid X=x_1, Z) > p(Y \mid X=x_3, Z)$$

The analysis of the relevant causal claims in the examples becomes unambiguous by interpreting the variables of the inequalities as follows:

For example 1	**For example 2**
Y=golf ball in the hole.	Y=Holmes crushed by stone.
x_1=squirrel's kick.	x_1=Watson's random push.

x_2=no kick.	x_2=Moriarty's well-aimed push.
x_3=squirrel's different more deviating kick.	x_3=Watson's push well-aimed away from Holmes.

Then from the first inequality $p(Y \mid X=x_1, Z) < p(Y \mid X=x_2, Z)$ one finds that x_1 events reduce the probability of Y to occur, thus x_1 events *do not probabilistically cause* Y in contrast to x_2 events. In terms of **example 1**, a squirrel's kick (x_1) reduces the probability of golf balls to drop into the hole (Y) in contrast to no kick (x_2). In terms of **example 2**, a random push (x_1) reduces the probability of crushing the target (Y) in contrast to a push well-aimed at the target (x_2).

Whereas, from the second inequality $p(Y \mid X=x_1, Z) > p(Y \mid X=x_3, Z)$ one finds that x_1 events increase the probability of Y to occur, thus x_1 events *probabilistically cause* Y in contrast to x_3 events. In terms of **example 1**, a squirrel's kick (x_1) increases the probability of the ball going into the hole (Y) in contrast to a different type of kick (x_3) which is more deviating than the first. And in terms of **example 2**, a random push (x_1) increases the probability of crushing the target (Y) in contrast to a well-aimed push away from the target (x_3).

As Hitchcock points out, different alternatives for X become significant depending on the appropriate context of the analysis (Hitchcock 1995, 269-272). Once contextual information is made explicit, the probabilistic theory can very well be used to analyse the allegedly problematic examples without reaching any of the counterintuitive results (Hitchcock 1995, 282). Thus it is possible to contest arguments using these examples to illustrate the limitations of the probabilistic theory of causation.

But what about the connection between general and singular causation? Notice that Hitchcock ultimately deals with the examples by making explicit the contrast class of the posited cause X, and then by offering two probabilistic interpretations about two distinct *general* (type-level) causal relations. So the question remains: is Hitchcock's refined probabilistic theory suited for analysing claims about *singular* causation as well? According to him, yes:

> This account applies at both the singular and general level. Singular causal claims describe probability relations between singular events (and also state that the named events occurred), while general causal claims describe formally analogous relations between generic event-types (Hitchcock 1995, 277).

The account is meant to apply to both notions by making the relevant specifications for the relata in terms of either "singular events" or "generic event-types". Thus both type of claims can be characterised within the same account. But even if Hitchcock's proposal also allows for the characterisation of claims about singular causation, *does it allow* for the assessment of whether it is true that a particular single event X has caused event Y or not? Hitchcock's account provides a probabilistic interpretation of cases of singular causation in terms of conditional probability functions, yet *it takes for granted that the singular cases are already true*. To make clearer the significance of this point, a brief elaboration on the details of Hitchcock's proposal will be helpful. Consider the following form of a probability function:

$$f_i(x) = p(Y_i \mid X_i = x)$$

The function is analogous to those employed for describing causal relations at the general-type level, but in this case, the indexing numbers 'i' attached to the variables stand for single units (or "individuals" in Hitchcock's account), thus the value of Y_i represents a single event occurring to unit u_i, such as "Pedro's developing lung cancer", and the value of X_i represents another single event occurring to u_i, such as "Pedro's smoking a certain amount of cigarettes per day". Following Hitchcock's account, claims about singular causation describe contrastive values of the above indexed type of conditional probability functions. For instance, the claim "Pedro's smoking 10 packs of cigarettes per day caused him to develop lung cancer (relative to not smoking)" would mean that: $p(Y_i \mid X_i = 10) > p(Y_i \mid X_i = 0)$, and that the two events standing as the causal relata have already occurred (see Hitchcock 1995, 279-280).

If the goal of this account is mainly to characterise token-event level cases in terms of probabilistic measurements, then Hitchcock's account offers a viable alternative to previous accounts. Nevertheless, it is not clear at all how this revised approach can be used in practice to connect both levels such that one can infer the truth of claims about singular events from claims at the general level. The problem I see in this respect comes from the formal assumptions that are required in order to define the proper conditional probability functions.

Broadly put, the assumptions required include idealisations about probability spaces for token events, in which, for instance, events can be repeated infinitely often, such that sequences and probabilities of token events X_i for each individual unit u_i can be well defined (see Hitchcock 1995, 277-278).

The major worry in relation to using this kind of idealising assumptions to formalise the account is that such assumptions are rather out of place when it comes to characterising the widely heterogeneous and in many cases unrepeatable events that are the subject matter of causal claims in actual social sciences (as the example in the previous chapter shows, see section 4.4). If there is no way to find a setting in the actual world at least vaguely resembling what the assumptions postulate in the formal setting, then any inference from the general level to the singular level could be formally sound and valid, but entirely unreliable in the real world.

But suppose, for the sake of the argument, that Hitchcock's analysis is practically applicable and suited to analyse causal claims at any level of analysis. What useful information can this account convey about the inferential relations between the two notions of singular (token-level) and general (type-level) causation?

The formal details and proof of the account are indeed indispensable for the accuracy of Hitchcock's proposal. As mentioned above, a crucial assumption behind his account is that the probability functions for singular and general causal claims are interrelated given particular postulations made about the structure of the *probability spaces* in which they are respectively defined. If the indexed conditional probability functions of the form $f_i(x) = p(Y_i \mid X_i = x)$ are all identical, i.e., if the causal influence of X to Y is identical from unit to unit, then it is possible to provide information about a "larger" probability space that would correspond to the not-indexed probability function $f(x) = p(Y \mid X = x)$, which is the kind of function reported by causal generalisations (see Hitchcock 1995, 278-281). Therefore, the probability function described by "For P, smoking causes lung cancer" can be constructed from information coming from the unit level probability space, but *only if* a "uniformity of nature" assumption is met: "that individuals will have identical probabilities of suffering from lung cancer, conditional upon their smoking the same amount, and being alike with respect to other causally relevant factors" (Hitchcock 1995, 281).

Hence what Hitchcock ultimately does is to develop a probabilistic requirement for characterising type-level causal claims, which is very similar to the *contextual unanimity condition* discussed in the previous section. Only that Hitchcock's version includes the refinement that a posited cause is always relative to another alternative cause, and provides a formal apparatus with which the probability values for causal claims at the type-level can be defined in terms of probability values at the token-level, which are also in principle well defined.

As valuable as all these achievements are for the goal of being able to analyse both notions of causation within the same probabilistic approach, there is a constraint on the inferential relations between the two levels built-in in the specification of the account. In order to infer the truth of a single-unit causal claim from the truth of a causal generalisation, the uniformity-of-nature assumption has to be true for the population of interest. In any other (more realistic) situation in which such assumption does not hold for the relevant population, Hitchcock's account fails to be informative about how to make inferences at the token level from true causal generalisations.

5.4. Saving the truth of causal generalisations is not enough

The main point to be taken from the previous sections is that inferences about causation at the unit level are dependent upon the particular specifications of the truth conditions used to establish causal generalisations. There are in fact *two main types of specifications* involved in the conceptual setting of causal generalisations which are (explicitly or implicitly) fixed a priori (i.e., before any empirical testing takes place), and which limit the inferences about whether 'X causes Y' in individual cases.

First type, conceptual assumptions about the causal characteristics of the relevant population $P = \{u_1, u_2, u_3,\ldots, u_n\}$ and about their specific unit-level causal backgrounds $Z = \{z_1, z_2, z_3,\ldots, z_n\}$. These assumptions define essentially how homogeneous (and on which relevant respects) the members of the population under study should be for the causal generalisation to be true. Hitchcock's "uniformity of nature" condition is an example of this kind of assumptions, for it states (as described in the previous section) that the same conditional probability function holds for all individuals u_i, at the unit level, given that they all are identically affected by the posited cause X, and given that they are all identical in all relevant Zs. This assumption is indispensable in the account to make the proper connection to the related general-level probability function (see Hitchcock 1995, 278-281). The uniformity of nature condition is in effect a description of how homogeneous certain causally relevant aspects of individuals u_i should be for the causal generalisation to be true about P. (As it will be illustrated in the following section, assumptions of this type are often required and used in empirical research in order to make standard causal inferential methods operational.)

Second type, the specification of the truth conditions for the causal generalisations in terms of causal influences at the unit level. These conditions define the criteria for discriminating spurious relations from genuine causation, but always in relation to a specific causal conception. This is precisely what the contextual unanimity condition, the Pareto condition, and the fair-sample condition are meant to do. They describe the sufficient and necessary criteria (in some ideal conditions) for a causal generalisation to be true. Since such criteria is defined in terms of the units u_i for which the posited cause X should increase the probability of the effect Y (conditional to their respective backgrounds z_i), these conditions—as described in section 4.2—allow entirely different inferences about causation at the unit level. (As it will be illustrated in the following section, standard inferential methods commonly used in empirical social research define truth conditions in this very same way, and thus they come with a particular notion of causation a priori fixed. As argued in chapter 3, different notions of causation in a causal claim in turn determine to different extents the inferences that are allowed from that claim.)

There is no clear answer—in the relevant philosophical accounts—to the important practical question of how to infer the truth of unit-level cases from the truth of causal generalisations. Most of the existing approaches focus on defining the appropriate conditions to warrant the truth of causal generalisations—which incidentally seems to be an attitude one also can find in empirical studies devoted to test posited socio-economic causal relations (see the examples in chapters 4 and 7).

During the philosophical debates in the 1980s about probabilistic causality, John Dupré (1984) famously suggested a hypothetical case (with the aim of undermining the contextual unanimity proposal) in which there is a rare condition present in some members of a population, such that if they have the condition, smoking make them less likely to develop lung cancer. In relation to Dupré's example, Ellery Eells made this remark:

> [A] person contemplating becoming a smoker, and trying to assess the health risks, should not be so concerned with the population frequency of that condition, but whether or not he has the condition. That is, the person should be concerned with which subpopulation he is a member of, the subpopulation of individuals with the condition (a population in which smoking is causally negative for lung cancer) or the subpopulation of individuals without the condition (a population in which smoking is causally positive for lung cancer) (Eells 1991, 104).

The idea emphasised in Eells's comment is comparable to the practical concern presented in the introduction of this chapter. From a policy-oriented point of view it would be much more useful to know about which particular conditions have to be met for a causal claim to be effective in relation to a concrete target of interest, rather than to know that a causal claim tends to be true in general. In other words, when a particular individual unit (e.g., an agent, a group, a country) has a decision to be made about achieving a certain practical goal on the basis of causal knowledge, the relevant concern is not necessarily (or exclusively) whether a causal generalisation is true or not for a whole population, but rather the concern is: given that the generalisation is true, how one can tell whether the posited causal relation will hold in an individual concrete case of interest.

Before moving on to the next section, let me put into words a brief preliminary reflection about the prospects of making policy analysis more reliable. Assuming for a moment that everything I say in this chapter is in fact perfectly well-known by philosophers of causality and probability experts, the issues I intend to emphasise are the following: are all the assumptions and conceptual specifications in the heart of causal inference always clear and explicit enough to all potential users of scientific results in the form of causal generalisations? Are all the potential users of such causal knowledge aware of all the intricacies and qualms about the outcomes of empirical scientific research?

Given that one knows that a causal generalisation has been assessed and established as true, it should be possible to make explicit all the relevant methodological assumptions that were required in the process of causal inference, such that all of them could be independently tested in relation to any concrete case of actual policy deliberation. One could increase the *level of confidence* in the expected effectiveness of a particular policy or strategy by—in addition of trusting that scientific causal claims are true—further evaluating whether (or how) the conceptual specifications employed in idealised settings also hold for the unit-level concrete cases of interest. Making all the relevant assumptions reasonably clear and explicit in this way—and then testing whether they hold or not for specific particular units in their particular background contexts—seems at least as a first approximation towards a procedure for obtaining more reliable information about whether a causal relation is true for particular target units.[46]

[46] This tempting and admittedly simple idea will be elaborated further in the context of the forthcoming chapters.

5.5. The potential outcomes framework of causal effects

I move now from the philosophical discussions on singular and general causation to the realm of the empirical methodology that social scientists use to make inferences about the truth of causal relations. In particular, I focus in this section on the counterfactual model of observational data analysis (see Morgan and Winship 2007), also known by statisticians as the potential outcomes framework of causal effects.[47] The aim is, on the one hand, to illustrate how conceptual and a priori methodological assumptions (similar to those described in the previous sections) are the backbone of one of the most successful methods of causal inference currently available in the social sciences. And on the other hand, to weigh up a key idea related to this framework, namely that once the assumptions required to establish the truth of a causal generalisation are made explicit, one can then proceed to assess whether they are valid, or reasonable, or realistic, or true about a particular individual member of a population and about its particular causal context.

The potential outcomes framework is often used in economic research to support policy recommendations such as those related to the OECD unemployment research, and to the research on the benefits of free trade briefly mentioned at the beginning of this chapter. As the study of these two cases shows (in chapters 4 and 7), when empirical methods are used in economics to establish causal generalisations, it is mostly only the final outcomes of the research (in the form of generalisations) that is offered to the policy makers as scientific guides for action.[48]

The model is composed by a population P of units u_i, such that P = $\{u_1, u_2, \ldots, u_n\}$; and variables representing the causal relata X and Y, which can take different values assigned by some measurement process for each given unit of analysis. Analogous to the logic of an experiment, a treatment variable and a response variable are defined on the basis of the particular aim of the investigation.

---

[47] Paul Holland (1986) refers to the framework as "Rubin's model of causal inference", since it was Donald Rubin who formalised it in a series of papers (see, e.g., Rubin 1974; 1977; 1978); while Rubin attributes the framework to Jerzy Neyman's original studies (see, e.g., Neyman 1935) on the "potential outcomes" of different treatments of land plots in agricultural experiments (see Rubin 2005, 324). For a "brief and selective history" of the potential outcomes framework, see Morgan and Winship 2007, chapter 1, sections 1.1 and 1.2.

[48] It could be argued that policy makers in fact demand such type of final results in the form of straightforward generalisations, so that they can use them quickly and easily as recipes for action without having to deal with the methodological complexities involved in their production.

Just as in Hitchcock's account, the effect of a certain cause is always the result of a contrast to an alternative cause (Holland 1986, 959). Causes are variables that can be interpreted as treatments to which the units of P can potentially be exposed to, and the response variable is a measurement of the difference between the outcomes produced by alternative treatments on the same unit. Assuming for simplicity a binary cause $X = \{x_1, x_2\}$, then for each unit u_i in P:

y_1 is the value of Y if u_i is exposed to x_1

y_2 is the value of Y if u_i is exposed to x_2

Then the unit-level causal effect of x_1 on u_i (relative to x_2) is given by the difference:

$d_y = y_1 - y_2$ unit-level causal effect

In principle, the main goal of the analysis is to measure the value of the causal effect at the unit-level, however this goal is frustrated by what Holland calls the "fundamental problem of causal inference": it is impossible to observe simultaneously on the same unit the value of y_1 and y_2 and, therefore it is impossible to observe the causal effect of x_1 on unit u_i alone (Holland 1986, 947).

A way of dealing with this problem is to focus on the population-level causal effect, rather than on the unit level, and to take for granted a number of a priori conceptual assumptions. Standard statistical analysis is used at the population level in order to calculate probabilities, distributions, and expected values. The probability of a variable taking a certain value is the proportion of units in which the variable takes that value, and expected values are calculated as average values over all members of P. The population-level causal effect is then defined as an average measure on the unit-level causal effects for P. The most usual average effect employed is the expected value of the difference $y_1 - y_2$ for all u_i. This specific causal concept is typically called the *average causal effect* (ACE) = $E(y_1 - y_2)$, which can also be expressed as:

$ACE = E(y_1) - E(y_2)$ average causal effect

Once defined in this way, the ACE can be estimated from observable data, since information from different units can be gathered to calculate $E(y_1)$ and $E(y_2)$ separately. The crucial move at this stage is that well-defined unit-level causal effects, which cannot be observed, are replaced by the population-level average causal effect which can be estimated. Causal inference proceeds from the observed values of X, either x_1 or x_2, and the observed values of the response variable Y, for each u_i. A causal generalisation obtained using this framework can then be interpreted as:

> For population P, $X = x_1$ (in contrast to $X = x_2$) has an average causal effect $d_y = y_1\text{-}y_2$ (see also Pearl 2000, 98; and Rubin 2005).

Notice that this account is analogous to the fair-sample account defended by Dupré, which uses an average effect in a "fair sample" of members of a population to infer whether there is a causal effect for the whole population (see Dupré 1984). And just as in the fair-sample account (reviewed in section 5.2), knowing that a generalisation about an ACE is true, it is not enough to automatically know what can be reliably inferred about the truth of any single-unit causal effect.

To see this more clearly, let me restate what is going on here: causal effects *cannot be observed* from a single unit (because of the fundamental problem of causal inference), but an ACE *can be estimated* from observed data on the outcomes of units exposed to x_1 and the outcomes of units exposed to x_2. Given certain assumptions (to be reviewed below), the ACE is then the *expected value* of the difference between being (potentially) exposed to x_1 and to x_2 in the population of study. However, from such expected value, it is not straightforward what the actual values of concrete unit-level causal effects would be for any individual unit.

The inferential process leading to an ACE requires a number of assumptions specifically intended to deal in one way or another with the fundamental problem of causal inference. To properly understand the meaning of an ACE one needs to understand the specific roles of the assumptions that are used to arrive at the final result. As it will be discussed below, some of them are indispensable to estimate an ACE in the first place, while also some of them are more essential than others in determining the inferential relations between the estimated ACE and unit-level causal effects. The main assumptions usually employed to estimate an average causal effect (ACE) are the following:

Temporal stability: The value of y_j, with j={1, 2}, is assumed to be independent of when the cause x_j is produced and measured in any particular unit u_i of P. The inferential framework assumes that the response on the same unit to the same cause would be the same response over time. Simply put, it does not matter at all when any particular unit is exposed to the posited cause (or to its alternative), the response will be the same for that same unit.

Causal transience: The value of y_j is assumed to be entirely independent from any prior exposure of u_i to the alternative cause, this means that the effect of the cause and the measurement that results in, say, y_1 does not affect u_i enough so as to affect the value of y_2 measured at a different point in time. Which is to say that it does not matter at all how many times any unit is exposed to the posited cause, the response will be the same each time for that same unit.[49]

Independence: The values of Y for each unit u_i are assumed to be independent of the process whereby units are assigned either to receive x_1 or x_2. A process of randomisation *correctly carried out* makes possible the independence between the assignment of X and the respective responses. If randomisation is possible, and thus the independence assumption is granted, then the ACE can be estimated from the available data on the observed variables.

Unit homogeneity: The value of the response variable y_1 is assumed to be equal for all units u_i, and the response variable y_2 is also assumed to be equal for all units u_i. Therefore, the unit-level causal effect d_y can be calculated by subtracting the y_2 observed in units that were exposed to x_2 from the y_1 observed in other units that were exposed to x_1.

Constant effect: the value of d_y on every unit of P is the same. If the variability of d_y is large among the members of P, then the ACE may not represent correctly the causal effect of specific units u_i. Simply put, the constant effect condition states that if every unit could be exposed simultaneously to x1 and to x2, *the difference* between the two outcomes will be the same for all units. This assumption can be seen as stating a form of causal homogeneity, and in this sense is analogous, e.g., to the contextual unanimity condition for the truth of causal generalisations (see section 5.2).

Stable unit treatment value assumption (SUTVA): the value of d_y of each unit u_i is assumed to be independent of any changes in the causal exposures to X of any of the other units in P (see Morgan and Winship 2007, 37-40). If this assumption does not hold, then the causal effect would be open to variation depending on how many members of P are exposed to the causal variable.

[49] Notice that if *temporal stability* and *causal transience* both hold, then $d_y = y_1 - y_2$ could be calculated by exposing the same unit u_i to x_1 (and measure its effect on Y) at some point in time, and then to expose it to x_2 (and measure its effect on Y) at a different moment.

From this list of assumptions, the first two—*temporal stability* and *causal transience*—are required to validate to a certain extent the inference from the unit-level causal effects y_d to the ACE at the population level. However, they are not indispensable in the sense that the ACE can be calculated from the existing data regardless of these conditions failing to hold strictly. They are mean to avoid that the ACE is biased. If temporal stability is not secured, the ACE could be including the effect of a confounder, i.e., the time at which units are exposed to X. If causal transience is not secured, then the number of times a unit has been exposed to X could be a confounder as well. In the ideal situation in which these assumptions hold, they make the ACE more reliable. Of course, in real-life cases, the situations described by these assumptions would seldom (if ever) be the case. Still, a way of understanding these assumptions from the point of view of a user of scientific causal results can be: the closest the actual situation of interest resembles situations of temporal stability and causal transience, the more reliable (or the less biased) the value of the ACE is as an estimate for the whole population.

Independence and *unit homogeneity* are indispensable to validate the inference of the ACE from the data on the unit-level observed outcomes. Independence is required to make sure that all potential confounders are distributed equally between the units exposed to cause x_1 and those exposed to x_2. In practice this is achieved by using different randomization techniques, and thus independence is dependent on the practical difficulties of actually achieving randomisation using the relevant technique in the population of interest. Independence is essential to secure that the average causal effect estimated is the genuine outcome of the posited cause X, rather than of any other potential disturbing factor. Lack of independence in practice, makes the ACE unreliable.

Unit homogeneity is a more debatable assumption. In practice, the higher the degree of unit homogeneity (i.e., the more similar the outcomes of exposure to x_1 or to x_2 are among all units), the better the ACE will represent the actual value of the unit-level differences between y_1 and y_2. Again this assumption is not indispensable to actually produce the estimate, but it makes the ACE more meaningful for potential inferences, so to speak. The better the ACE represents the actual outcomes of exposure to X for all units, the more confidence one can have on the inferences about unit-level causal effects.

Constant effect and *stable unit treatment value assumption* (SUTVA) are also meant to make the ACE a better representation of the actual difference between exposure outcomes (to the posited cause X). But what is more, they are indispensable for the validity of any inference about unit-level causal effects from the estimated value of an ACE. Thus these assumptions are extra important for practical and policy purposes.

As Holland explicitly puts it, the constant effect assumption "makes the value of the average causal effect relevant to every unit and, therefore, allows [the ACE] to be used to draw causal inferences at the unit level" (Holland 1986, 949). Hence, even when an ACE is perfectly well estimated, if the constant effect assumption does not hold in reality, one cannot make reliable inferences about causal effects for any particular unit of interest.

Similarly, the SUTVA states that the outcomes of causal exposure in the units are independent among each other. If SUTVA does not hold, the ACE is unstable among units in a very unsystematic way. Broadly put, the causal effect for each unit will be dependent on how many other units have been exposed to cause X, which will make the ACE entirely unreliable for inferential purposes. The example of causal generalisations on the benefits of free trade provides a good illustration of the significance of this assumption: the causal effects of a trade liberalisation policy in a country (almost by definition) cannot be entirely independent of the trade liberalisation policy implemented in other countries. In relation to trade liberalisation policies, it is precisely the concurrence of causal interventions in different units (countries) which determines to a great extent the resulting causal effect (see chapter 7).

The point of making these assumptions explicit is not to simply question them as unrealistic or false, but rather to understand that they are part of the meaning of a causal generalisation about an ACE. Accordingly, any user or consumer of this kind of causal knowledge could benefit from taking them seriously. As I have been repeating all along this chapter, the meaning of a causal generalisation depends (among other things) on the conceptual assumptions made a priori to characterise it. These assumptions carry essential information about when, where, and how causal concepts are supposed to obtain in the real world.

As Holland suggests, it would be "a mistake to conclude from the Fundamental Problem of Causal Inference that causal inference is impossible. What *is* impossible is causal inference without making untested assumptions" (Holland 1986, 959). This is perfectly acceptable when investigating or testing causal relations in isolation, e.g., when doing positive scientific economics (as described in chapter 1). However, such untested assumptions should later be put to the test whenever the aim is to implement an effective policy in a concrete target situation, i.e., when engaging in *the art of economic policy making* (see chapter 1). The potential outcomes framework is a method that focuses on establishing empirical scientific knowledge in the form of "causal generalisations", however to be useful for concrete policy purposes, its assumptions should be explicit and clear to the users of the science so that they can be tested in the context of the particular situations of interest.

The first point that should be made explicit about the potential outcomes framework is its emphasis on measuring the effects of causes rather than on investigating what the causes of given effects are (Holland 1986, 959). The method is meant to offer estimates of causal effects given that previous causal knowledge about the hypotheses under investigation is available or presupposed by the researchers. On this, Morgan and Winship comment that:

> To offer a precise and defendable causal effect estimate, a well-specified theory is needed to justify assumptions about underlying causal relationships. And, if theory is poorly specified, or divergent theories exist in the relevant scholarly community that support alternative assumptions about underlying causal relationships, then alternative causal effect estimates may be considered valid conditional on the validity of alternative maintained assumptions (Morgan and Winship 2007, 30).

Hence, any potential user of the results of this research should be well aware that the specifics of the theory and those of any underlying causal relations on which scientific causal hypotheses rely have to be independently tested. The fact that empirical results are sensitive to the theoretical presuppositions underlying empirical research is well known by scientist, but commonly neglected in the process of policy making (this issue will be reconsidered in chapter 6).[50]

For policy purposes, whenever the potential outcomes framework is employed to evaluate a causal generalisation, and then subsequently the same generalisation is employed as an input in a policy deliberation for the attainment of a desired effect in a particular unit-system (e.g., a city, a country, a community, or the like), the following steps will be advisable to take:

[50] The focus on the measurement of the effects of causes can be considered as a significant limitation of the potential outcomes framework for policy purposes, since in the policy making process the interest is not only on the exclusive causal influence of one cause on its effect, but rather on what all the relevant causes for the production of the desired effect are and whether they are or not present in the particular case of interest. Nevertheless, the framework is often employed to evaluate the "effectiveness" of policy claims (see Morgan and Winship 2007, 17-21).

a) Review exactly which methodological assumptions were made in order to deal with the fundamental problem of causal inference.

b) Evaluate how large the variability of the posited ACE is assumed to exist among the different units of P (this can be done in practice via sensitivity analyses, yet these analyses are not always offered to the policy makers as relevant information together with the research results).

c) Check out whether the assumptions that were made to calculate the ACE hold for the particular unit-system for which the policy recommendation is intended, for the degree to which these assumptions hold or fail to hold for the relevant unit will indicate how reliable the inference made about this unit can be.

The steps presented above are not enough for securing the truth of policy claims, for instance more has to be done in terms of getting the proper evidential support for policy claims. But carefully performed, these basic steps can help increase the reliability of the inference of particular single-unit policy claims from the truth of scientific causal generalisations.

5.6. Concluding remarks

In this chapter, I focused on the question of how causal claims about individual units of a population can be warranted by the truth of causal generalisations. After considering traditional probabilistic approaches to causality, I argued that the existing philosophical accounts about how to understand causal generalisations mainly focus on establishing the truth conditions for the causal claims, but do not say much about the inferential relations between true generalisations and claims about particular cases (i.e., the ones relevant for policy recommendations). The main point put forward is that the criteria for the truth of causal generalisations (defined under different assumptions) enables and constrains the inferences that are allowed from the general (or average) level to the individual unit case.

The analysis also shows that for every proposed specification of truth conditions for causal generalisations there are always some (more or less implicit) methodological assumptions that are required for the claims to be true. From this insight, I have proposed that to make better informed inferences from causal generalisations to concrete individual cases, these assumptions should be made explicit, and it should be investigated case by case how they hold in concrete contexts of application. The main idea behind my proposal can be expressed in simple language as follows: if you want to infer a concrete policy recommendation form a causal generalisation, tell me which assumptions were made to establish the truth of the generalisation at the conceptual level, and then I will tell you which aspects of your particular and concrete case have to be checked to make your inferences more reliable.

Finally, I illustrate the relevant methodological assumptions that have to be made explicit for concrete policy purposes by using the potential outcomes framework for causal inference. As it will be made clear with the case study in chapter 7, assumptions like unit homogeneity, constant effect, or SUTVA are usually required in empirical research to establish the truth of causal generalisations. Therefore, the implications of whether these assumptions actually hold or not in concrete cases of application have to be properly inspected each time one makes an inference to a concrete new case. In the following two chapters, I will refer to this framework again, and will elaborate in more detail on the proposal here only sketched. In particular, I will explain how the different sources of evidential support for policy making are related to testing whether methodological assumptions hold or not for target unit cases.

CHAPTER 6

EVIDENCE FOR CAUSALITY AND EVIDENCE FOR POLICY

In the previous chapter, I have argued that the conceptual assumptions required for defining and assessing the truth of causal generalisations determine the inferences that can be made about the truth of unit-level policy recommendations. If the population of reference, *P*, is not causally homogeneous in the relevant respects, then the truth of a causal generalisation does not automatically allow *reliable* inferences about the truth of the posited causal relation for particular units of *P* (let alone for units belonging to other populations different from *P*). In this chapter, I investigate in more detail the difficulties of inferring unit-level policy recommendations from general causal claims, and the ways in which empirical evidence can be used to deal with such difficulties.

I defend the claim that there are different roles for empirical evidence in relation to the process of policy making, which are often conflated in academic discussions on the practical relevance of economic science: one is the straightforward and direct role in establishing economic causal generalisations, another is a more indirect and often complex role in the process of making inferences about effective policy strategies on the basis of causal knowledge. The distinction between these two roles of evidence is mainly methodological and follows—as I will make clear in this and the next chapter—from the distinction between *scientific* inquiry and *policy* making defended in chapter 1. Evidence plays different functions in *the process of testing scientific causal claims* and in *the process of inferring effective policies* because these two activities have different epistemic aims which require rather different appraisal procedures.

There are some quite comprehensive philosophical accounts explicitly intended to tackle in general the use of evidence and scientific knowledge in relation to policy with a fairly normative attitude (see, e.g., Kitcher 2001; Douglas 2009; Mitchell 2009). In a much more focused line of research, philosophers interested in studying how scientific knowledge is employed to shape policy have directed their attention to a relatively recent case: the evidence-based movement. Since the 1980s, researchers in medical science (and later in other disciplines) have been advocating a more evidence-based approach to scientific practice, in contrast to traditional approaches which have been mainly based either on pure theory or on what are often considered low-quality and unreliable forms of evidence, e.g., narrative studies, unsystematic experience-based accounts, common sense, informal conventions, expertise, and so forth (see, e.g., Cochrane 1972; Guyatt, et al. 1992; Sackett, et al. 1996; Sackett, et al. 1997; Petty 2006).

In a nutshell, the main idea motivating the whole evidence-based approach is that empirical sciences should devote more effort to improving and systematising their evidence-gathering methods and evidence-evaluating standards. The aim is to make such methods and standards as "scientific" as possible, with "scientific" here meaning: conducive to high levels of accuracy combined with a minimum amount of subjective influence over the results (see Worrall 2007; Stegenga 2011).

Most philosophical discussions have focused on probing a number of procedural difficulties related to the evidence-based movement. They explore questions like: how can we comparatively evaluate the epistemic weight of different types of evidence? How can evidential rankings be objective? Is the *best* available evidence *always* the best in relation to all types of scientific problems? Is all the evidence labelled as 'non-scientific' dispensable? (see, e.g., Barton 2000; Worrall 2002, 2007; Ashcroft 2004; Vandenbroucke 2004, 2008; Borgerson 2009; Howick 2011; Solomon 2011; Stegenga 2011, 2013). In a rather huge contrast, it is difficult to find any piece of philosophical research discussing or offering a full-fledge investigation of the question: *assuming the best evidential methods and standards have been employed, is it really the case that having evidence-based scientific knowledge leads to more effective policies?*

An exception to this last remark can is the work of Nancy Cartwright. She has written in a variety of academic settings about the methodological and philosophical problems related to how the evidence-based movement provides or fails to provide good basis for effective policy. In her view, the evidence-based movement that has been taking over policy-oriented scientific research in the last decades (especially in health and medical practice) has wrongly placed a great deal of trust on particular evidential methods, such as randomised controlled trials (Cartwright 2009, 2010). But these types of methods are exclusively successful "at establishing *efficacy*: whether a treatment causes a given outcome in the selected population under the selected circumstances". When for policy purposes "we are interested in *effectiveness*: What would happen were the treatment to be introduced as and when it would be in the population of interest" (Cartwright 2009, 131).

In relation to the social sciences, Julian Reiss has recently argued specifically for a more evidence-based economic science (see Reiss 2004; and 2008, chapter 1), with the intention of supplementing what he calls a theory-based orthodoxy in current economics. In his view, evidence-based economics should integrate "individual socio-economic expertise with valid external evidence from systematic research relevant for the purpose at stake" (Reiss 2008, 13). In accordance with the general premises of the movement, in as much as causal claims in economics are supported by the best available evidence, by means of using the best available evaluating techniques, then they would more likely be claims about *genuine* causal relations. Nevertheless, even if an evidence-based economics of this sort could help making *scientific causal knowledge* more reliable, it remains to be seen whether (and how exactly) evidence-based economics would also help making *economic policy* more reliable.

In the first section, I unravel the main tenets of Cartwright's ideas on *evidence for use* and her ideas on the evidence-based movement (section 6.1), then I discuss some merits and deficiencies of her approach (section 6.2). In the third section, I suggest a way to rethink the ways in which evidence and evidential methods connect to *causal efficacy* and to *policy effectiveness* (section 6.3). In section four, I single out three methodological differences between causal hypotheses and policy hypotheses (section 6.4). In the fifth section, I adapt an existing typology of causal claims to characterise more precisely the notion of a 'policy hypothesis' (section 6.5). In the final section, I provide a sketch of what I call *the basic problems of policy inference* and briefly elaborate on the roles that empirical evidence plays and could play in dealing with such problems (section 6.6).

6.1. Deconstructing Cartwright's view on evidence for use

Nancy Cartwright has been concerned with the relation between scientific practice and public policy for over four decades. The scope of her ideas and worries on the topic go from calls to fellow philosophers of science to become more involved in practical social concerns (e.g., 1974, 1979, 1999, 2006, 2012a) to thorough criticisms of scientific methods of causal inference revealing how they fail to be of any use (or to provide reliable guidance) for practical action (e.g., 2001, 2007, 2010, 2011, 2012b). In particular—and in relation to this chapter—she has written on the uses and misuses of evidence and evidential techniques by researchers and policy-oriented organisations related to the so-called evidence-based movement. Without intending to be exhaustive, I will discuss here some of her most specific ideas about the use of causal inference and evidence for policy purposes.[51]

Cartwright's ideas about the use of evidence for policy can be grouped into two main lines of argument: The first is a general criticism of causal theories and methods of causal inference on the basis of their inherent limitations to provide reliable support to policy recommendations. The second is a much more specific assessment of the evidence-based movement, which focuses to a great extent on criticising the unjustified epistemic priority given to randomised control trials (RCTs) as the best evidence for policy purposes. These are two complementary but distinct discourses in Cartwright's theory of *evidence for use*. To use her own labels, I will refer to the first kind as criticisms against the "hunting of causes", and to the second as criticisms against the "using of causes".

6.1.1. On the hunting of causes

Cartwright's first significant attempt to characterise the relation between causal knowledge and practical relevance occurs in "Causal laws and effective strategies" (Cartwright 1979). In this well-known article, she presents a proposal of a theory for hunting causes and then attempts to formally connect it to practical guides for action.

In the first part, she draws on existing accounts of probabilistic causation (e.g., Suppes 1970) in order to develop a non-reductive account which could allow reliable discrimination of spurious correlations from genuine causation. Her theory is based upon conditioning on additional background knowledge and her version of a *contextual unanimity condition* (see previous chapter, section 5.2), so that "C causes E if and only if C increases the probability of E in every situation which is otherwise causally homogeneous with respect to E" (Cartwright 1979, 423). The formal presentation of her theory includes some of the usual conceptual assumptions about the whole set of relevant causal factors, other than C, in order to warrant the causal inference in an ideally conceived setting. These assumptions presuppose some causal notions, hence why Cartwright presented her account as a *non-reductive* theory of causality, which was later promoted under the slogan: "no causes in, no causes out" (see Cartwright 1989, chapter 2).

In the second part of the article, she offers a probabilistic conception of what would count as an *effective strategy* S in relation to a practical goal G. Her

[51] These ideas are presented, e.g., in Cartwright 2006, 2009, 2013; Cartwright and Efstathiou 2011; Cartwright and Stegenga 2011.

probabilistic definition of S has, among other allegedly desirable qualities, the virtue of being compatible with the following common intuition:

> There is *a natural connection* between causes and strategies that should be maintained: if one wants to obtain a goal, it is a good (in the pre-utility sense of good) strategy to introduce a cause for that goal. So long as one holds both the simple view that increase in conditional probability is a sure mark of causation and the view that conditional probabilities are the right measure of effectiveness, *the connection is straightforward* (Cartwright 1979, 431-432; emphasis added).

There is not much use in analysing the formal details of Cartwright's probabilistic theory for the purposes of this chapter. On the one hand, she has currently moved away from defending her old probabilistic account and has developed instead a tendency-law account of what she calls "capacities" (see Cartwright 1989, chapter 4). More recently she has also advocated a (radical) pluralistic account of causality (see Cartwright 2004). Furthermore, in her discussions on *evidence for use*, she casts some doubt on the potential policy virtues of the probabilistic account: when she criticises the major theories of causality for being incapable of producing causal knowledge that connects to real-life effective strategies (e.g., Cartwright 2001; Cartwright and Efstathiou 2011), one of the main targets of her attack is the probabilistic account.

On the other hand, regardless of the particular qualities and merits of Cartwright's theory of causality, her analysis of the connection between causal claims and effective strategies obviously falls into what Leuridan, Weber, and Van Dyck (2008, 298) characterise as "the standard view on the practical value of causal knowledge", namely the over-simplistic intuition that the practical relevance of a causal claim lies entirely in the fact that intervening on causes is a good way to bring about desired effects (I have argued against the *standard view* on several dimensions in chapters 2, 3, and 4).

In her more recent publications, however, Cartwright holds a position that is much more critical of the *standard view*. For example, her criticisms on why invariant accounts of causality fail to be relevant to support reliable policy interventions is entirely built on the idea that the *standard view* is problematic. Her criticism is, broadly put, that even if these causal accounts use the notion of *ideal* invariant interventions to establish causality, it does not automatically follow that the *concrete* interventions required to deal with a particular policy problem in real life will be warranted on the same grounds (see, e.g., Cartwright 2002; Reiss and Cartwright 2004).

In one of her first presentations of her *evidence-for-use* view, Cartwright (2006) claims that most methods of causal inference employed for *hunting* causes are successful in doing so, only because they are "clinchers", i.e., their causal conclusions can be deduced from their premises; yet they fail to provide guidance about *how to use* such conclusions. There are "on offer right now a lot of alternative accounts of what causality consists in: probabilistic theories of causality, invariance accounts, manipulation theories, causal process theories, and so on. Each, it turns out, is closely associated with one or another well-known method for establishing causal conclusions" (Cartwright 2006, 988), but these accounts do not tell much about *how* to apply such methods in real cases of interest, or about *when* it is appropriate to do so.

For example, the "fashionable Bayes-nets methods [...] lay down three assumptions about causality, then show that any time causes meet these three conditions, their methods will not give erroneous results" (Cartwright 2006, 988-989). When one takes a closer look at these three assumptions required to warrant causality, namely faithfulness, the causal Markov condition, and minimality, one then notices—as Cartwright does—that they rarely hold for or correspond to anything in reality: "Are there any even rough identifying features a system may have that will give us a clue that it is faithful or satisfies causal Markov or minimality?" (2006, 989).

Cartwright point is similar to what I have claimed in the previous chapter: the specifications of the truth conditions for causal generalisations come with a built-in characterisation of causality in the form of conceptual assumptions about causal criteria and relevant causal backgrounds. Similarly, methods of causal inference such as Bayes-nets methods (criticised by Cartwright) or the potential-outcomes framework (described in the previous chapter, see section 5.5) heavily rely on conceptual assumptions such as *faithfulness* and *minimality* (the former) or *independence* and *unit homogeneity* (the latter) to be successful in estimating causal effects. However, these assumptions wind up restricting the inferences one is allowed to make about causal effects for particular units of interest in concrete real-world situations.

6.1.2. On the using of causes

The case of the evidence-based movement allowed Cartwright to accomplish two things: on the one hand, philosophy of science had in fact a lot to contribute to the discussion of evidence-based medical practice at the end of the 1990s. Thus by engaging with a detailed analysis of the methodological and philosophical caveats of the movement, Cartwright was able to preach by example and direct her philosophical efforts towards questions of social policy relevance, just as she had been urging other philosophers to do. On the other hand, the overemphasised devotion to RCTs as the gold standard of scientific evidence within the evidence-based movement provided Cartwright with the perfect example of a method of causal inference that is very successful in hunting causes, while at the same time has a very limited potential for being informative about policy targets in non-experimental settings.[52]

For the purposes of this chapter, let me single out three general premises that align most evidence-based ventures:

> 1. Empirical evidence (and evidential methods) can be objectively ranked in terms of epistemic weight.
>
> 2. When a decision about accepting or rejecting a hypothesis has to be reached, it has to be based on the best available evidence.
>
> 3. Improving evidential standards by using the best methods will result in scientific knowledge that is more reliable to support practical and policy applications.

As regards to the original evidence-based movement in medical practice most discussions focus on premises 1 and 2. The main target of criticisms has been the implicit and not thoroughly justified view that in all objective rankings, randomised controlled trials (RCTs) are always the best available type of evidence (together with meta-analyses and systematic reviews, which combine results from RCTs). However, as most critics have argued, it is not obvious why RCTs would necessarily have a higher epistemic weight in comparison to other types of evidential sources (e.g., Worrall 2002, 2007; Ashcroft 2004; Borgerson 2009; Howick 2011).

There are some well-known methodological problems related to randomised experimentation. In theory, randomisation needs to be *perfect* in order to secure that all relevant causal factors apart from the causal variable under assessment are distributed equally among the sub-population to be

[52] Note that RCTs are not the most common method of causal inference in economics. Especially in policy-oriented branches of empirical (or applied) economics, like health economics, monetary economics, public economics, labour economics, international trade economics, economic growth, industrial organisation, environmental economics, and the like. Econometric regressions and simulations have been the preferred evidential tool in these sub-disciplines. However, there are some recent serious efforts to increase the use of experimental and quasi-experimental evidence to deal with policy issues in line with the evidence-based movement, for instance, in development economics, see Duflo, et al. 2004; Banerjee 2007; Banerjee and Duflo 2011; Cohen and Easterly 2009; and List 2011.

tested. In practice, the larger the sample used for a particular study, the more likely that randomisation distributes all the relevant known and unknown confounders equally among the different study groups. Thus while an infinite sample would be the conceptual benchmark for an ideal randomised control trial, actual RCTs can seldom use very large samples (see Reiss 2013, 203-204).

Moreover, practical difficulties in implementing perfect blind experimentation, or in monitoring the experiments without influencing them, are common methodological problems that could contribute to bias the results (see Shadish, et al. 2002; Guala 2005; Reiss 2013, chapter 10). And even in an impeccably well randomised, properly double blinded, and well monitored RCT some unknown confounders could "by chance" be unbalanced between the treatment and the control group producing unexplained significant differences in the result (see Worrall 2007, 1004-1006).

All these well-known facts about RCTs mean that, in practice, RCTs for many causal relations of interest might not be properly feasible, accurate, or reliable, simply because randomisation is not always sufficient to provide full control over confounders. This opens the discussion about the virtues of alternative non-experimental types of evidence, and about the supposedly objective grounds on which evidence rankings are constructed. It is just not clear why alternative sources of evidence are necessarily of less epistemic value than RCTs in all conceivable situations.[53]

Yet Cartwright's criticism of RCTs takes a clear detour from the discussions on evidential rankings and gold-standard status, to rather focus on the third premise of the evidence-based movement presented above, namely: the idea that using high quality evidential methods, such as RCTs, will produce more reliable strategies or policies to achieve concrete goals. Her main concern can be put as follows: even if one takes for granted that RCTs are successful methods for testing causal relations, the question is how can one tell whether the results of RCTs are of any relevance for the effectiveness of a proposed policy? (see Cartwright 2009, 127-128).

According to Cartwright, there are two criteria for good evidence: "First, the evidence must be credible: Evidence claims should be likely to be true. Second, the full body of evidence should make the conclusion probable, or probable enough given the size of the policy bet" (2009, 128). In relation to policy

[53] There are a number of studies on the epistemic virtues of observational studies (e.g., Black 1996; Benson and Hartz 2000), expert knowledge (e.g., Selinger and Crease 2006; Collins and Evans 2007; Martini and Boumans 2014), case studies (e.g., Gerring 2007; Ruzzene 2012), and other forms of qualitative evidence (e.g., Silverman 2001 [1993]), which contribute to cast doubt on the gold-standard status of RCTs.

effectiveness, she argues, the aim is to assess whether a particular "Treatment T will result in outcome O when implemented when and how it will be in the target situation/population" (2009, 129).

The evidence-based movement claims that RCTs are good evidence to support claims about implementing 'T in order to bring about outcome O', but Cartwright argues that RCTs only support "claims of one particular form, essentially, 'T causes O in particular circumstances X in particular population Φ'" (2009, 129). But if one wants to bring about O in a completely different population from Φ, how can one be sure that T will be effective there as well? This question is the centrepiece of Cartwright's *evidence for use*, which can be fairly put (using her terminology) in terms of another slogan: "we need to know not only what works, but what works for us".

Cartwright and many others often refer to this issue as the *external-validity problem*: "for what other populations can we expect these same conclusions to hold?" (2006, 986). Is it possible to extrapolate a result obtained in a particular controlled experimental setting to a different (non-experimental) setting and expect the same result to obtain? In line with her general criticisms of the methods for causal inference commented above, Cartwright claims that RCTs are good evidential methods for hunting causes, because they are *clinchers*:

> Given the right definition of 'ideal', it is possible to show that in an 'ideal' RCT a positive result deductively implies the conclusion under test: If there is a higher probability of O in the treatment group in an 'ideal' RCT than in the control group, it follows deductively that T causes O in the experimental population under the experimental conditions (Cartwright 2009, 129).

Yet, RCTs are just not good at warranting that their results will hold when the posited treatments are implemented in different populations or sub-populations on current normal conditions. In Cartwright's view, to make the move from the experiment to the real life situation "an inference ticket" is needed (2009, 131). But where can one get an inference ticket?

6.2. What works and what doesn't work in Cartwright's account

I take Cartwright's criticism of RCTs as parallel and, to some extent, complementary to my argument in the previous chapter. The problem of inferring a specific effective treatment for a concrete target from a RCT is analogous to the problem of inferring a specific policy recommendation for a concrete target from a causal generalisation. Thus I entirely agree with her general diagnosis: methods of causal inference commonly used in scientific research are often "clinchers" of causality, and they require certain unrealistic assumptions to be successful in evaluating causal results, which limits in significant ways the reliability of using any of such results as an effective guide for action in the real world. Good evidence for causation is not necessarily good evidence for policy. This is definitely an extremely relevant issue that has not received enough attention neither from academics nor from policy makers. Nevertheless, her proposals seem a bit lacking about how to deal with the issue.

The *evidence-for-use* view as put forward by Cartwright offers a well-developed diagnosis of the situation: it provides a systematic and quite persuasive account about the problem of using causal knowledge for policy purposes in the simplistic way suggested by the *standard view*, and then the problem is nicely traced back to the intrinsic features of existing methods of causal inference (mainly RCTs). But there is never a clear answer to the question: what to do then?

The issue is how can one know whether the result of a RCT performed in population A (in an experimental setting) will also obtain in population B (the non-experimental target population of interest)? Cartwright's answer then is that "[w]e need to construct a variety of causal scenarios" (2009, 135). Unfortunately, there is no elaboration of what exactly these causal scenarios are or how one is meant to construct them. On which causal grounds would these different scenarios be built? One could perhaps look at further causal information about population B before making any inference about whether the result of an RCT performed in A will obtain in B as well. But where should one look for such further causal information?

Cartwright's general idea is still very much in the right direction. The point she tries to emphasise with the idea of constructing alternative causal scenarios seems to be that, independently of how they are to be produced, one should think of them as different arguments that support a hypothesis H throughout different warranting steps. Then all the relevant steps leading to the ending point of a causal scenario need to be evaluated separately and one by one in order to assess whether the posited cause T will bring about effect O in different settings:

> A claim that supports a step in a scenario is relevant to H only if the scenario starts with T, leads to O or not O, and itself has sufficient probability to be taken seriously. But whether this last is true will depend on how well supported other steps in the scenario are, and that depends on what other claims support these steps" (2009, 135).

Indeed this procedure, whenever practicable, could be helpful to detect whether there is any logical flaw somewhere in the line of argument leading from the premises to the particular outcome of a potential scenario. Thus one could rule out from the set of available causal scenarios a number of them as invalid or incorrect. As potentially useful as this might seem, it still does not tell us "what will work for us", it only tells us "what will not work at all".[54] But how do we choose the right causal scenario to settle policy evaluations?

A more worrying aspect in Cartwright's evidence-for-use account occurs in relation to the rather striking claim that: "causation is in trouble". In a joint article, Nancy Cartwright and Sophia Efstathiou (2011) argue that "philosophic—and economic—accounts of causality" are in trouble because they are directly based upon the logic of specific inferential methods, such as experimental and counterfactual methods (Cartwright and Efstathiou 2011, 223-224). As Cartwright has argued before, the main problem with such existing accounts of causality is that they do not offer a "bridge" between *hunting* and *using* causes, yet here the argument goes a bit farther:

> What assures us that the knowledge we hunt at such great effort and cost can be put to the uses we want to make of it? To be *practicable* a theory of causation must *simultaneously* ground how we label features as 'causes' and the inferences we make once the label is attached. So we need a theory of causation that gives us *in one fell swoop* both *methods for inferring* causes and *methods for using* them (Cartwright and Efstathiou 2011, 224; emphasis added).

What exactly the aim behind this suggestion is or how exactly a proper theory of causation should then look like are issues not outlined in the article. Yet, there are some hints that Cartwright's account of *capacities* could be a suitable candidate to start building an ideal theory of causation that is good both for hunting and for using.

---

[54] An alternative way of understanding (or of developing further) Cartwright's suggestion could be to use causal models to generate economic policy simulations. Simulations could include all known relevant causal factors as variables with some representation of their specific causal influences and then, by including possible values, or probabilities of taking any values, for certain variables in the model, one could compute a hypothetical estimate of particular policy outcomes (see Adelman 1988; Gilbert and Troitzsch 1999; Law and Kelton 2000). Another form of generating "causal scenarios" has been proposed to model complex multilevel causal structures (with nested causal relationships) in the form of recursive Bayesian networks (RBNs). This modelling proposal is still a conceptual formalism with not many explored concrete applications; however, the final aim is to provide quantitative information about probabilities, as well as qualitative information about mechanistic structure and causal relations, which could then be used for *prediction*, *explanation* and *control* of actual systems (see Williamson and Gabbay 2004; Casini, et al. 2011).

The theories of causality considered by Cartwright as problematic are listed as: invariance interventionist accounts, "accounts based on 'Galilean' experiments (Giere 1979), Lewis-style ('miracle'-based) counterfactuals (Lewis 1973)", and probabilistic accounts (Cartwright and Efstathiou 2011, 224). Cartwright has made similar criticisms already in her earlier assessments of probabilistic, interventionist, and Bayes-nets methods for causal inference (see Cartwright 2007).

Again, all these theories and methods are problematic because they are good for testing "efficacy" claims, but not "effectiveness" claims. To move from efficacy to effectiveness, Cartwright argues, "we need an inference ticket. For a long time I have been selling inference tickets underwritten by an ontology of capacities" (2009, 131). Capacities are not traditional causal laws, but a sort of latent causal contributions which are inherent to certain factors independently of any contextual circumstances.

> The notion of capacity has three elements: (1) potentiality (capacities describe what a factor can do in the abstract, not what actually happens in full empirical reality); (2) causality (capacity claims are not claims about coassociation but about what results a factor can *produce*); and (3) stability (the ability to produce the effect in question must persist across some envisaged variation of circumstance) (Cartwright 1998, 45).

To know that X *has the capacity to produce* Y is, according to Cartwright, more useful for policy purposes than knowing that X *causes* Y, because the former claim implies that the ability of bringing about the effect is present in many different situations. Capacities are like Millian tendencies in that *their influence is always present* (whenever they are activated), yet their total outcome could be counterbalanced by other causal influences (disturbing factors) affecting the same target (see Cartwright 1998). Thus even if X has the capacity of producing Y, it will only result in the production of Y if X obtains together with the right configuration of additional causal factors that affect Y. Capacities are said to explain better than "causal laws" what underlies the process that brings about certain effects. Since one can be sure that a "capacity will produce its contribution whenever it is present (or is properly triggered), knowledge of capacities can be extremely useful" (Cartwright and Efstathiou 2011, 233). But how can one acquire knowledge about capacities in the first place?

One would expect that the usual methods of causal inference could be a way to find about capacities. For example, the result of a controlled experiment could be informative about the existence of a capacity that works (or manifests) in a particular setting of controlled conditions. But how can one know in which other settings it will manifest again or not? If this interpretation is correct, then all the problems related to the available methods of causal inference that Cartwright has criticised will apply to finding capacities as well.

As a matter of fact, capacities are said to be different from causal laws, among other things because a capacity can be measured "in various experiments but the experiments themselves are not enough to tell us that there is a capacity to be measured in the first place" (Cartwright and Efstathiou 2011, 233). So the manifestation of a capacity in an experiment is not enough to know it is there yet. But then if the usual methods of causal inference are not good at all to find capacities, how can one know that there is a capacity somewhere at all; a capacity which we can be sure "will produce its contribution whenever it is present"?[55]

Even accepting—for the sake of the argument—that having knowledge of capacities is possible, it is still not obvious how exactly capacities are supposed to be employed to dealing with the external validity problem. How can capacities be put into use to achieve effectiveness? Is the idea that by taking into account *all* pertinent capacities acting together in a particular situation, then one could secure the effectiveness of a policy implementation? These interesting insights definitely call for further elaboration.

After a whole section devoted to highlighting how capacities are much better than causal laws in several respects, Cartwright and Efstathiou make the following comment:

> [T]he most we can definitely predict with the kind of knowledge we usually have in these situations [i.e., situations in which we know there is a capacity] is that the [posited cause] […] *may very well* [bring about the expected effect] […]. So, even with capacities, predictive power is weak. But that is not the point here (Cartwright and Efstathiou 2011, 234).

But then what was the point? Was not the point to offer an account of causality that could offer *simultaneously* a way for *hunting* causes and a way for *using* them? In a subsequent paragraph the authors suggest that capacities are at least partially or potentially useful for the task (in an indeterminate way),

[55] Sherrilyn Roush makes essentially the same observation about Cartwright's capacities proposal when she comments that verification of capacities could be just as difficult as verification of causal factors (see Roush 2009, 141-142). Capacities are helpful to illustrate better "the transferability problem" (i.e., the external validity problem) and "what the problem of evidence-based implementation is. However, to take it as a solution requires supposing we have knowledge of capacities, and this is as hard as the problem was in the first place" (Roush 2009, 142).

and thus an account of capacities is still a better option than the existing causal accounts which are hopeless as *inference tickets* to move from efficacy to effectiveness:

> What matters for our worries about causal laws is that causal laws and capacities are entirely distinct. This is so even if one has a very different account of causal laws from Cartwright's. None of the accounts of causal laws currently discussed in philosophy look anything like an account of capacities; nor do our standard methods for testing causal laws serve well for establishing capacities. Capacities can be of use for predicting what happens when we set policy or build new technological devices. That, however, does not salvage causal laws. In introducing capacities Cartwright never meant to undermine causal laws. But focusing on the distinction between capacities and causal laws points up the problem in bold relief: we are very good at finding out about causal laws; but once we have done so, of what possible use are they? (Cartwright and Efstathiou 2011, 234).

Cartwright and Efstathiou's conclusions seem a bit misled in relation to the issue at hand. The alleged problem with theories of causality which makes the authors claim that "causality is in trouble" is that no theory offers an account that is *simultaneously* successful to hunt and to use causal knowledge effectively.[56] But apart from mentioning that "capacities" could be a useful notion that "points up the problem [of moving from efficacy to effectiveness] in bold relief", there is no elaboration on how exactly a proper causal theory is supposed to achieve this required move. Much more puzzling, perhaps: there is no elaboration whatsoever on how the proposed desiderata for what counts as a *proper* or "practicable" causal account is justified. Why does a "practicable" theory of causality *need* to offer "in one fell swoop" *both* inferential methods for testing causality and methods for practical use? Why simultaneously? And how would such a theory work?

My own suggestion about how to think of this issue follows form asking the question: is it really indispensable to get a causal theory that does it all? Is it really the case that existing causal theories as they stand are of no use to cope with this policy relevant issue? Contrary to Cartwright and Efstathiou's opinion, my suggestion is that it is not necessary to have a theory that is simultaneously

[56] The idea of simultaneously knowing how to asses and how to use causal claims has been Cartwright's crucial desiderata for causal theories already in most of her earlier criticisms of causal accounts: "Perhaps it seems an unfair criticism of our philosophic accounts to say they are thin on use. […] The problem is one we can see by comparing Hoover's approach to Simon with mine. What we need is to join the two approaches in one, so that we *simultaneously know* how to establish a causal claim and what use we can make of that claim once it is established" (Cartwright 2007, 5; emphasis added).

good for hunting and for using causal knowledge. Instead, we should understand that the process of discovering, justifying, and evaluating *causal relations*, on the one hand, and the process of conceiving, justifying, and evaluating *policy strategies*, on the other, *are two distinct endeavours* with different epistemic goals and thus contrasting appraisal methodologies. This distinction compels us, not to have a unique theory that simultaneously deals with the two endeavours, but to use all the relevant available causal knowledge (including, causal theories, inferential methods, causal concepts, background and contextual knowledge, and so forth) to deal separately with each of these two endeavours in the best possible way in relation to their respective aims. I will elaborate on this distinction in the following sections of chapter 6 and will offer a detailed illustration in chapter 7.

6.3. Evidence for causal relations and evidence for policy strategies: a preliminary distinction

The two endeavours of *hunting* causes and *using* them require testing two different types of hypotheses: hypotheses about causal influence (of the form 'X causes Y'), and policy hypotheses (of the form 'X is an effective policy to attain Y'). Testing both types of hypotheses means testing them for their truth. Even if there is a broad sense in which both types of hypotheses are "causal", the first are about *causal efficacy* whereas the second are about the *effectiveness* of strategies or policies in relation to concrete ends. Thus, in contrast to Cartwright's view, what needs to be recognised is that available theories and methods of causal inference play *different evidential roles* when the hypothesis under evaluation is a claim about causal efficacy and when it is a claim about policy effectiveness.

One and the same piece of evidence or a particular evidential method could play a specific role in evaluating a causal hypothesis and a totally different role in evaluating a related policy hypothesis. To clarify this, consider the following typology of three concepts of evidence, proposed by Julian Reiss:

> **Prima facie evidence**: "An observational or experimental outcome *e* is prima facie evidence for *h* if and only if knowing *e* is relevant (in a sense to be determined in a concrete context on the basis of background knowledge) for believing or acting on *h*" (Reiss 2008, 7).
>
> **Valid evidence**: "*e* is valid evidence for *h* if and only if *e* is prima facie evidence for *h* and all known sources of error in inferring from *e* to *h* have been controlled for" (Reiss 2008, 8).

Sound evidence: "An experimental or observational outcome *e* is sound evidence for an *h* if it is valid evidence for *h* and *h* is the relevant hypothesis for attaining purpose *p*" (Reiss 2008, 10).

This distinction only offers a very broad way of epistemically distinguishing types of evidence, since there could be further meaningful ways to sub-differentiate pieces of evidence within the same basic type (e.g., a piece of *valid evidence* could be "better" or "more reliable" than other pieces of *valid evidence*). But overall, the idea behind this typology is that the relevance of any piece of evidence has to be always evaluated in relation to a particular purpose. Accordingly, Reiss summarises the typology relative to different purposes: "prima facie evidence gives a licence to *investigate*; valid evidence gives a licence to *believe*; and sound evidence gives a licence to *act*" (Reiss 2008, 11).

Reiss's typology might seem quite clear-cut at first sight. If correct, one could immediately jump to the conclusion that what we need in order to connect scientific causal knowledge to effective policy making is to make sure we always employ *sound* evidence to support scientific hypotheses. *Sound* evidence would thus be, at least in principle, a solution to the *evidence-for-use* problems singled out by Cartwright. Unfortunately, the distinction is not as neat as it seems.

The move from prima facie to valid evidence is fairly unproblematic. For instance, let *e* be the observation of a correlation between X and Y, then *e* can be taken as prima facie evidence for hypothesis *h* that 'X causes Y'. This then allows (or induces) us to *investigate* further about *h*. Now let *e** be a properly designed and well performed randomised controlled trial (RCT) used to control for all potential sources of error in inferring from *e** to the same *h*, then *e** is valid evidence for *h*, and *e** allows us to *believe* in *h*. Putting this example in terms of statements we can see it as:

Hypothesis h: X causes Y.
Claim e: X has been observed to be correlated to Y.
*Claim e**: X has been observed to cause Y in an RCT.

Given that *e** is valid evidence for *h*, let us put forward a further claim such as:

Credence claim: There is valid evidence to believe that X causes Y.

Now the move from valid evidence to sound evidence is a bit more problematic. Notice that, in the previous example, *e* and *e** are distinct in relation to the same *h*. It is precisely that *e** is epistemically better than *e* in supporting *h* what makes the distinction reasonable in the first place. But, in accordance to Reiss's definitions, the move from valid evidence to sound evidence is not achieved by improving the quality of the evidence, but by making sure that "*h* is the relevant hypothesis for attaining purpose *p*" (Reiss 2008, 10). The notion of sound evidence seems to add a further dimension for the appraisal of evidence in addition to *evidential weight*, namely a *relevance* dimension. This breaks the continuity among the three types of evidence in terms of them providing different levels of epistemic support for *h*, and makes the notion of sound evidence rather ambiguous.

Given *hypothesis h* mentioned above, i.e., that 'X causes Y', then according to Reiss's definitions *at least* the following four claims need to be true for a piece of evidence to count as *sound evidence*:

*Claim e**: X has been observed to cause Y in a well performed RCT.
Credence claim: There is valid evidence to believe that *h*.
Purpose claim: There is a purpose *p* which is to bring about Y.
Relevance claim: *h* is the relevant hypothesis for attaining purpose *p*.

The *credence claim* is true in virtue of *claim e** being true; if *claim e** would be false (and there would be no other piece of evidence that could be taken as valid), then there would be no valid evidence for believing *h* to begin with. The *purpose claim* is a description of a postulated goal which could easily be tested as true or false in the context of concrete cases. So for the sake of the argument, let us take the first three claims as true. The question then is: how can one then tell whether the *relevance claim* is true?

By construction in this case, *h* is 'X causes Y', and *p* is 'bringing about Y'. Thus it seems tempting to say that *h* is obviously the relevant hypothesis for attaining purpose *p*. However, in real life things would almost never be as straightforward. Rarely practical purposes would coincide exactly with the posited effect of a causal claim. For instance, a purpose *q* could require a compound of effects Y, Z, and W, each one brought about in a particular way so as to generate a particular interaction among the three effects. In such a case, we could still claim that *h* is a relevant hypothesis to attain purpose *q*, and given that there is valid evidence to take *h* as true, then all the requirements for having sound evidence would be fulfilled. Yet it is not clear that such a case would offer any licence to act at all on the basis of *h*, for to bring about Y is clearly not enough to attain purpose *q*. But goals in policy making are usually extremely more complex than these conceptual illustrations. Not only they often are *multidimensional* (in the sense described in chapter 4, section 4.1), but also they commonly require the inputs from other known and unknown causal influences and additional background specifications to be effectively achieved (see chapter 2, section 2.2). In most concrete cases, policy goals would require a complex concurrence of causal effects triggered by some specific multidimensional (or compound) interventions.

It could be argued that the definition of sound evidence requires not that *h* is *the relevant* hypothesis, but only *a relevant* hypothesis for attaining *p*. Yet this makes the notion of sound evidence either trivial or more ambiguous, since everything can be relevant to everything in one way or another. But even if one concedes that a causal hypothesis for which there is already valid evidence could be unambiguously judged as relevant (in some well-defined way) to a particular purpose *p*, this does not straightforwardly license any definite acting upon that hypothesis. In most real life cases, the *h* tested for causal efficacy in the context of scientific research is simply a different hypothesis from the *h** tested for policy effectiveness.

As I have argued in the first part of this thesis, knowing that 'X causes Y' is true does not allow one to believe that 'bringing about X is an effective strategy to bring about Y'. Confusing between these two different hypotheses is precisely the kind of mistake that can follow from uncritically endorsing the *standard view* (see chapter 2). The implicit idea, in Reiss's notion of sound evidence, that the *h* to be tested for *validity* can simultaneously be the same *h* that would be relevant to effectively attain a practical purpose *p* is reminiscent of the basic intuition captured in the *standard view*.

In short, having valid evidence for *h*, plus knowing that *h* is relevant for 'bringing about Y' could be labelled as 'sound evidence' if that is preferred, and yet that *does not tell us anything* about whether the (entirely different) hypothesis 'bringing about X is an effective strategy to bring about Y' is true, let alone whether one could reliably act upon it.[57]

[57] It is not that Reiss's *sound evidence* necessarily implies the *standard view*, but the fact that the "relevance relation" between *h* and *p* and the characteristics of "purpose *p*" are left vague leaves the notion open to interpretations directly based on the *standard view*. The ambiguities in the characterisation of this notion risk mistakenly taking *valid* evidence for a causal claim as *valid* evidence for a claim about an effective strategy, when—as I will elaborate further—these two claims are in all significant respects different hypotheses.

My suggestion about how adapt this typology would be to drop the notion of ‘sound evidence’ all together, and simply make explicit that whenever causal knowledge is to be used as a basis for policy there are (at least) *two types of hypotheses under assessment*, a causal hypothesis and a policy hypothesis. Then for each of these types of hypotheses there can be *prima facie* or *valid* evidence. We can then focus on how to use our knowledge on causal inference to evaluate each type of hypotheses in the most appropriate way. To see this more clearly, consider the following two types of hypotheses:

h: ‘X causes Y’

*h**: ‘bringing about X is an effective strategy to bring about Y’

As mentioned before, claims like *h* are for most scientific purposes evaluated for *causal efficacy*. In relation to this type of endeavour, one can have prima facie evidence for *h*, which in turn will trigger further testing using customary methods of causal inference. Then assuming the causal tests are properly implemented, the results would provide valid evidence for believing *h*. But for policy purposes the type of hypothesis to be assessed is not *h*, but one like *h**, which is to be tested for *effectiveness* and on rather different grounds from those used to test *h*.

In relation to testing hypothesis *h**, one can have prima facie evidence for *h**, which could be, incidentally, a causal claim like *h*. More precisely, given that there is properly established scientific causal knowledge in the form of claims like *h* (all of them well-grounded on valid evidence), and given that such causal knowledge is *relevant* to the attainment of a particular policy goal that relates to Y, then all the available well-grounded causal knowledge can at the most aspire to count as *prima facie* evidence for *h**.

This is to say, even if there is valid evidence for *h*, *h* is only prima facie evidence for *h**.[58] At the most, *h* (the fact that there is valid evidence to believe *h*) gives a license to investigate *h**, but neither a licence to believe nor to act upon *h**. In order to be licensed to act upon *h** what is required is to get in turn valid evidence for believing that *h** is true. I will say more about the type of

[58] In turn, given that there is valid evidence to believe *h**, for example, after an actual policy strategy has been implemented in a concrete unit of analysis and the intended goal has been effectively attained, then such evidence for *h** should not be automatically taken as valid evidence for *h*. Unless all potential sources of error have been controlled for in relation to *h*, valid evidence for *h** could amount at best to be prima facie evidence for *h*.

evidence that can amount to valid evidence for claims like *h** after exploring further the differences between claims like *h* and *h** in the following section.

6.4. Differences that matter between causal and policy claims

I have been saying that claims about *causal efficacy* are different from claims about *policy effectiveness*. But how exactly are they different types of claims? And more importantly, in which way are the suggested differences significant to the problem of connecting causal knowledge to effective strategies? Broadly put, the methodological differences that I mean to point out determine the evidential standards upon which hypotheses are to be evaluated and accepted as true.

Consider the following two examples of hypotheses commonly studied and discussed in international trade economics:

h: Trade liberalisation causes economic growth.

*h**: Implementing a trade liberalisation policy in a country is a good strategy to obtain economic growth.

The first claim has the form of a causal generalisation, while the second has the form of a typical policy recommendation. These two propositions are similar in that they both can be taken as empirical hypotheses about a relation (yet to be specified) between trade liberalisation and economic growth. Both can be evaluated or recognized as true (or false) after considering all the pertinent evidence in accordance to the suitable appraisal criteria. But the similarities end there.

The statements above are different at least in three significant respects: **a)** the implicit epistemic aims for which they are studied, **b)** the standards for what counts as valid evidence in their assessment, and **c)** the intended scope of application. Making explicit these three dimensions is essential—or so I suggest—for a proper understanding of the roles of causation and evidence in real-life cases of socioeconomic policy making.

a) Different epistemic aims: this difference has been already put forward and discussed in a variety of ways all along this chapter, so I will just briefly restate it. The types of relations to which the claims refer are of a different sort. *h* is a proposition about the existence of a causal relation between some posited causal relata. The immediate epistemic aim behind investigating it is to find out whether the causal relation is actually there, and as much as possible about its whole causal structure and about how it works. In contrast, h^* is a proposition about the effectiveness of a particular policy in relation to a concrete desired outcome. The immediate epistemic aim is rather instrumental, i.e., to find out whether the posited policy will be successful—and if possible the extent to which it will be successful—to achieve an intended practical goal.

b) Different testing methodologies: h is true if 'trade liberalisation' is *causally efficacious* in relation to 'economic growth', where 'causally efficacious' is to be defined in accordance to any accepted criteria for causation put forward by traditional theories of causality (see chapter 3, section 3.1). In contrast, h^* is true if the implementation of a trade liberalisation policy in a concrete context of application is *effective* in relation to the attainment of its goal. The testing criteria for 'effectiveness' have to be set in terms of a certain level of confidence (to be contextually characterised) that the policy will generate the desired economic growth, given all the relevant background causal factors and conditions (and potential by-products) in the intended concrete target situation.

Causal efficacy and policy effectiveness thus require different evidential standards and procedures in order to be assessed. As it has been discussed in the literature on causation and causal inference, the criteria for establishing the truth of a causal hypothesis involve procedures that can help distinguishing cases of genuine causation from spurious relations (see, e.g., Mill 1843, 3.8; Simon 1954; Spirtes, et al. 1993; Pearl 2000; Hoover 2001; Shadish, et al. 2002). Whereas the suitable criteria for evaluating the truth of a policy recommendation should involve any procedures that could help warranting the achievement of the concrete desired effect. These procedures cannot be reduced exclusively to scientific methods, but to any formal or informal epistemic activity that could provide useful information about the reliability of the particular policy under deliberation in the particular context of implementation.

A more detailed elaboration on the complexities of the policy-appraisal methodology will have to wait until the concrete example analysed in the following chapter. However, I can state here that there is one clear procedural aspect that differentiates the methodology for testing causal hypotheses from the more intricate methodology for testing policy hypotheses. As mentioned before, testing causal efficacy requires the use of our theories and methods of causal inference with the aim of *controlling for* (abstracting from, screening off, or isolating from) *all potential sources of error*; whereas testing the effectiveness of a policy hypothesis requires the use of our causal theories and all useful evidence-gathering methods with the aim of *collecting and evaluating all the* *relevant causal information about any potential errors and disturbing causal factors* that could influence the intended policy effect in the concrete target population. Broadly put, in the first case the testing methodology consists in *isolating* a causal influence from all possible disturbances, whereas in the second case the methodology consists in *de-isolating* it to evaluate all the potential interactions in the context of a target situation.

c) Different intended scope of application: generality is a desirable property of causal claims, whereas specificity is often a desirable property of policy recommendations. "Trade liberalisation causes economic growth" is a causal claim that is meant to be true in general, i.e., for a rather large population of application (with countries being the units of the population in this case), even if the members of the population are to a great extent causally heterogeneous (see chapter 4, section 4.4). In contrast, the policy hypothesis "implementing a trade liberalisation policy is a good strategy to obtain economic growth in a country" is a unit-level type of claim in the sense that it is meant to apply to a single unit (or a specific sub-population).

Generality (broadly understood) is commonly regarded as a theoretical virtue of scientific knowledge. For example, in the appraisal of competing scientific explanations, a more general explanation is—other theoretical virtues being equal—often taken as a better explanation (see, e.g., Kitcher 1981; Cartwright 1983; Strevens 2004). Similarly, claims about scientific causal relations that are true along a wide variety of cases and situations are preferred for explanatory purposes over claims that are only true about a few individual or isolated events (see, e.g., Mitchell 2000; Woodward and Hitchcock 2003; Hitchcock and Woodward 2003).

For policy purposes, there seems to be also a certain demand for scientific knowledge of general validity in order to inform and guide the deliberations and recommendations made by policy authorities (for instance, the OECD and the countries' economic authorities in the example analysed in chapter 4). However, the final aim of a policy-making process is to bring about local effects for a particular unit-system in a particular context. Policy hypotheses like h^*, namely "trade liberalisation policy is a good strategy to bring about economic growth in a country" are unit-level claims. Thus, what really matters in relation to a policy hypothesis like h^* is not to know that a related scientific causal generalisation about free trade is true with a high degree of generality (which would be mere prima facie evidence for h^*), but whether there is *specific* valid evidence to believe that the unit-level claim *h** will be true for the concrete target unit in which a policy is to be implemented.[59]

6.5. Characterising policy hypotheses: a concrete proposal

What type of claims exactly are those that I have been calling 'policy hypotheses'? To understand in a more systematic way the differences between causal and policy hypotheses, let me briefly refer to a—frequently ignored—typology of causal claims put forward by Christopher Hitchcock.

I have discussed (in chapter 5) Hitchcock's probabilistic account of causality with the aim of understanding better the inferential connections between general (or type-level) causation and singular (or unit-level) causation. In the context of that discussion, Hitchcock argues that most traditional views on the meaning of causal claims conflate two different distinctions into the general-singular typology: on the one hand, there is a distinction between "narrow" and "wide" causal relations; and on the other, a distinction between what he labels "actual causation" and "causal tendencies" (Hitchcock 2001a, 219-220).

The *narrow-wide* distinction is based on the specification of the units of the population for which a causal claim is meant to be true. Hitchcock presents it as follows: "narrow" causality refers to causal relations that are true "within a single individual or a homogeneous population", whereas "wide" causality refers to relations that are meant to be true "within broader, more heterogeneous populations" (Hitchcock 2001a, 220).

The *actual-tendency* distinction is less clear-cut. At first, it seems to be based on whether X and Y have occurred or not. Cases of actual causation are thus relations between events that have actually obtained, whereas in the cases of so-called causal tendencies, the relata need not be actualised events (Hitchcock 2001a, 220). It is not entirely clear why Hitchcock takes the notion of "tendency" as the alternative to "actual". The only explicit comment on this respect appears in a footnote in which he says that the terminology for this distinction is taken from I. J. Good's (1961-1962) causal account, in which in turn 'causal tendencies' are understood as probabilistic regularities.

[59] I am taking for granted that causal generalisations are often the final goal of most scientific research and that they are among the preferred forms of scientific knowledge used as the basis of policy recommendations. This is the case in the two paradigmatic case studies I have chosen (see chapters 4 and 7). The same qualm applies for the idea that policy hypotheses are commonly claims about concrete policies being applied in the context of a specific target unit. Causal claims studied in scientific research could be about particular cases as well, like for example causal claims studied and assessed in history. And policies might very well be intended to attain certain abstract goals at a highly general level, like for example to reduce the level of inequality or of poverty in the whole planet. I simply do not focus on such special cases here.

Hitchcock's notion of causal tendency is indeed compatible with an interpretation of it as a probabilistic causal relation, such that whenever he writes that X "tends to cause" Y, he basically means that X *increases the probability* of Y.

An aspect that Hitchcock's actual-tendency distinction is intended to underline is that the truth of claims about actual causation is independent of the truth of claims about causal tendencies. A single case of actual causation does not entail the truth of a probabilistic causal tendency, for causal tendencies are relations among repeatable events (Hitchcock 1995, 265). The claim "Pedro's last dance caused his death" would be according to Hitchcock's distinction a claim about actual causation that could very well be true, and yet the causal tendency "dancing causes death" need not be true at all. Similarly, a claim about a causal tendency might be true, even if no related case of actual causation has ever occurred. For instance, "eating one kilogram of uranium-235 causes death" reports a true causal tendency in virtue of certain features of humans and the nature of nuclear reactions, yet no death has ever been caused in this way (see Hitchcock 1995, 265).

Hitchcock's main goal after presenting these distinctions is to argue that they cross-classify into *four distinct categories* of causal claims (see Hitchcock 2001a, 219-220):

(1) narrow-actual causation,
(2) narrow-causal tendencies,
(3) wide-actual causation, and
(4) wide-causal tendencies,

which, in Hitchcock's view,[60] correspond to *four distinct causal concepts* that are respectively reported by claims like the following:

(1) Pedro's smoking caused his lung cancer.
(2) Pedro is the sort of person for whom smoking tends to cause lung cancer.
(3) Last year, thousands of cases of lung cancer were caused by smoking.
(4) Smoking causes lung cancer.

[60] For an argument against Hitchcock's fourfold distinction as an unnecessary multiplication of causal concepts, see Jacob 2006.

This seems reasonably neat, yet a few qualms should be raised before using this distinction to clarify the nature of policy hypothesis. For instance, it is not clear from Hitchcock's definitions whether the essential basis for the narrow-wide distinction is the number of units for which the claim is meant to be true (Hitchcock 1995) or the fact that the units are causally homogenous or heterogeneous to a certain degree (Hitchcock 2001a). It seems safe to say at least that whether a claim is narrow or wide in Hitchcock's sense would depend entirely on the specification of the relevant population of application of a causal claim. As it was argued in chapter 4, a causal claim '(For *P*), X causes Y' could have different meanings depending not only on which causal concept is singled out as the relevant causal relation, but also on the specifics of the often implicit clause '(For *P*)'.

It is also not clear whether the actual-tendency distinction might be conflating some further distinctions. If by "actual causation" one understands a *success verb*—as Hitchcock suggests—then opposite notions to "actual" that are more appropriate than "tendency" could be, for example, "potential", "hypothetical", or "counterfactual" causation (this point is particularly important to characterise policy recommendations, so I will comment further on it below). On the other hand, if by "tendency" one presupposes a probabilistic relation—as Hitchcock does—then a more appropriate opposite notion to "probabilistic" could be "deterministic" (see Hausman 2010, 47-48).

Furthermore, it is not clear whether causal generalisations need to be always interpreted as "probabilistic" tendencies, their meaning can be analysed employing other non-probabilistic approaches, as it is done in some recent accounts on the meaning of "generic claims" (see, e.g., Leslie 2007) or of "ceteris paribus" generalisations (see, e.g. Earman and Roberts 1999; Strevens 2012). Claims such as "mosquitoes cause malaria" or "sharks eat people" are normally taken, *for practical purposes*, as uncontroversial true causal generalisations, regardless of the very small probability that a mosquito carries the relevant parasite or that a shark attacks and eats a human, hence the traditional probabilistic interpretations might not be enough to grasp all there is to the meaning of certain practically useful generalisations (see Leslie 2007).

According to Hitchcock's classification, causal generalisations like 'trade liberalisation causes economic growth' can be understood "as reporting a wide causal tendency" (Hitchcock 2001a, 220). As mentioned above, this means that they report probabilistic causal relations (between repeatable events) which are meant to be true for relatively broad heterogeneous populations. But which one of Hitchcock's notions suits better the form and contents of policy recommendations such as ‘implementing a trade liberalisation policy in a country is a good strategy to obtain economic growth’?

So far (and in previous chapters) I have been describing policy hypotheses as claims that are meant to be true for specific unit members of a population. Thus they can be taken as *narrow* claims in the sense that they are about one or perhaps a few causally homogeneous units. In the case at hand, the trade-liberalisation-policy hypothesis is intended to be true for a particular country. In so far as the claim need not in principle be about repeatable events, but only about a one-time event, the claim is not meant to report a “causal tendency”. Is it then a claim about “actual causation”? Not necessarily, since the posited causal effect in the policy hypothesis has not yet been actualised, but it is only *expected* to obtain after the policy is implemented. When testing a policy hypothesis, the question is not whether an instance of X *has caused* an instance of Y, but rather whether an instance of X *will cause* an instance of Y in the target setting.

The interpretation of “actual causation” as a *success verb* (as Hitchcock puts it) is thus not entirely appropriate for the relation reported by a policy recommendation. ‘For unit *u*, X *caused* Y’ is a generic form for claims about *narrow-actual causation*, but the type of claim to be characterised and to be assessed in policy-making cases has instead a form like ‘For unit *u*, X *will cause* Y’. Hence, in addition to the four notions proposed by Hitchcock, let me introduce a supplementary type of causal relation called ***narrow-potential causation*** which in turn is reported by claims of the form ‘For unit *u*, X *will cause* Y’. This additional notion seems to suit better the form and contents of typical policy recommendations.[61]

6.6. The basic problems of policy inference

I have argued (in chapter 5) that the main issue to be dealt with in relation to the process of designing and implementing effective policies on the basis of economic science is contained in the question: how can one infer the truth of unit-level causal claims from the truth of type-level causal generalisations? I

[61] Perhaps for some philosophical purposes to differentiate between ‘X *caused* Y’ and ‘X *will cause* Y’ could be somewhat irrelevant, and both claims could simply be treated as reporting the same kind of “causal concept” (and perhaps even be conceptually given the same necessary and sufficient basic truth conditions). However, if the purpose is to understand as clearly as possible the process of inferring policy recommendations that are *expected to be effective* at some point in the future, then the distinction between *actual* and *potential* is in fact a crucial issue.

will refer to the issue raised by this question as the *tendency-to-unit problem of policy inference*.

Given the following examples of claims *h* (about a wide-causal tendency) and *h** (about narrow-potential causation):

h: For *P*, trade liberalisation (*TL*) causes economic growth (*EG*).

*h**: For *u**, implementing a trade liberalisation policy (*TL* = *tl**) *will cause* economic growth (*EG* = *eg**).[62]

The tendency-to-unit problem of policy inference can be phrased as: How can one infer claims about *narrow-potential causation* like *h** from claims about *wide-causal tendencies* like *h*?

Assuming the following evidential claims are true:

Prima-facie-evidence claim:
TL has been observed to be correlated to *EG*.

Valid-evidence claim:
TL has been observed to cause *EG* in a number of well-performed econometric studies.[63]

Then the following *credence claim* is also true:
There is *valid evidence* to believe *h*, i.e., that *TL* causes *EG*.

The *credence claim* is precisely what most of the existing evidential base of empirical studies on the benefits of free trade can warrant at best (see chapter 7 for details). But contrary to what many agents involved in economic policy analysis and deliberation assume, all that empirical evidential base, together with any ideological or conceptual arguments that could contribute to support the truth of *h*, are *nothing but* prima facie evidence for the policy hypothesis *h** which says that: 'For *u**, a trade liberalisation policy (*TL* = *tl**) *will cause* economic growth (*EG* = *eg**)'.

The *tendency-to-unit problem of policy inference* is akin to the problem of *external validity* emphasised by Cartwright and many others as one of the main

[62] Where the superscript '*' refers to a specific policy-target situation.
[63] In the case of international trade liberalisation, most of the empirical research consists in econometric studies using national or cross-national databases. Some of these studies can be very local, i.e., they investigate the causal effects of trade liberalisation at the sector, regional, or national level; while others are rather broad in their scope trying to estimate average causal effects for a large amount of (causally heterogeneous) countries. For more details and references to these empirical studies, see chapter 7.

problems of RCTs, and by implication of the evidence-based movement (see sections 6.1.2 and 6.2 above). The issue raised by Cartwright is: how is it possible to reliably infer that a result that *obtained* in a particular controlled experimental setting *will obtain* in a different (non-experimental) setting? As it has been noted by some authors, the so called external validity problem is not exclusive to RCTs (see, e.g., Roush 2009, 139; Reiss 2013, 205). Any empirical study (experimental or not) that is restricted to analysing data about a particular case in a specific (controlled or natural) setting can be subject to the same inferential misgivings. This issue can also be put in terms of the causal notions depicted in the previous sections in order to understand the relevant evidential relations. Let me refer to the issue as the *unit-to-unit problem of policy inference* and describe it as follows.

Given the following examples of claims h^c (about narrow-actual causation) and h^* (about narrow-potential causation):

> h^c: For u^c, implementing a trade liberalisation policy ($TL = tl^c$) *caused* economic growth ($EG = eg^c$).[64]
>
> h^*: For u^*, implementing a trade liberalisation policy ($TL = tl^*$) *will cause* economic growth ($EG = eg^*$).
>
> With $u^c \neq u^*$ and assuming that the following evidential claims are true:
>
> *Prima-facie-evidence claim*:
> TL has been observed to be correlated to EG (in country u^c or in any other country that is not u^*).
>
> *Valid-evidence claim*:
> ($TL = tl^c$) has been observed to cause ($EG = eg^c$) in a well-performed empirical study using all the relevant data about country u^c.
>
> Then the following *credence claim* is also true:
> There is *valid evidence* to believe that TL caused EG in country u^c.

Again, valid evidence for h^c can only aspire to be *nothing but* prima facie evidence for h^*. Thus the usual external-validity concerns for policy purposes can be put in terms of the unit-to-unit inferential problem:

---

[64] Where the superscript '^{c}' refers to the *controlled setting* of a well-performed empirical study.

The unit-to-unit problem of policy inference: how can one infer claims about *narrow-potential causation* like h^* from claims about *narrow-actual causation* like h^c?

Both issues are highly problematic for the inference of effective policies from causal knowledge, *particularly when the populations of interest are causally heterogeneous*. The specifics of the evidential base can certainly be much more complicated than the rough outline presented here. For instance, a large amount of studies about the same unit of analysis u^c (or about causally homogenous units) could be combined in order to give "stronger" evidential support to a *distinct* causal hypothesis such as '*TL tends to cause EG* in country u^c'.[65] In Hitchcock's typology this would count as a claim reporting a *narrow-causal tendency*, which in turn could be taken as prima facie evidence for h^*. Then again, regardless of there being valid evidence allowing us to believe that there is a *narrow-causal tendency* from *TL* to *EG* in u^c (or in countries that are causally homogenous to u^c), a variation of the *tendency-unit problem of policy inference* would emerge as soon as anyone would want to infer a unit-level policy recommendation for country u^* on the basis of the *causal-tendency* claim about u^c.

So how to get *valid evidence* for unit-level policy hypotheses like h^*? I will only begin delineating an answer to this question by means of the specifics of the case study in the following chapter. Suffice it to say here, the valid evidence required to be allowed to believe in policy hypotheses—like h^*—requires the collection and processing of a whole new evidential base. Such new evidential base will be composed by all available pieces of information that could support further *credence claims* about the status of all the relevant background and disturbing causal factors (and unintended *by-products*) that could have an effect on the causal influence of *TL* on *EG* in the specific target-unit of interest.

6.7. Concluding remarks

A causal generalisation is a proposition of the form "X tends to cause Y", whereas a policy recommendation is a proposition of the form "X is an effective strategy for Y in a specific causal context". These two types of propositions are often put forward as similar hypotheses in need of evidential support; however,

[65] This is essentially what systematic reviews and meta-analyses are meant to do with the results of experimental studies, and meta-regressions with the results of econometric studies. As it will be discussed in the following chapter, these type methods of evidence amalgamation are commonly used to offer more *robust* evidential support to economic causal generalisations (i.e., claims about *wide-causal tendencies*).

they differ in three important respects: First, the two types of hypotheses commonly aim at different purposes: establishing scientific knowledge and designing policy recommendations in relation to concrete goals. Second, the criteria for establishing the truth of causal knowledge involve procedures that distinguish cases of genuine causation from spurious relations, whereas the criteria for the truth of policy recommendations involve procedures that warrant the achievement of a concrete desired effect in a concrete context of application. And third, generality is a desirable property of scientific causal knowledge, whereas specificity is often a desirable property of policy recommendations.

In so far as causal claims differ from policy recommendations in aims, testing criteria, and intended scope, then the required evidence (and the ways it is used) to support them also differs. This implies different methodological approaches. Economic research that aims at supporting causal hypotheses requires evidential methods that generate results in isolation from all potential sources of error, e.g., confounders, experimental bias, institutional frameworks, and so on (see chapter 2). Thus, a causal hypothesis "X causes Y" within a particular causal structure including other causal factors Z requires evidence that X causally affects Y independently of any influence from the relevant set of factors Z. In contrast, research that aims at supporting policy recommendations requires the use of evidential methods to generate results not only about causal relations in isolation, but also about which potential sources of error (from all the known and unknown factors Z) are present and active in the intended concrete context of policy application. Thus, the assessment of a policy hypothesis "X is effective to generate Y in a concrete context" requires *valid* evidence that X *tends to cause* Y, plus *valid* evidence about whether all the relevant factors Z (pertaining to the relevant causal structure) are or are not active in the concrete intended context of application.

CHAPTER 7

THE ROLE OF EVIDENCE IN ECONOMIC POLICY MAKING: FROM CAUSAL GENERALISATIONS TO TRADE POLICY

One of the main points I have been making in this thesis is that accepting as true a causal generalisation is not sufficient to infer the truth of any policy recommendation that derives from it. Causal hypotheses like "Trade liberalisation causes economic gains" require different evidential support from the support required by policy hypothesis like "For country A, increasing trade liberalisation will cause economic gains". As suggested in the previous chapter, having *valid* evidence to believe in a causal hypothesis is not sufficient at all to believe in (let alone to act upon) any purported policy implication. To confuse these two types of hypotheses or, more precisely, to take the available *valid* evidence for the former as *valid* evidence for the latter is a big and rather common mistake in economics.

Presented this boldly, it might seem obvious that a causal claim of the form 'X causes Y' and a policy recommendation of the form 'implement X in order to attain Y' are not the same type of hypotheses, and that their evidential requirements thus differ. Still, the confusion is pervasive in the economic literature, even if in a fairly subtle way: knowledge that 'X causes Y' is true is frequently taken as "valid" (and enough) evidence for believing also that 'X is an effective strategy to attain Y' in a given situation. In what follows, I illustrate this problem in relation to the scientific research on the economic benefits of trade liberalisation, and suggest a potential way to attain the proper evidence to support the effectiveness of economic policy recommendations.

In section 7.1, I offer some examples of the implicit intuition that testing causal generalisations is relevant because, if they are found to be true, then automatic policy implications would follow. In section 7.2, I offer a brief summary of the history of new international trade theory, only to point out why such a new theoretical development has not been able to settle the debate on the benefits of free trade in the theoretical arena. Sections 7.3 and 7.4 describe the general hypotheses commonly studied in the empirical literature on the benefits of trade. Sections 7.5 and 7.6 contain the bulk of the discussion of variables and the econometric methods used in the existing empirical research. Section 7.7 discusses three methodological aspects of the empirical literature that I claim have negative (or at least ambiguous) consequences for policy purposes. In section 7.8, I build a link between this case study and several aspects already argued for in the previous chapters of this dissertation, and also illustrate a potential way of doing things better, i.e., a form of using more efficiently the available causal knowledge and evidence that economic science has to offer in order to improve the reliability of the process of policy making.

7.1. FREE TRADE CAUSES ECONOMIC BENEFITS: A SCIENTIFIC GENERALISATION OR A POLICY GUIDELINE?

Both theoretical and empirical studies on the economic effects of trade liberalisation usually devote a few lines or paragraphs to state their policy relevance. Such types of comments are often bold and brief, and occur in the introduction or the concluding sections of a study. The aim is to offer the reader at least a hint of the way in which the research and the results of the research could have some real-life policy implications.

The various ways in which international economics articles are justified on (superficially argued) policy-relevance grounds have been exposed by William Milberg (1996). He identifies a number of rhetorical devices used by professional economists to make their articles look more "policy relevant". Without going into the details of all distinct rhetorical moves, the point to keep in mind is that most of the scientific research on the economic effects of trade liberalisation embraces the idea that studying the topic is important because the results would have straightforward implications to guide trade policy.

In an article summarising the theoretical progress in so-called 'new international economics', Robert Feenstra (2006) begins by reminding the reader of the main practical impetus behind the development of new trade theories:

> [T]he models of economies of scale and monopolistic competition were conceived with a very practical application in mind, namely, *the gains that would result from large-scale tariff reductions*. Whether from multilateral tariff reductions under regional trade agreements, *these models predicted gains from trade over and above the gains from specialization in conventional models* (Feenstra 2006, 617-618; emphasis added).

This is a standard idea among the supporters of the new international economics: the suggestion that the theory is practically relevant because it explains why "tariff reductions" to international trade are good. The approach is commonly justified by claiming that a theory that includes economies of scale and monopolistic competition is not only intuitively more realistic, but also better supported by historical evidence, and therefore such approach allegedly *explains* and *predicts* better than previous accounts the main benefits of trade liberalisation policies (see Feenstra 2006).

In empirical studies, the policy-relevance motivation is commonly stated more concisely and essentially as follows: it is important to empirically test causal hypotheses on the effects of trade liberalisation, because implications for policy making would automatically follow. For instance, in a rather influential study, Francisco Rodríguez and Dani Rodrik (2001) state explicitly what they take as the main question motivating empirical research on international trade:

> Do countries with lower policy-induced barriers to international trade grow faster, once other relevant country characteristics are controlled for? We take this to be the central question of policy relevance in this area. To the extent that the empirical literature demonstrates a positive causal link from openness to growth, *the main operational implication is that governments should dismantle their barriers to trade* (Rodríguez and Rodrik 2001, 264; emphasis added).

As I will emphasise below, Rodríguez and Rodrik's article is in fact mainly devoted to provide methodological criticisms of the empirical research on the benefits of trade liberalisation. Still, the quote above shows that, even in an otherwise critical article, the authors are very much in agreement with the widespread conception within the discipline that the *causal generalisation* "trade liberalisation causes economic growth" is a crucial research hypothesis—"once other relevant country characteristics are controlled for"—because to the extent to which it is empirically demonstrated, then the *policy recommendation* that "governments should dismantle their barriers to trade" would naturally follow.[66]

In addition to academic articles, there are also specially commissioned policy-oriented research studies and reports put together by international organizations such as the World Bank, the OECD, the IMF, and the like, which are explicitly meant to inform and offer policy prescriptions to authorities in real economies. Usually presented in the form of systematic reviews, these institutional reports tend to summarise in a few lines the main conclusions of many different pieces of available scientific research. The final product is a

---

[66] To be clear, I am not claiming that Rodríguez and Rodrik (or any other theoretical or empirical international economist) ultimately believe that *causal generalisations* automatically warrant *policy hypotheses*. I only want to emphasise that whenever economists superficially justify the relevance of their scientific research by hastily stating the potential (usually abstract) policy implications of their results, they let the door wide open for confusion.

(selective) compilation of empirical results presented as an evidential base supporting the view that trade liberalisation is a beneficial trend to be followed by most national economies in order to prosper. Some examples:

> More open and outward-oriented economies consistently outperform countries with restrictive trade and foreign investment regimes (OECD 1998, 36).

> Policies toward foreign trade are among the more important factors promoting economic growth and convergence in developing countries (IMF 1997, 84).

In this kind of reports, the jump from a scientific evidential base to the formulation of straight-forward policy recommendations is rather obvious. After all, one of the main services that such international organizations are meant to provide to policy makers is precisely some guidelines that could be straightforwardly implemented to improve socioeconomic performance (see also the case in chapter 4).

The main (positive) goal of pointing out the existence of this confusion—i.e., the unjustified jump from well-grounded scientific knowledge to the inference of policy recommendations—is to explore the potential ways to correct it. Thus, the constructive idea motivating this chapter is to explore in which ways policy recommendations can be better supported by well-grounded scientific knowledge so as to better warrant their effectiveness.

7.2. New trade theory and the benefits of free trade

During most of the 20th century, international trade theory was based upon the assumptions of *constant returns to scale* and *perfect competition,* and the idea that the notion of *comparative advantage* was the crucial determinant of trade patterns among nations. Since its origins, the idea of comparative advantage has been presented as an argument intended to illuminate the benefits of international trade and the disadvantages of enforcing a closed economy (see Ricardo 1821 [1817], chapter 7; Mill 1909 [1848], book III, chapter 17).

The basic argument states that trade between two nations is beneficial (to both the nations) even when one country would have an absolute advantage at producing all goods in relation the other country, as long as each country can produce certain goods in a more efficient way than other goods. In such a situation, each country would obtain greater "economic gains" from specialising in the goods it can produce most efficiently (and then buy the remaining of the domestically demanded goods from abroad), in comparison with the gains each country would obtain from diversifying production and closing their borders for trade.

After the first presentation of these ideas in the writings of David Ricardo, the theory of comparative advantage had further theoretical refinements during the 20th century (e.g., Heckscher 1919; Ohlin 1933; Stolper and Samuelson 1941; Rybczynski 1955), yet the substantial assumptions of traditional trade theory, i.e., constant returns to scale and perfect competition, remained always at its core. However, trade models based on these assumptions and the notion of comparative advantage as the main determinant of trade patterns have been criticised as extremely unrealistic and empirically void.

During the late 1970s, some international economists started developing models which abandoned the old assumptions of constant returns to scale and perfect competition, and which included instead the allegedly much more realistic assumptions of the monopolistic competition models (see, e.g., Helpman 1981; Krugman 1979, 1980, 1981; Lancaster 1980). These modifications gave rise to the "new trade theory". In this new theoretical approach, comparative advantage is no longer the only or even the main determinant of trade patterns among countries; instead *increasing returns of scale* are taken as the main driver of international specialization.

Broadly put, monopolistic competition models contain two key assumptions: First, firms included in an industry are able to differentiate their product from the products of each other, and thus each firm can be taken as a "monopoly" of the particular good it produces. Second, firms in the same industry take the prices of each other as given (in contrast to traditional oligopolistic models in which the pricing decisions of firms are interdependent). The name of these models comes precisely from the fact that each firm is assumed to behave as a "monopoly" of its own differentiated product, even if goods offered by rival firms within the industry are still similar to the extent that they can be substitutes of one another. Accordingly, firms can be represented as monopolists competing with each other, and the demand for each commodity is modelled as depending on the quantities and prices of all the non-identical substitutes produced by other firms in the industry.[67]

Just like the classical account, new trade theory has also been used to build arguments in favour of free trade (e.g., Krugman 1993; Feenstra 2006), but the main point that its proponents customarily stress is that new trade theory provides a much more realistic account of how trade patterns among countries

[67] For a fairly simple account of these models, see Krugman and Obstfeld 2009, chapter 6; for a more advanced and formal approach, see Feenstra 2004, chapter 5.

actually developed. Nevertheless, the new allegedly more realistic theory also brought some tension to the scientific consensus on the benefits of free trade, since dropping the assumption of perfect competition precluded the possibility of securing Pareto optimal outcomes in the models. Thus while patterns of trade can be theoretically explained and modelled overall, firms and industries are heterogeneous and consequently the welfare effects of liberalisation need not be always beneficial for all.[68]

As Krugman would later recollect, the same "new trade theory" that "legitimized imperfect competition" making arguments in favour of trade more realistic "also opened the door to possible arguments for government intervention" (Krugman 1993, 363). Indeed, as soon as the "new trade theory" was put forward and recognised, proposals about the protection of strategic sectors in the economy with the aim of achieving Pareto optimal outcomes also emerged (see, e.g., Brander and Spencer 1985).

From an exclusively theoretical perspective, new trade theory provides no ultimate answer to the debate of free trade versus protectionism; the theory is in fact theoretically compatible with both types of policy positions: trade liberalisation can have economic benefits and yet "strategic" interventions could help contribute improving social welfare by inducing a more efficient distribution of those gains. The theory exclusively shows that, in a world of imperfect competition, increasing returns to scale, and other particular economic complexities, both policy strategies—free trade and protectionism—are sub-optimal. Consequently, most of the serious academic debate on the benefits of free trade has moved from appeals to theoretical accounts towards arguments based on empirical evidence. But can the debate be settled then with empirical research?

7.3. Empirical evidence on the benefits of free trade: what are the hypotheses under evaluation?

For the most part, the empirical investigation of the benefits of free trade consists in econometric estimations of causal effects of a stylised variable that stands for "trade liberalisation" upon a set of other variables that are typically considered good economic indicators of economic gains (I will refer to them below). Even if all these studies analyse a number of different specific hypotheses, whenever they are used to argue in favour of free trade, typically their final outcomes are piled together as a supporting evidential base for a

[68] For a discussion on how trade liberalisation cannot secure Pareto optimal outcomes, see Driskill 2012.

general causal hypothesis such as: “Trade liberalisation causes economic gains” (see, e.g., Irwin 2009). Such a claim is of course too general to provide any useful understanding of the effects of trade liberalisation, or any specific guideline for policy implementation, so let us look in more detail at the specific hypotheses commonly studied.

The hypotheses appraised in the empirical literature are typically of the generic form:

‘(For *P*), TL causes Y’

Where *P* is the population of all countries or some large subset of countries, TL is a variable for “trade liberalisation”, usually defined as a dummy (more on this below), and Y can be any of the usual macroeconomic indicators of national “economic gains”, such as growth per capita, investment, income equality, prices, and the like, as measured by accepted international standards. The variables selected as the posited effects Y vary rather largely among different studies, which is unsurprising given the wide variety of notions that can fall under the term “economic gains”.

The usual aims of empirical work are either to openly test the implications of current trade theory or to explore the potential effects of free trade using historical data. In new trade theory, for instance, monopolistic competition models can be used to formally derive the following hypotheses: expanding production of certain goods to a larger scale than domestic demand will generate efficient cost reductions, will increase the variety of available goods in the domestic market, and will reduce the level of prices in the trading countries (see Feenstra 2006). Consequently, a portion of empirical studies have been primarily designed to test whether increments in foreign trade actually have these implications, and thus in principle to test the validity of new trade theory.

Other empirical studies in international economics focus on testing the direct effects of trade liberalisation on macroeconomic variables. These studies are still theory-based in very basic ways (as I will clarify below), still their primary goal is not to test the validity of underlying theories, but the causal influence of trade liberalisation (TL) on different observable economic variables (Y) in isolation from other disturbing and contextual factors.

The following is a non-exhaustive list of some of the main economic gains (Y) that have been empirically evaluated and supported by existing studies:

1. TL causes increases in *economic growth* (GDP/capita).[69]

2. TL causes increases in *investment* (INV).[70]

3. TL causes increases of the *proportion of trade* with foreign countries (usually called "openness ratio").[71]

4. TL causes reductions of the *prices of goods* (P), given increasing returns to scale.[72]

5. TL causes increases of the *variety of goods* available to consumers.[73]

6. TL causes increases of *economic competition* which causes the self-selection (and survival) of the most efficient firms.[74]

7. TL causes *convergence of wages*, which leads to reductions of income inequalities.[75]

8. TL causes reductions of *unemployment* (U).[76]

9. TL causes transferences of *foreign technologies*.[77]

7.4. Making causal generalisations explicit for policy purposes

Obviously the hypotheses above have to be true (or believed to be true according to certain scientific standards) in order for any user of the science to even begin considering their policy implications. Since my aim here is to investigate the evidential requirements for achieving policy effectiveness (in addition to having access to well-established scientific causal generalisations), I will assume that the causal hypotheses studied and accepted in the empirical literature have been well-established and thus they can be trusted as true claims.

---

[69] See Edwards 1992; Dollar 1992; Sachs and Warner 1995; Greenaway, et al. 1997; Ades, et al. 1999; Wacziarg and Welch 2008.
[70] Mainly via foreign direct investment (FDI); see Baldwin and Seghezza 1996; Wacziarg and Welch 2008.
[71] See Wacziarg and Welch 2008.
[72] See Harris 1984; Smith and Venables 1988; Badinger 2007.
[73] See Feenstra 1994; Hummels and Klenow 2005; Broda and Weinstein 2006.
[74] See Coe, et al. 1997; Bernard, et al. 2003; Eaton, et al. 2004; Trefler 2004; Feenstra and Kee 2008.
[75] See Ben-David 1993; Frankel and Romer 1999; Winters, et al. 2004.
[76] See Krueger 1983; Milner and Wright 1998; Falvey 1999.
[77] Mainly foreign capital equipment and foreign research; see Keller 2004; Madsen 2007.

If a causal claim of the form 'TL causes Y' is taken as true, then in order to be able to make any potential policy recommendation one needs information about the following two aspects:

a) The exact meaning of implementing a particular change in TL (a TL reform, or any other concrete policy intervention in TL). Users of the science and policy makers will need to know this to a certain level of precision in order to be able to properly intervene on TL in particular countries in an effective way.

b) The meaning of the specific causal notion that has been tested and established as true in the relevant empirical research, since—as shown in chapters 3 and 5—different notions of causation allow or constrain the inferences about potential practical applications and concrete policy implementations of the scientific results.

7.5. The meaning of changes in 'trade liberalisation'

The first thing that comes to mind when one thinks of trade liberalisation is the elimination of taxes to international exchange. That is somewhat correct. Trade liberalisation reforms almost always include the reduction or elimination of commercial tariffs and quotas. That is also the way in which typical theoretical analyses and simulations conceptually study variations in trade liberalisation: one or a number of variables represent international tariffs so that their burden on the whole economy or on different economic sectors can be analysed. But when it comes to the real world, what exactly is the meaning of 'trade liberalisation'? What kind of change in a real economy corresponds to a change in the 'trade liberalisation' variable?

In order to do empirical research from available datasets, international economists have defined a variable usually labelled "openness", which essentially measures the liberalisation level (in terms of the amount of restrictions to trade) in a particular country for a certain year. Since many countries have explicitly started up liberalisation reforms at some point during the last decades, researchers have mainly focused on characterising two essential aspects of liberalisation, namely: some definite *criteria* for what counts as an "open" (or a "closed") economy, and precise *liberalisation dates* for all countries that have launched a liberalisation reform.

The *openness criteria*, on the one hand, allow for straightforward comparisons, typically cross-country comparisons, of several economic indicators (Y) between “close” and “open” economies, for fixed periods. On the other hand, reliable data on *liberalisation dates* allows for comparisons of the economic trends and development within each particular country “before” and “after” liberalisation reforms were implemented.

The most famous and still widely used *openness criteria* was developed by Jeffrey Sachs and Andrew Warner (1995). The authors constructed a dummy variable (using an ‘international comparisons’ dataset by Summers and Heston 1991) and classified countries as either “close” or “open” based on five different dimensions. Accordingly, a country is taken as “closed” if it displays *at least one* of the following five economic features (i.e., the dummy variable takes the value of 0 for a country in a particular year if at least one of the five features below is true for that country in that particular year):

1. Average tariff rate level of 40% or more.
2. Nontariff barriers covering 40% or more of all trade.
3. Black-market exchange rate at least 20% lower than the official exchange rate.
4. State monopoly on exports.
5. Socialist (centrally-planned) economic system.

Collecting precise *liberalisation dates* is less straightforward, since usually different procedures are employed. A relatively simple method is to use an *ex-ante* approach and consider, for instance, the statements of intent made by countries when a World Bank Structural Adjustment Loan (SAL) is granted. The date in which the loan begins is then taken as the starting point of a liberalisation reform (see, e.g., Harrigan and Mosley 1991; World Bank 1993; Greenaway, et al. 2002).

The alternative is to use an *ex-post* approach, in which a set of economic characteristics are evaluated in a number of countries along a certain period in order to determine significant changes in their openness conditions. For example, following this approach, Dean, Deasi, and Riedel (1994) inferred liberalisation dates for 32 countries from the 1980s to the beginning of the 1990s by analysing in detail their socioeconomic history and focusing on four variables: changes in average tariffs, changes in quotas, export taxes, and foreign exchange restrictions (Dean, et al. 1994, 11-14). Similarly, in the study by Sachs and Warner (1995) mentioned above, the authors inferred liberalisation dates for 111 countries using their own slightly different openness criteria.

An additional variable related to liberalisation that is commonly included in empirical studies is the *openness ratio*. This is a measure of the volume of foreign trade of a country relative to its national product, thus it is simply calculated as a ratio of the imports plus exports to the GDP (in a particular year). Note that the "openness variable", as defined for example by the Sachs-Warner criteria, essentially measures the amount of trade barriers a country has in place, whereas the "openness ratio" refers to the amount of foreign trade a country experiences.

Changes in liberalisation levels as reflected in the "openness variable" need not coincide with changes in the "openness ratio", for example, a reduction in the level of trade tariffs (a change in the openness variable) may or may not result in an increase of the actual volume of foreign trade (openness ratio). The openness ratio could be used as a measure of the actual causal efficacy of a trade liberalisation reform on traded volumes. While the volume of foreign trade (openness ratio) may change for other reasons different from changes in the level of foreign-trade barriers.

There is no definite agreement on the best way to measure "trade liberalisation", so a plurality of accounts remains (see Rodríguez and Rodrik 2001; Greenaway, et al. 2002). I will simply emphasise here a particularly problematic issue from a policy-making perspective in relation to the measurement of variable TL. It could be the case that the posited cause "trade liberalisation" is measured in some clear and explicit way, and yet be entirely uninformative about what exactly a policy maker is supposed to change or intervene upon in order to affect TL in a desired way. For instance, assume the Sachs-Warner criteria is accepted, then how do one intervene on TL? Should tariffs be reduced, should non-tariff barriers, should the monopoly of the state on exports be dissolved? Which of the five dimensions in the measurement has to be affected and in which way in order to attain a desired change in Y? Should it be a combination of changes in all five dimensions?

7.6. Empirical methods and implicit connotations of causation

The two most common empirical approaches to estimate the effects of trade liberalisation are cross-sectional and time-series analyses.

Cross-sectional analysis (also called cross-country analysis when units are countries) has been widely used in trade and economic growth research, especially during the 1980s and 1990s. In this type of studies, a set of explanatory variables (containing TL) and a dependent variable (Y) are included in the specification of a regression equation and the influence of each variable on the selected Y is estimated in a particular year for all sampled countries. Using matching techniques, it is possible to compare a group of countries that have experienced a significant trade liberalisation reform (TL = 1) with a group of countries that have not (TL = 0). Assuming the model specification is correct (i.e., assuming that the regression equation includes variables for all relevant causal factors that have an effect of Y in addition to TL), then the result is an estimate of an *average causal effect* of TL on Y, inferred from cross-sectional comparisons between countries that were liberalised or non-liberalised during the same period.

Time series analysis offers, in contrast, estimates of the causal effect of trade liberalisation (TL) within one and the same country throughout any number of consecutive years. Using data on "liberalisation dates", it is possible to evaluate trends and significant "jumps" in the evolution of the economic indicators as a consequence of trade-liberalisation reforms within a single country. In such cases, TL is set to 0 for all years before the liberalisation date of a particular country, and 1 afterwards. The result is an estimate of an *average causal effect* of TL on Y, inferred from differential comparisons of the same country's variations in Y from one year to the next one throughout a set of consecutive years.

Empirical studies can also combine both cross-country and intertemporal analyses whenever panel datasets are available. A panel dataset encompasses information on a relatively large number of socioeconomic characteristics of different countries for different years. Applying the openness criteria to distinguish *liberalised* and *non-liberalised* countries, plus information from a panel dataset, it is possible to estimate in the same study separate regression equations for different groups of countries. For instance, using a sub-sample including only countries that have experienced trade reforms allows before-and-after comparisons of the effects of TL; while using a sample including liberalised and non-liberalised countries allows for comparisons between the different trends followed by open and closed economies (see, e.g., Greenaway, et al. 1997).

Different studies differ from one another in rather small technical details. However, it is fair to say that the empirical research on trade liberalisation provides results mainly of the two kinds described above: 1) estimates of causal effects obtained from comparisons across different countries at the same time, and 2) estimates of causal effects from within-country comparisons at different times.

The immediate consequence of using different methodological approaches is, not surprisingly, that *the estimated causal effects are always different.* The specific assumptions required by each approach (for instance, assumptions made about country homogeneity, or about potential confounders, or about uniformity background characteristics, and so on) would normally determine to a certain extent the meaning of the resulting causal notion under evaluation (see chapter 5). But if the testing methodology of choice affects the meaning of the causal concepts under evaluation, do these different methods test the validity of the same causal hypothesis or rather different ones altogether? Let me try to answer this question by analysing the methods employed and their assumptions in more detail.

7.6.1. Controlling for confounders to test causal generalisations

To get *valid* evidence for a causal hypothesis like 'TL causes Y' some inferential method that allows control for all (or as many as possible) potential confounders that could have an effect on Y is required. This is a traditional approach of scientific inquiry, i.e., to investigate about the workings of causal relations in isolation (see Mill 1843, 3.8; Mäki 1992a; Reiss 2008).

In cross-country empirical studies on the effects of trade liberalisation, one of the main ways to control for potential confounders consists in including all the relevant variables in the regression equation. The standard procedure to decide which variables are included is to look at what accepted growth theory has to say. Typically, the starting point is a 'core' new growth theory model "of the type which has now become standard" (Greenaway et al. 2002, 234).[78]

For example, in the already mentioned highly-quoted article by Sachs and Warner (1995),[79] the authors used Barro's (1991) growth specification as a baseline for their first regression, and after a few variations they end up with the following variables:

Dependent variable:
Y: Real GDP per capita

[78] In the empirical international trade literature, a "standard" specification refers to one in line with the models put forward by Romer (1990) and Barro (1991), and the specification-search results of Levine and Renelt (1992). See Hoover and Perez 2004 for a detailed discussion of different specification-search methodologies to find robust determinants of economic growth for cross-country regressions.
[79] 5474 citations in Google Scholar; 223 in Web of Science (by February 2016).

Explanatory variables:

X_1: Sachs-Warner openness dummy variable (TL)

X_2: Ratio of real gross domestic investment to real GDP (INV)

X_3: Population density

X_4: Secondary-school enrolment rate

X_5: Primary-school enrolment rate

X_6: Ratio of government consumption to GDP

X_7: Extreme political repression and unrest

X_8: Number of revolutions per year

X_9: Number of assassinations per capita per year

Controlling for this list of variables, Sachs and Warner's cross-country analysis provided an estimate of a positive causal effect from TL to GDP per capita, using data on 111 countries for a period from 1970 to 1990. The estimate of a magnitude of 2.44 means that *on average* countries classified as open have experienced an economic growth of 2.44 percentage points higher than countries classified as closed.[80]

In 2001, Rodríguez and Rodrik published a very influential critical review—already mentioned in the first section—on the state of the empirical research on the economic effects of trade liberalisation.[81] They reviewed in detail some of the most important articles at the time (Dollar 1992; Sachs and Warner 1995; Ben-David 1993; Edwards 1998; Frankel and Romer 1999; and more briefly: Lee 1993; Harrison 1996; Wacziarg 2001) and discussed thoroughly a number of methodological issues in the literature. According to the authors, the nature of the relationship between TL and growth had not been properly assessed and remained "far from being settled on empirical grounds" (Rodríguez and Rodrik 2001, 266).

The two main criticisms were directed to the Sachs-Warner openness criteria, and to the use of cross-country analysis. On the first issue, they conclude that the Sachs-Warner openness indicator "yields an upward-biased estimate of the effects of trade restrictions" (Rodríguez and Rodrik 2001, 282), and that ultimately it is "so correlated with plausible groupings of alternative

[80] The same set of control variables has been used with a few variations in subsequent — empirical research, see, e.g., Dollar 1992; Ben-David 1993; and Edwards 1998. Greenaway et al. 2002 used the same model specification in order to test the robustness of TL on GDP per capita to different ways to obtain liberalisation dates; and Wacziarg and Horn Welch 2008 replicate Sachs and Werner's result using the exact same econometric specification and a new dataset.

[81] 3766 citations in Google Scholar; 120 in Web of Science (by February 2016).

explanatory variables [...] that it is risky to draw strong inferences about the effect of openness on growth based on its coefficient in a growth regression" (Rodríguez and Rodrik 2001, 292). On the reliance on cross-country regressions, they argue that the static estimates of cross-sectional analysis could mask dynamic variation of causally relevant characteristics of countries, which in turn could make the estimate non time-invariant. And overall they claim that:

> For the most part, the strong results in this literature arise either from obvious misspecification or from the use of measures of openness that are proxies for other policy or institutional variables that have an independent detrimental effect on growth. When we do point to the fragility of the coefficients, it is to make the point that the coefficients on the openness indicators are particularly sensitive to controls for these other policy and institutional variables (Rodríguez and Rodrik 2001, 315).

As a consequence of Rodríguez and Rodrik's criticisms, subsequent studies reduced their reliance on cross-country estimation in favour of other within-country estimation techniques that could be used to analyse the effects and trends over time of trade reforms, without relying so much on choosing the adequate controls for comparing heterogonous countries. Panel data analyses in which cross-country estimations were complemented and combined with within-country estimation techniques started to appear. In particular adaptations of methods following the design-based logic of randomisation (and of the potential outcomes framework), such as *difference-in-differences* and *fixed effects* analyses have since then become more popular.

One of the most influential empirical studies, after Rodríguez and Rodrik's (2001) critique, has been Wacziarg and Horn Welch's (2008) "Trade liberalization and growth: new evidence".[82] The authors do three things in this article: first, they update the Sachs-Warner openness classification using a more comprehensive database; second, they use a more complete database to replicate (and revise) Sachs and Warner's cross-sectional positive results; and the "third and most important goal is to exploit the timing of liberalization in a within-country setting to identify the changes in growth, investment rates, and openness [ratio] associated with discrete change in trade policy" (Wacziarg and Horn Welch 2008, 189).

Using an extended dataset (from the 1970s until the end of the 1990s) for 141 countries with corrected liberalisation dates, Wacziarg and Horn Welch tested the robustness of cross-sectional analysis estimates for different

[82] 1296 citations in Google Scholar; 175 in Web of Science (by February 2016).

decades. They found that the Sachs-Warner results did not hold for the 1990s because relevant country characteristics indeed vary through time, confirming thus some of Rodríguez and Rodrik's concerns about the potential fragility of cross-country estimates.

The authors then used the same dataset to estimate the within-country effects of trade liberalisation (TL) on GDP per capita, investment, and volumes of trade relative to GDP (openness ratio), and found that, in contrast to the cross-country results, "the results based on within country variation suggest that over time the effects of increased policy openness within countries are positive, economically large, and statistically significant" (Wacziarg and Horn Welch 2008, 189).

The specification of the regressions in the analysis of within-country effects is entirely different from those use in the cross-country analysis in many technical respects. But the main difference to be noticed here is that, in contrast to adding explanatory variables to control for confounders, in the within-country methodology there are no added variables. Instead, the estimates derive entirely from the average differences in the level of Y from one period to the next using variations in the dummy variable TL to capture the difference between pre- and post-liberalisation periods.

No confounders are explicitly controlled for since essentially what these methods attempt is to measure an average causal effect of all the changes from one period to the next, i.e., the differences in Y from time t_1 to time t_2, and then from t_2 to t_3, and so on, for all years in the dataset for each country, then a final general estimate is calculated as an average over all country-specific causal effects. For instance, to estimate the within-country effects of TL on GDP, Wacziarg and Horn Welch used the following equation:

$$\log y_{it} - \log y_{it-1} = \alpha_i + \beta LIB_{it} + \varepsilon_{it}$$

where y_{it} is GDP per capita in country i at time t and LIB_{it} (their variable for TL) takes the value of 1 if t is greater than the liberalisation year.[83] Using this equation, then their estimate for the period from 1950 to 1998 was of 1.42 percent points of average difference in growth between liberalised and non-liberalised countries.

Wacziarg and Horn Welch's results are presented as revised results on the positive economic effects of TL, after taking into account the methodological

[83] Residuals are modelled so as to include country and time fixed-effects, see Wacziarg and Horn Welch 2008, 199-202.

concerns pointed out by Rodríguez and Rodrik (2001). Their within-country analysis purportedly complements cross-country findings by offering an account of how economic variables like growth and investment have evolved from year to year in countries that experienced liberalisation reforms and in countries that remained closed.

While this move apparently deals with within-country variation concern, difference-in-differences analysis assumes that by averaging over the differences in Y of all different countries from which some liberalised at some point (i.e., switched from TL = 0 to TL = 1) and others did not, this can be treated as the outcome of a random assignment to TL = 0 or 1 and consequently an *average causal effect* can be estimated. This amounts to following the logic of the potential outcomes framework (see chapter 5, section 5.5), plus an additional assumption often called "parallel trend assumption" which states that the average change observed in the "control" group (of non-liberalised countries) represents the *counterfactual trend* for the "treatment" group (of liberalised countries) if there were no "treatment". So ultimately, the causal effect is a difference between the observed value of Y in liberalised countries and what the value of Y *would have been* with parallel trends had there been no liberalisation.

Similarly, fixed effects techniques explicitly assume that a number of causal relevant causal characteristics are time-invariant during different periods for the same country thereby making the before-and-after comparisons meaningful. Using fixed-effects amounts then to secure a similar situation to the one expected when a "temporal stability" assumption is made in the potential-outcomes framework (see chapter 5, section 5.5).

7.6.2. Different causal notions and different practical relevance

All methods—cross-country or within-country—have their respective merits and deficiencies depending on the context of the case at hand. Cross-country methods have the advantage of allowing the researchers to control explicitly for potential disturbing factors (which is useful insofar as reliable background knowledge about those factors is already available), but if countries have a high variability over time the estimates will be biased. Within-country methods have the advantage of exploring trends that different countries experience before and after trade reforms have taken place, and trends followed by countries that have remained closed, but causally relevant heterogeneity across countries could always be hidden in the final average results.

Let us take as a given that the empirical methods used to test the causal effects of free trade are indeed good means to control for potential errors or cofounders whenever properly implemented. Therefore, one can take the existing empirical evidential base (as many economists do) as *valid* evidence for the causal generalisation "trade liberalisation causes economic growth". Is the evidence also enough to act upon the causal hypothesis on the benefits of trade liberalisation?

As discussed in chapter 3, causal claims can have different practical implications depending on their specific meaning. Economists have never been explicit about the causal theory they endorse whenever they evaluate causal hypothesis about macroeconomic variables. Still by looking at the empirical studies it seems fairly clear that the causal notion under evaluation could be analysed in terms of a probabilistic theory of causality (see also chapter 5).

Mechanistic causal information is also sometimes employed, but mainly in the form of background knowledge in order to facilitate the estimation of probabilistic results. For example, brief explanations about the underlying mechanism are sometimes quoted to justify the chosen model specification of the regression equations. Some other times—as I will describe briefly in the next section—mechanistic background information is called upon in order to explain away exceptions and outliers of the general results. Thus, mechanistic causal information plays at most an indirect role in supporting posited probabilistic causal claims (see Reiss 2015; Lehtinen 2016).

Within a probabilistic theory of causality, there is still room for a number of distinct causal concepts related to different definitions and criteria about what is to count as a causal relation.[84] As can be noticed from the methods described in the previous section, the criteria used in the empirical research on the benefits of free trade resembles Dupré's (1984) notion of a "fair sample" approach to probabilistic causality (see chapter 5, section 5.2). Moreover, some of the techniques employed are explicit attempts to analyse the existing panel datasets as if they were the result of randomised experimental designs, and thus the causal concepts estimated in these studies are types of *average causal effects* (ACEs).

It should be clear, however, that the specific causal notions referred to by the resulting estimated average causal effects are heavily determined by the method (cross-sectional or within-country analysis) employed, and by the assumptions required to achieve those estimates. The meaning of the claim "TL

[84] See the discussion on conceptual causal pluralism in chapter 3, section 3.2, and on distinct causal criteria in the probabilistic account in chapter 5, sections 5.2 and 5.4.

causes Y" can vary significantly depending on the method employed and the assumptions made to establish it, which is another way to say that its potential policy implications may vary significantly as well (see chapter 3).

As I have suggested in the previous chapters, different causal concepts need not have the same practical relevance. In this case, the methods and assumptions made so as to estimate the causal effects of trade liberalisation determine (enable and constrain) the potential inferences about practical applications that can be made from the results. Even if the available empirical evidence is *valid* enough to allow us to believe in the truth of the posited scientific causal hypothesis "Trade liberalisation causes economic gains", this still does not answer the question: Is one allowed to act upon it? Shall we implement a trade policy reform in our country in order to grow faster? I will elaborate in the next section on the problematic implications that these empirical results can have for policy makers trying to infer effective policy recommendations.

7.7. Study-specific hypotheses versus causal generalisations

There are three aspects of the empirical studies supporting the economic gains of trade liberalisation that I want to point out.

The first is that there is no single variable that accurately measures 'trade liberalisation'. This has been the focus of most methodological discussions among economists working on empirical international economics. From a scientific perspective, the aim has been to conceptualise the best compound variable that could account for what economists consider a close economy (e.g., Sachs and Warner 1995). In many cases, it is a search for the best proxy variable for TL (see World Bank 2008). This has resulted in several empirical evaluations of the robustness of the economic effects to different measurements of TL (see Greenaway, et al. 2002). As long as the proposed measures are shown to be reasonable indicators (or signs) of trade liberalisation reforms, they are taken as acceptable proxies to test generalisations.

From a policy-oriented perspective, however, the main problem is that there is no clear way of intervening on a multidimensional variable such as the one currently in use. What exactly should a policy maker do in order to induce a change on the Sachs-Warner notion of TL? Under that understanding of TL, any policy intervention should include a combination of changes on at least "the level of tariffs", "non-tariff barriers", and "the black market exchange rate". Nevertheless, the existing empirical evidential base is completely silent about which precise combination of changes in the different dimensions of TL should be implemented to induce a desired effect in the level of economic growth or any other relevant outcome variable. Thus, even if the variables currently in use to stand for TL levels were acceptable to test the validity of causal generalisations, they would still be rather ambiguous in relation to the purposes of policy makers.

Secondly, any coefficient obtained using cross-country and within-country methods, regardless of how statistically significant and unbiased, ultimately is an average-causal-effect type of causal concept inferred from a large and wide number of comparisons of different sorts at the unit-country level. Now, from a policy-making perspective, these results (even if they could very well represent the best valid evidence for believing in a causal generalisation like "TL causes Y on average in a certain positive magnitude") tell nothing about the particular policy strategy that should be implemented in a particular economy to effectively achieve the desired change in Y. Well-established empirical estimations of average causal effects could be taken as *valid* evidence for *scientific causal generalisations*, but they only amount at best to be *prima facie* evidence for *specific policy recommendations* (see chapter 6, sections 6.3, 6.4, 6.5, and 6.6).

As I have suggested before, these econometric techniques are not perfect tools to control for the influence of other relevant causal factors, and are vulnerable to ignore underlying heterogeneity of background conditions or of single-unit causal effects. Most of the intrinsic methodological problems with such estimation techniques are rather well-known by econometricians and empirical economists. The merit of many contributions often consists of some technical improvement to deal with such methodological problems, e.g., sensitivity tests, multiple regressions to test for robustness of different approaches and specifications, or case-study approaches to deal with (or explain away) single-unit heterogeneity.

Similarly, the authors of the empirical literature seem to be aware of the fact that the specific conceptual and methodological assumptions in the econometric techniques restrict in different ways the inferential power of their results. Rodríguez and Rodrik's (2001) critical article is a straightforward example of this awareness, but also other studies less methodologically focused and that are explicitly meant to support the positive economic effects of trade liberalisation are usually careful to add here and there qualifications about the interpretation of their results outside the abstract scientific realm. For instance, the existence of concurrent policies is commonly recognised as one of the potential constraints for the effectiveness of concrete trade policies in the real world (Wacziarg and Horn Welch 2008, 206-207; Greenaway et al. 1997, 1886; and Greenaway et al. 2002, 233).

In relation to the potential heterogeneity of single-unit effects that can always be hidden in the general average estimates, for example, Wacziarg and Horn Welch (2008) write after presenting their positive results:

> [T]he extent to which per capita income growth changed after trade reforms varied widely across countries. While the average effect obtained in the large sample is positive, roughly half of the countries experienced zero or even negative changes in growth following liberalization. [...] generalisations about the factors that may explain these differences are difficult to draw. The institutional environment of countries, the extent of political turmoil, the scope and depth of economic reforms, and the characteristics of concurrent macroeconomic policies all seem to have a role to play, to varying degrees in different countries. While this article paints a picture that is highly favorable to outward-oriented policy reforms on average, it cautions against one-size-fits-all policies that disregard local circumstances (Wacziarg and Horn Welch 2008, 189-190).

Unfortunately, none of these type of qualms seems to make it to the public debates or the policy-making arena. Even in semi-academic divulgation texts, such as the popular *Free trade under fire*, Douglas Irwin (2009) summarises the recent empirical results (including Wacziarg and Horn Welch's results) in favour of trade liberalisation, with no mention of all the qualms and potential biases of average results. Instead, he flatly concludes that "despite shortcomings in method and measurement, cross-country and within-country studies support the conclusion that economies with more open trade policies tend to perform better than those with more restrictive trade policies" (Irwin 2009, 54).

Even if econometricians and many economists know and understand fairly well the inherent "shortcomings" of their methods, ignoring these shortcomings when communicating economic results to the users of the science could turn into potentially dangerous misuses of these results. This can happen whenever the results are not properly communicated to (or properly understood by) policy makers together with all the specifications and qualifications inherent to the particular estimation techniques employed in the empirical evaluations.

The third aspect to be noted is what can be called the "saving-the-generalisation attitude". In the empirical research on the benefits of TL, the more general a result is (for instance, by using the largest available dataset), the better. Then when there are any potential exceptions to an established causal generalisation, the scientific priority is to save the truth of the generalisation by explaining away the outlaying cases.

This attitude is compatible, to some extent, with the methodological distinction (I have argued for in chapter 1) between doing economic *science* and doing economic *policy*, and the fact that these two endeavours have distinct aims. Broadly put: the "science" aims at studying causal relations in isolation (in the abstract) and at establishing causal generalisations, whereas the "art" of policy making aims at finding the most effective means to achieve specific effects in specific situations, not in general or on average, but in concrete unit-level cases (see chapters 5 and 6).

Empirical economists are in principle simply doing their jobs when they focus on establishing average results that are as general as possible. Nevertheless, it is entirely unjustified to suggest that such causal generalisations can be straightforwardly useful for guiding or supporting effective policy recommendations, as it is suggested in some of the quotations presented in the first section of this chapter.

As mentioned before, systematic reports put together by international economic organisations usually follow the same "generality attitude" when they put together compendiums of empirical studies as the evidential base that supports the general hypothesis that "trade liberalisation causes economic gains". However, each empirical study refers to a certain specific hypothesis of the form "trade liberalisation causes Y" (were Y stands for any of the specific economic indicators listed in section 7.3). Furthermore, every single study could refer to a specific causal connotation which is determined to a great extent by the estimation technique and its assumptions, the specification of the model, the specification of the relevant variables. Also the particular data sets employed in the study certainly determine the population of application.

Again, there is in principle no problem with considering the particular empirical studies with different specific hypotheses about, say, the effects of trade on economic growth, as "valid evidence" for the causal generalisation "trade liberalisation causes economic". However, without the proper information and understanding about the precise way in which the available empirical evidence is obtained for each of the study-specific hypotheses, policy makers would be in danger of inferring ineffective, or plainly wrong, policy prescriptions to attain particular policy goals. The existing empirical literature conforms a valid evidential base for scientific causal generalisations, but only prima facie evidence for policy action. So how can one get valid evidence for policy hypotheses?

7.8. On the methodological contrast between economic science and economic policy making

Evidence plays different functions in the process of testing scientific causal generalisations and in the process of evaluating effective policy recommendations, because these two activities have different epistemic aims which require different appraisal procedures. Consider the following two hypotheses:

> h_1: For *P*, trade liberalisation causes economic growth.
>
> h_2: For country *u*, implementing a trade liberalisation policy will generate economic growth.

The empirical evidential base of hypothesis h_1 is at best prima facie evidence for policy hypothesis h_2 about country *u*. To get valid evidence for h_2, the method of isolation is no longer useful. The evidential base that could count as valid evidence for h_2 requires putting together literally everything we know about the relevant disturbing factors that can potentially affect a specific policy intervention intended in country *u*. That is, all sorts of specific contextual conditions of country *u*: institutional, geographical, political, the socioeconomic situation of its inhabitants, the composition of its economic sectors, cultural features, relevant historical events, a detailed account of concurrent planned policies, other countries' trade policies, and so on and so forth.

A good place to start would be to make explicit the assumptions used to justify the empirical base supporting h_1. The assumptions required to make work the evidential methods used to test the generalisations can be used as a starting point for a list of features that have to be further tested in order to secure the effectiveness of specific interventions.

Consider, for instance, some of the basic assumptions commonly made in econometrics based on the potential-outcomes framework (as discussed in chapter 5, section 5.5):

> **Temporal stability**: The value of a single-unit effect is assumed to be independent of when the cause is produced and measured in any particular unit of *P* (see Holland 1986, 948).

The framework assumes that the response on the same unit to the same cause would be the same response over time. Simply put, it does not matter when any particular unit is exposed to the posited cause (or to its alternative), the response will be the same for that same unit. In relation to trade liberalisation, this assumption has to be questioned. One of the problems with existing empirical evidence has been precisely that the effects of TL on economic growth seem to vary depending on the decade in which the TL reform was implemented (see Wacziarg and Horn Welch 2008, 202-206). Empirical economists have developed new techniques to analyse these time variations when they calculate average effects, however further understanding on the specific causes of this time variation will be useful for country-specific policy purposes.

> **Stable unit treatment value assumption (SUTVA)**: the value of the causal effect of each unit u_i is assumed to be independent of any changes in the causal exposures to TL of any of the other units in *P* (see Morgan and Winship 2007, 37-40).

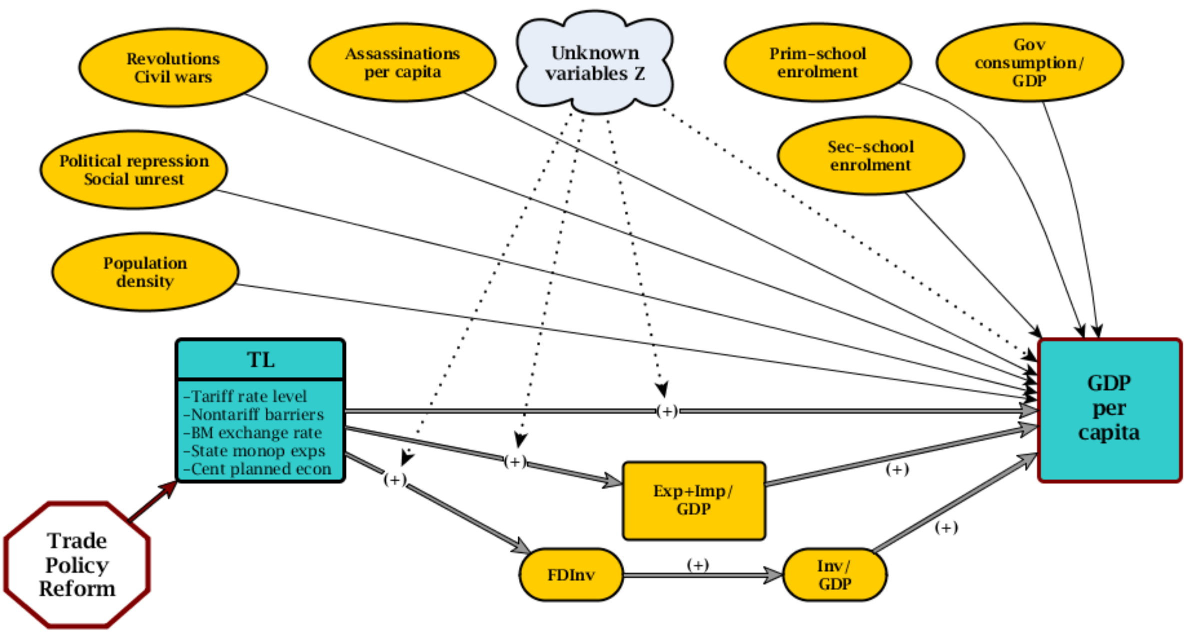
If this assumption does not hold, then the estimated causal effect could change depending on how many members of *P* are exposed to the causal variable. Again, this is a useful assumption when estimating average causal effects, but it seems particularly problematic specially in relation to trade liberalisation. There are some studies, in fact, showing that the effects of trade liberalisation for some Asian countries in the 1960s were positive and high, in part because the majority of countries in the region at that moment had closed economies. Whereas, the effects of trade liberalisation for Latin-American countries in the 1980s and 1990s had ambiguous effects, especially for the countries that started liberalisation policies after most economies were already open. Trade liberalisation is simply a variable for which the SUTVA does not apply, since the specific causal effect will always depend on what the other trading countries (at least in the economic region of country *u*) will do.

Whether the two assumptions just mentioned (or any other similar assumption made for the purposes of empirical testing of causal generalisations) are true or not about real countries determines how much economists can trust the truth of the causal relation in general. In as much as the assumptions are not true about a particular country (even if the *average causal effect* is positive for *P*), then it is necessary to perform additional explorations of the conditions which could preclude (or enable) the realisation of the intended effect in that particular country.

A reasonable place to begin the local empirical investigation for potential disturbing factors is the list of (observable) variables that were controlled for in the empirical investigations using regression analysis. Assuming there are good reasons to believe the specifications used were correct, then policy makers can take seriously some of the parameters (or at least the signs) estimated in such equations and run further more localised empirical investigations. Local and regional studies are required to explore the actual mechanisms that can constrain (or enable) the intended actual effect in the particular country in which a liberalisation policy is to be implemented.

Consider the list of variables (see section 7.6.1) that were controlled for in the empirical studies by Sachs and Warner (1995) and by Wacziarg and Horn Welch (2008). Following those accounts, one can draw a causal diagram like **D1**.[85]

Diagram D1



All the variables in this diagram in addition to TL have an effect on GDP per capita. They were explicitly controlled for in the empirical tests. An evaluation of the potential effectiveness of a trade liberalisation policy in a particular country will require an evaluation of all these causal relations in the local context of the intended policy. Moreover, it will be useful to search for any additional information about the cloud-shaped set labelled "unknown variables Z" (which include such things as institutional framework, cultural aspects, or potential influences from concurrent policy reforms), perhaps coming from

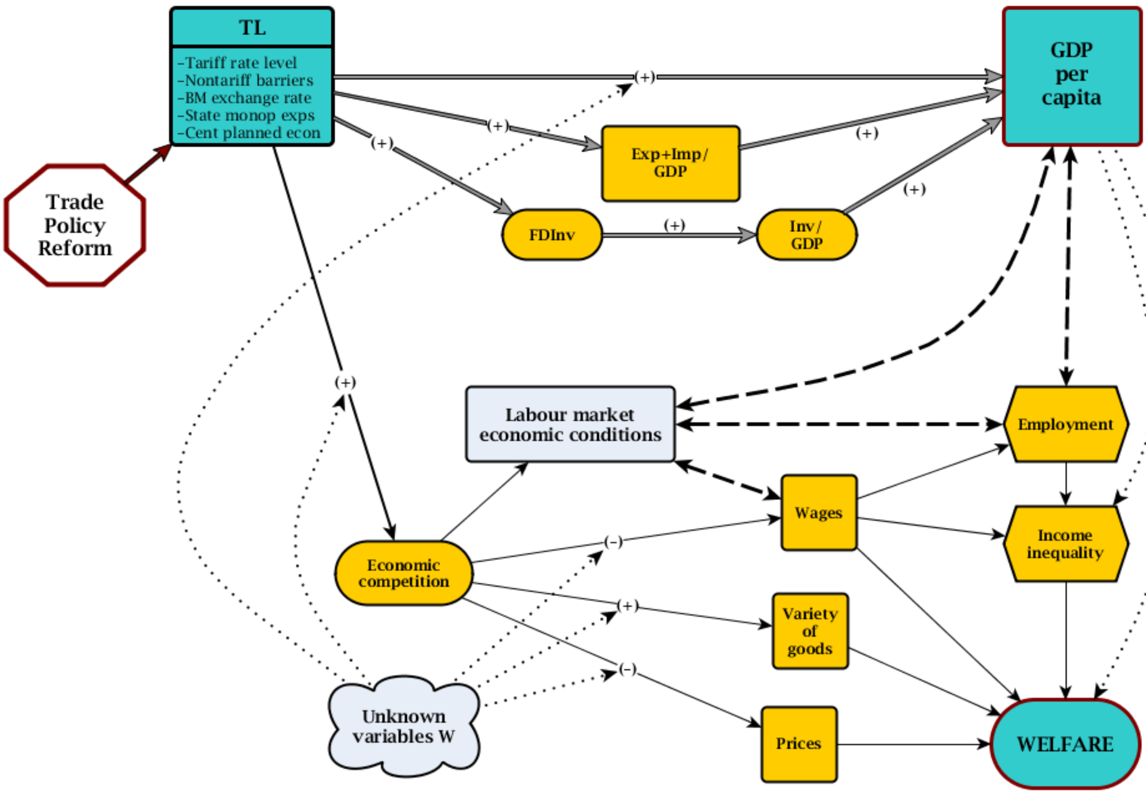
[85] Causal diagram **D1** shows the main variables controlled for in cross-country regressions in typical empirical studies on the benefits of trade liberalisation. The sources are the works cited in the previous sections of this chapter.

experts in other disciplines, e.g., sociology, psychology, political sciences, environmental sciences, history, and so on.

The point is that the methods employed to estimate the *average causal effects* are intended to isolate the causal relation from all other potential influences of disturbing factors. But it is precisely information about all those potential disturbing factors that is now required to evaluate the effectiveness of a particular policy hypothesis in a particular country.

There is also empirical evidence supporting (to different levels of success) the general suggested mechanisms of new trade theory (see Feenstra 2006). On the basis of that research, a similar picture like that of **D1** can be drawn and used as another guideline for *local* empirical investigations in relation to the implementation of concrete trade liberalisation policies.

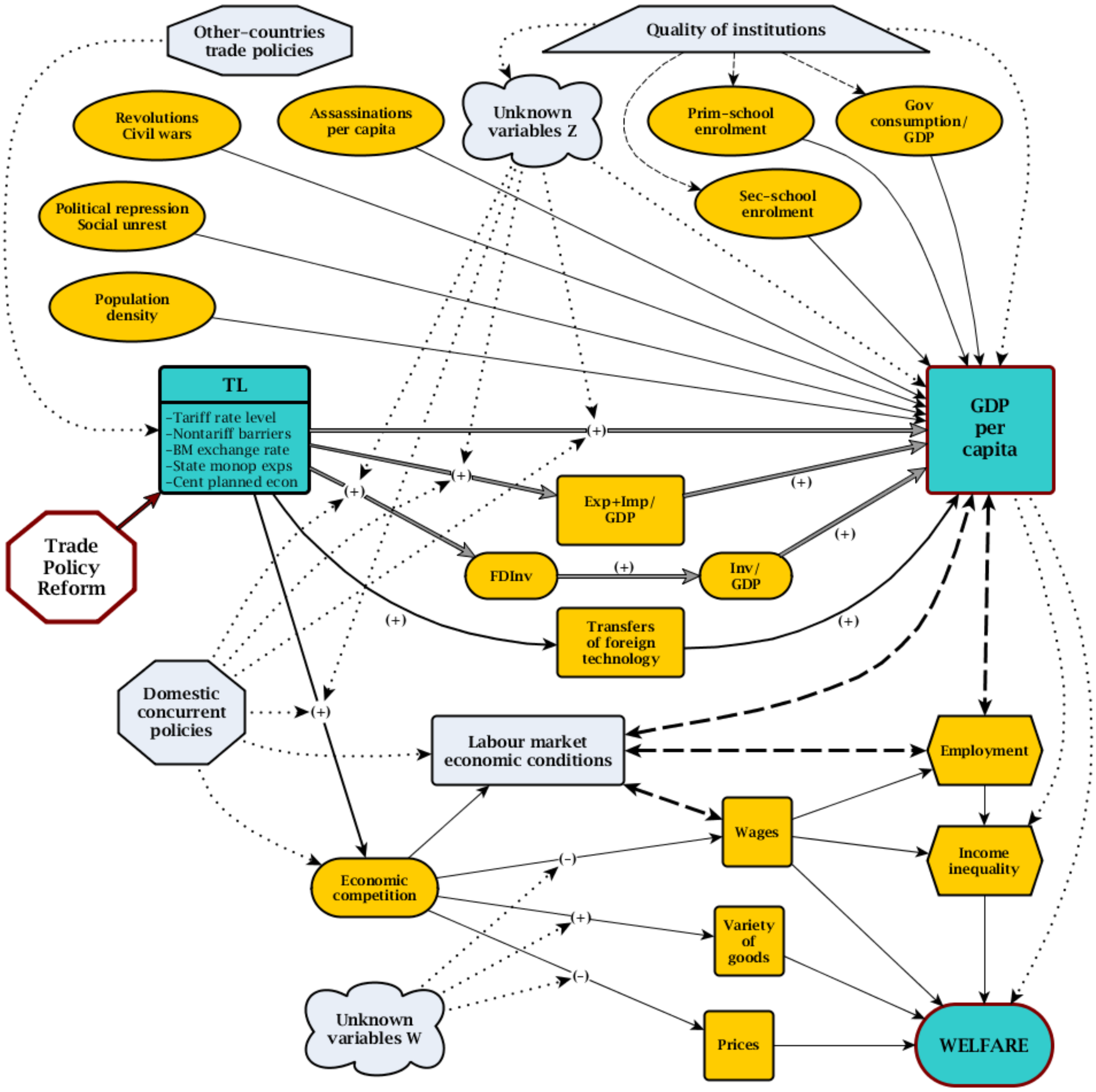
Diagram D2



On the basis of the diagram above, a policy maker should have to evaluate the status all the relevant nodes (and the values of variables W) and to evaluate any existing evidence on the specific causal effects that such variables have (or have had) on the causal relation under consideration, i.e., TL causes GDP per capita. As it has been discussed by some authors (see, e.g., Driskill 2012) increases in economic growth are never automatic improvements on welfare.

This is reflected in the right-bottom corner of diagram **D2**. New international trade economics provides some insights on the possible mechanisms that could lead to welfare distortions, but the point here is that depending on the concrete policy aims, all the relevant nodes and causal connections in diagrams **D1** and **D2** should be evaluated at a local level.

Diagram D3



Local and regional studies have become more popular after the year 2000. These studies show (ex post) some of the contextual variables that have had a distorting effect in particular regions after trade liberalisation reforms have been implemented (see, e.g., Easterly, et al. 2003; Baylis, et al. 2012). Among the disturbing factors typically mentioned are: not sufficient investment in education; not sufficient innovation, and infrastructure; poor quality of national institutions; and (the previously mentioned) the state of specific complementary policy reforms (fiscal, energetic, communications, and the like). Without trying to be exhaustive, **D3** above is my attempt of compiling in a causal diagram most of the relevant causal interactions in relation to trade liberalisation that can be found in the existing empirical literature. A careful analysis of all the nodes and posited causal connections is required in order to get *valid* evidence for supporting a concrete trade liberalisation policy intervention in a particular country.

7.9. Concluding remarks

I have analysed in detail some of the most influential studies in empirical international economics on the benefits of free trade. The studies show that there is valid evidence of a positive *average causal effect* from a variable defined as "trade liberalisation" to other well-known economic indicators such as GDP per capita, investment, and so on. Nevertheless, I have argued that there are at least three methodological aspects of this literature that make the results ambiguous or inadequate for policy purposes.

1) The variable that is used to measure trade liberalisation is rather ambiguous from a policy-making point of view. It is usually defined as a multidimensional variable that allows measurement, yet it is not clear how a feasible policy intervention (let alone the right one) is supposed to be derived or inferred from it.

2) The specific assumptions about the causal criteria and about the controlling of background conditions that are implicit in the econometric methods used restrict the inferences that can be made about concrete policy applications of the empirical results. It is unclear the precise way in which a causal claim about the existence of an average causal effect of a certain magnitude is supposed to be applied in the design and implementation of a policy reform for a particular country.

3) The general scientific results obtained in these studies are sometimes complemented with analyses of outlying cases, sensitivity analyses, robustness tests, nevertheless in most cases the aim is to save the established truth of a causal generalisation, i.e., to explain away the exceptions such that they can be ignored. From a policy-making perspective, an attitude much more useful would be to offer a clear and explicit delimitation of all the relevant factors, contextual characteristics, and background conditions that interact and

concurrently enable the occurrence of a causal effect. Such information could be used as a basic guideline for policy applications.

A final and significant observation is that most empirical economist working on the research reviewed here seem to be very well-aware of the implicit assumptions and the related limitations that such assumptions impose to making practical inferences about what would happen in specific country economies. Nevertheless, the clarifications, qualms, and qualifications required to properly understand the empirical results seldom reach the actual users of these research results (or at least the systematic reports often used to inform of the available scientific results). There is a huge amount of information implicitly and explicitly contained in the specifics of empirical testing methodologies (just as it is also sometimes implicit in the details of theoretical studies), such information properly organised and presented can be of really great practical value for policy makers trying to attain practical goals and effective policy strategies in real and particular socio-economic situations.

REFERENCES

Adelman, Irma. 1988. Simulation models. In *The new Palgrave: a dictionary of economics*, eds. J. Eatwell, M. Millgate, and P. Newman. London: Macmillan, 340-342.

Ades, Alberto F., and Edward L. Glaeser. 1999. Evidence on growth, increasing returns and the extent of the market. *Quarterly Journal of Economics*, 114 (3): 1025-1045.

Alexandrova, Anna. 2006. Connecting economic models to the real world: game theory and the FCC spectrum auctions. *Philosophy of the Social Sciences*, 36 (2): 173-192.

Andersen, Hanne, and Susann Wagenknecht. 2013. Epistemic dependence in interdisciplinary groups. *Synthese*, 190 (11): 1881-1898.

Angner, Erik. 2006. Economists as experts: overconfidence in theory and practice. *Journal of Economic Methodology*, 13 (1): 1-24.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: an empiricist's companion*. Princeton (NJ): Princeton University Press.

Ashcroft, Richard E. 2004. Current epistemological problems in evidence based medicine. *Journal of Medical Ethics*, 30 (2): 131-135.

Badinger, Harald. 2007. Has the EU's single market programme fostered competition? Testing for a decrease in mark-up ratios in EU industries. *Oxford Bulletin of Economics and Statistics*, 69 (4): 497-519.

Backhouse, Roger. 2004. Serving two masters: economic methodology between philosophy and practice. In *Economic policy under uncertainty*, eds. P. Mooslechner, H. Schuberth, M. Schürz. Cheltenham (UK): Edward Elgar, 180-190.

Baldwin, Richard E., and Elena Seghezza. 1996. Testing for trade-induced investment-led growth. *NBER Working Paper* 5416. National Bureau of Economic Research, Cambridge, MA.

Banerjee, Abhijit V. (ed.). 2007. *Making aid work*. Cambridge (MA): MIT Press.

Banerjee, Abhijit V., and Esther Duflo. 2011. *Poor economics: a radical rethinking of the way to fight global poverty*. New York: Public Affairs.

Baron, Marcia W., Philip Pettit, and Michael Slote (eds.). 1997. *Three methods of ethics: a debate*. Oxford: Blackwell publishing.

Barro, Robert J., 1991. Economic growth in a cross section of countries. *Quarterly Journal of Economics*, 106 (2): 407-443.

Barton, Stuart. 2000. Which clinical studies provide the best evidence? The best RCT still trumps the best observational study. *British Medical Journal*, 321 (7256): 255-256.

Baumol, William J., and Alan S. Blinder. 1997. *Economics: principles and policy*. New York: Harcourt, Brace & Jovanovich.

Baylis, Kathy, Rafael Garduño-Rivera, Gianfranco Piras. 2012. The distributional effects of NAFTA in Mexico: evidence from a panel of municipalities. *Regional Science and Urban Economics*, 42 (1-2): 286-302.

Beebee, Helen. 2007. Hume on causation: the projectivist interpretation. In *Causation, physics, and the constitution of reality: Russell's republic revisited*, eds. Huw Price, and Richard Corry. Oxford: Oxford University Press, 224-249.

Begg, David, Stanley Fischer, and Rudiger Dornbusch. 1994. *Economics*. London: McGraw-Hill.

Ben-David, Dan. 1993. Equalizing exchange: trade liberalization and income convergence. *Quarterly Journal of Economics*, 108 (3): 653-679.

Benson, Kjell, and Arthur J. Hartz. 2000. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 342 (25): 1878-1886.

Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum. 2003. Plants and productivity in international trade. *American Economic Review*, 93 (4): 1268-1290.

Black, Max. 1964. The gap between "is" and "should". *The Philosophical Review*, 73 (2): 165-181.

Black, Max. 1970. *Margins of precision: essays in logic and language*. Ithaca and London: Cornell University Press.

Black, Nick. 1996. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, 312 (7040): 1215-1218.

Blanchard, Olivier. 2006. European unemployment: the evolution of facts and ideas. *Economic Policy*, 21 (45): 5-59.

Blaug, Mark. 1992. *The methodology of economics: or how economists explain*. Cambridge: Cambridge University Press.

Boeri, Tito, and Jan van Ours. 2008. *The economics of imperfect labor markets*. Princeton: Princeton University Press.

Borgerson, Kirstin. 2009. Valuing evidence: bias and the evidence hierarchy of evidence-based medicine. *Perspectives in Biology and Medicine*, 52 (2): 218-233.

Brander, James A., and Barbara J. Spencer. 1985. Export subsidies and international market share rivalry. *Journal of International Economics*, 18 (1-2): 83-100.

Brandom, Robert. 2007. Inferentialism and some of its challenges. *Philosophy and Phenomenological Research*, 74 (3): 651-676.

Brandt, Allan M. 2007. *The cigarette century: the rise, fall, and deadly persistence of the product that defined America*. New York: Basic Books.

Broda, Christian, and David E. Weinstein. 2006. Globalization and the gains from variety. *Quarterly Journal of Economics*, 121 (2): 541-585.

Campaner, Raffaella, and Maria Carla Galavotti. 2007. Plurality in causality. In *Thinking about causes: from Greek philosophy to modern physics*, eds. Peter Machamer, and Gereon Wolters. Pittsburgh (PA): University of Pittsburgh Press, 178-199.

Cartwright, Nancy D. 1974. How do we apply science? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1974: 713-719.

Cartwright, Nancy D. 1979. Causal laws and effective strategies. *Noûs*, 13 (4): 419-437.

Cartwright, Nancy D. 1983. *How the laws of physics lie*. Oxford: Clarendon Press.

Cartwright, Nancy D. 1989. *Nature's capacities and their measurement*. Oxford: Oxford University Press.

Cartwright, Nancy D. 1998. Capacities. In *The handbook of economic methodology*, eds. John B. Davis, D. Wade Hands, and Uskali Mäki. Northampton (MA): Edward Elgar Publishing, 45-48.

Cartwright, Nancy D. 1999. *The dappled world: a study of the boundaries of science*. Cambridge (UK): Cambridge University Press.

Cartwright, Nancy D. 2001. What is wrong with Bayes nets?" *The Monist*, 84 (2): 242-264.

Cartwright, Nancy D. 2002. Against modularity, the causal Markov condition, and any link between the two: comments on Hausman and Woodward. *British Journal for the Philosophy of Science*, 53 (3): 411-453.

Cartwright, Nancy D. 2004. Causation: one word, many things. *Philosophy of Science*, 71: 805-819.

Cartwright, Nancy D. 2006. Well-ordered science: evidence for use. *Philosophy of Science*, 73 (5): 981-990.

Cartwright, Nancy D. 2007. *Hunting causes and using them: approaches in philosophy and economics*. Cambridge (UK): Cambridge University Press.

Cartwright, Nancy D. 2009. Evidence-based policy: what's to be done about relevance? *Philosophical Studies*, 143 (1): 127-136.

Cartwright, Nancy D. 2010. What are randomised controlled trials good for? *Philosophical Studies*, 147 (1): 59-70.

Cartwright, Nancy D. 2011. Predicting 'it will work for us': (way) beyond statistics. In *Causality in the sciences*, eds. P. M. Illari, F. Russo, and J. Williamson. Oxford: Oxford University Press, 750-768.

Cartwright, Nancy D. 2012a. Will this policy work for you? Predicting effectiveness better: how philosophy helps [Presidential Address]. *Philosophy of Science*, 79 (5): 973-989.

Cartwright, Nancy D. 2012b. RCTs, evidence, and predicting policy effectiveness. In *The Oxford handbook of philosophy of social science*, ed. Harold Kincaid. New York: Oxford University Press.

Cartwright, Nancy D. 2013. 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.

Cartwright, Nancy D., and Eileen Munro. 2010. The limitations of randomized controlled trials in predicting effectiveness. *Journal of Experimental Child Psychology*, 16 (2): 260-266.

Cartwright, Nancy D., and Jacob Stegenga. 2011. A theory of evidence for evidence-based policy. In *Evidence, inference and enquiry*, eds. Philip Dawid, William Twining, and Mimi Vasilaki. Proceedings of the British Academy, 171. Oxford: Oxford University Press, 289-319.

Cartwright, Nancy D., and Sophia Efstathiou. 2011. Hunting causes and using them: is there no bridge from here to there? *International Studies in the Philosophy of Science*, 25 (3): 223-241.

Casini, Lorenzo. 2012. Causation: many words, one thing? *Theoria*, 74 (2): 203-219.

Casini, Lorenzo, Phyllis McKay Illari, Federica Russo, and Jon Williamson. 2011. Models for prediction, explanation and control: recursive Bayesian networks. *Theoria*, 26 (1): 5-33.

Claveau, François, and Luis Mireles-Flores. 2014. On the meaning of causal generalisations in policy-oriented economic research. *International Studies in the Philosophy of Science*, 28 (4): 397-416.

Cochrane, Archie L. 1972. *Effectiveness and efficiency: random reflection on health services*. London: Nuffield Provincial Hospitals Trust.

Coe, David T., Elhanan Helpman, and Alexander W. Hoffmaister. 1997. North-south R & D spillovers. *The Economic Journal*, 107 (440): 134-149.

Cohen, Jessica, and William Easterly. 2009. *What works in development? Thinking big and thinking small*. Washington (DC): Brookings Institution.

Colander, David. 1992. The lost art of economics. *Journal of Economic Perspectives*, 6 (3): 191-198.

Colander, David. 1994. The art of economics by the numbers. In Backhouse, R. (editor). *New directions in economic methodology*. London: Routledge, 35-49.

Colander, David. 2001. *The lost art of economics: essays on economics and the economics profession*. Cheltenham: Edward Elgar.

Colander, David. 2005. The making of an economist redux. *Journal of Economic Perspectives*, 19 (1): 175-198.

Colander, David. 2008. The making of a global European economist. *Kyklos*, 61 (2): 215-236.

Colander, David. 2010. The economics profession, the financial crisis, and method. *Journal of Economic Methodology*, 17 (4): 419-427.

Collins, Harry, and Robert Evans. 2007. *Rethinking expertise*. Chicago: University of Chicago Press.

Collins, John, Ned Hall, and Laurie Paul. 2004. *Causation and counterfactuals*. Cambridge (MA): MIT Press.

Dasgupta, Partha. 2005. What do economists analyse and why: values or facts? *Economics and Philosophy*, 21 (2): 221-278.

Davidson, Donald. 1967. Causal relations. *The Journal of Philosophy*, 64 (21): 691-703.

De Vreese, Leen. 2009. Disentangling causal pluralism. In *Worldviews, science and us: studies of analytical metaphysics: a selection of topics from a methodological perspective*, eds. Robrecht Vanderbeeken, and Bart D'Hooghe. Singapore: World Scientific Publishing Co., 207-223.

Dean, Judith M., Seema Desai, James Riedel. 1994. Trade policy reform in developing countries since 1985: a review of the evidence. *World Bank Discussion Paper* 267. International Bank of Reconstruction and Development/The World Bank, Washington, DC.

Doll, Richard, and A. Bradford Hill. 1950. Smoking and carcinoma of the lung: preliminary report. *British Medical Journal*, 2 (4682): 739-748.

Doll, Richard, and A. Bradford Hill. 1954. The mortality of doctors in relation to their smoking habits. *British Medical Journal*, 1 (4877): 1451-1455.

Dollar, David. 1992. Outward-oriented developing economies really do grow more rapidly: evidence from 95 LDCs, 1976-1985. *Economic Development and Cultural Change*, 40 (3): 523-544.

Douglas, Heather E. 2009. *Science, policy, and the value-free ideal*. Pittsburgh: University of Pittsburgh Press.

Dowe, Phil. 2000. *Physical causation*. Cambridge: Cambridge University Press.

Dowe, Phil. 2001. A counterfactual theory of prevention and "causation" by omission. *Australasian Journal of Philosophy*, 79 (2): 216-226.

Driskill, Robert. 2012. Deconstructing the argument for free trade: a case study of the role of economists in policy debates. *Economics and Philosophy*, 28 (1): 1-30.

Drummond, Michael F., Mark J. Sculpher, George W. Torrance, Bernie J. O'Brien, Greg L. Stoddart. 1987. *Methods for the economic evaluation of health care programmes*. Oxford: Oxford University Press.

Ducasse, Curt John. 1926. On the nature and observability of the causal relation. *Journal of Philosophy*, 23 (3): 57-68.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2004. Randomized evaluations of interventions in social service delivery. *Development Outreach*, 6 (1): 26-29.

Dupré, John. 1984. Probabilistic causality emancipated. In *Midwest studies in philosophy IX: causation and causal theories*, eds. P. A. French, T. E. Uehling, Jr., and H. K. Wettstein. Minneapolis: University of Minnesota Press, 169-175.

Earman, John, and John Roberts. 1999. Ceteris paribus, there is no problem of provisos. *Synthese*, 118: 439-478.

Eaton, Jonathan, Samuel Kortum, and Francis Kramarz. 2004. Dissecting trade: firms, industries, and export destinations. *American Economic Review*, 94 (2): 150-154.

Easterly, William, Norbert Fiess, and Daniel Lederman. 2003. NAFTA and convergence in North America: high expectations, big events, little time. *Economía*, 4 (1): 1-53.

Edwards, Sebastian. 1992. Trade orientation, distortions and growth in developing countries. *Journal of Development Economics*, 39 (1): 31-57.

Edwards, Sebastian. 1998. Openness, productivity and growth: what do we really know? *The Economic Journal*, 108 (447): 383-398.

Eells, Ellery. 1987. Probabilistic causality: reply to John Dupré. *Philosophy of Science*, 54 (1): 105-114.

Eells, Ellery. 1991. *Probabilistic causality*. Cambridge: Cambridge University Press.

Falvey, Rod. 1999. Trade liberalization and factor price convergence. *Journal of International Economics*, 49 (1):195-210.

Feenstra, Robert C. 1994. New product varieties and the measurement of international prices. *American Economic Review*, 84 (1): 157-177.

Feenstra, Robert C. 2004. *Advanced international trade: theory and evidence*. Princeton (NJ): Princeton University Press.

Feenstra, Robert C. 2006. New evidence on the gains from trade. *Review of World Economics*, 142 (4): 617-641.

Feenstra, Robert C., and Hiau Looi Kee. 2008. Export variety and country productivity: estimating the monopolistic competition model with endogenous productivity. *Journal of International Economics*, 74 (2): 500-518.

Fleck, Susan, and Constance Sorrentino. 1994. Employment and unemployment in Mexico's labor force. *Monthly Labor Review*, November 1994, 3-31.

Frankel, Jeffrey A., and David Romer. 1999. Does trade cause growth? *American Economic Review*, 89 (3): 379-399.

Freeman, Richard B. 2005. Labour market institutions without blinders: the debate over flexibility and labour market performance. *International Economic Journal*, 19 (2): 129-145.

Friedman, Milton. 1953. The methodology of positive economics. In *Essays in positive economics*, Milton Friedman. Chicago: University of Chicago Press, 3-43.

Gasking, Douglas. 1955. Causation and recipes. *Mind*, 64 (256): 479-487.

Gerring, John. 2007. *Case study research: principles and practices*. Cambridge (UK): Cambridge University Press.

Giere, Ronald. 1979. *Understanding scientific reasoning*. New York: Holt, Rinehart & Winston.

Gilbert, Nigel, and Klaus G. Troitzsch. 1999. *Simulation for the social scientist*. Buckingham (UK): Open University Press.

Glennan, Stuart. 2010. Singular and general causal relations: a mechanist perspective. In *Causality in the sciences*, eds. P. M. Illari, F. Russo, and J. Williamson. Oxford: Oxford University Press, 789-817.

Glymour, Clark. 1997. A review of recent work on the foundations of causal inference. In *Causality in crisis?* eds. Vaughn McKim and Steven Turner. Notre Dame (IN): University of Notre Dame Press, 201-248.

Godfrey-Smith, Peter. 2009. Causal pluralism. In *The Oxford handbook of causation*, eds. Helen Beebee, Christopher R. Hitchcock, and Peter Menzies. Oxford: Oxford University Press, 326-337.

Goldman, Alvin I. 2001. Experts: which one should you trust? *Philosophy and Phenomenological Research*, 63 (1): 85-110.

Good, Irving John. 1961-1962. A causal calculus I-II. *British Journal for the Philosophy of Science*, 11 (44): 305-318; 12 (45): 43-51; Errata and corrigenda, 13 (49): 88.

Greenaway, David, Stephen J. Leybourne, David Sapsford. 1997. Modelling growth (and liberalization) using smooth transitions analysis. *Economic Inquiry*, 35 (4): 798-814.

Greenaway, David, Wyn Morgan, Peter Wright. 2002. Trade liberalisation and growth in developing countries. *Journal of Development Economics*, 67 (1): 229-244.

Grüne-Yanoff, Till. 2011. Mismeasuring the value of statistical life. *Journal of Economic Methodology*, 16 (2): 109-123.

Guala, Francesco. 2005. *The methodology of experimental economics*. Cambridge: Cambridge University Press.

Guyatt, Gordon, John Cairns, David Churchill, Deborah Cook, Brian Haynes, et al. 1992. Evidence-based medicine: a new approach to teaching the practice of medicine. *JAMA: The Journal of the American Medical Association*, 268 (17): 2420-2425.

Hall, Ned. 2004. Two concepts of causation. In *Causation and counterfactuals*, eds. John Collins, Ned Hall, and Laurie Paul. Cambridge (MA): MIT Press, 225-276.

Hands, D. Wade. 2001. *Reflection without rules*. Cambridge: Cambridge University Press.

Hands, D. Wade. 2009. The positive-normative dichotomy and economics. In *Philosophy of economics*, ed. Uskali Mäki; vol. 13 of *The handbook of the philosophy of science*, eds. D. M. Gabbay, P. Thagard, and J. Woods. Amsterdam: Elsevier, 219-239.

Hansen, Fredrik. 2011. The *Stern Review* and its critics: economics at work in an interdisciplinary setting. *Journal of Economic Methodology*, 18 (3): 255-270.

Hardwig, John. 1985. Epistemic dependence. *Journal of Philosophy*, 82 (7): 335-349.

Hardwig, John. 1991. The role of trust in knowledge. *Journal of Philosophy*, 88 (12): 693-708.

Harberger, Arnold C. 1993. The search for relevance in economics. *The American Economic Review*, 83 (2): 1-16.

Harberger, Arnold C. 2004. Issues of economic policy and economic growth. *Working Paper*. University of California, Los Angeles.

Harrigan, Jane, Paul Mosley. 1991. Evaluating the impact of world bank structural adjustment lending: 1980-87. *Journal of Development Studies*, 27 (3): 63-94.

Harris, Richard. 1984. Applied general equilibrium analysis of small open economies with scale economies and imperfect competition. *American Economic Review*, 74 (5): 1016-1032.

Harrison, Ann. 1996. Openness and growth: a time-series, cross-country analysis for developing countries. Journal of Development Economics, 48 (2): 419-447.

Hart, H. L. A., and Tony Honoré. 1985. *Causation in the law*. Oxford: Clarendon.

Hausman, Daniel M. 1992. *The inexact and separate science of economics*. Cambridge (UK): Cambridge University Press.

Hausman, Daniel M. 1998. *Causal asymmetries*. Cambridge: Cambridge University Press.

Hausman, Daniel M. 2005. Causal relata: tokens, types, or variables. *Erkenntnis*, 63: 33-54.

Hausman, Daniel M. 2010. Probabilistic causality and causal generalizations. In *The place of probability in science*, eds. Ellery Eells, and James H. Fetzer. Dordrecht (NL): Springer, 47-63.

Hausman, Daniel M., and Michael S. McPherson. 2006. *Economic analysis, moral philosophy, and public policy*. Cambridge: Cambridge University Press.

Haynes, Brian. 1999. Can it work? Does it work? Is it worth it? The testing of healthcare interventions is evolving. *British Medical Journal*, 319 (7211): 652-653.

Heckman, James J. 2000. Causal parameters and policy analysis in economics: a twentieth century retrospective. *Quarterly Journal of Economics*, 115 (1): 45-97.

Heckman, James J. 2005. The scientific model of causality. *Sociological Methodology*, 35 (1): 1-98.

Heckscher, Eli F. 1919. The effects of foreign trade on the distribution of income. *Ekonomisk Tidskrift*, 21: 497-512.

Heim, Irene, and Angelika Kratzer. 1998. *Semantics in generative grammar*. Malden (MA): Blackwell.

Helpman, Elhanan. 1981. International trade in the presence of product differentiation, economies of scale, and monopolistic competition: a Chamberlin-Heckscher-Ohlin approach. *Journal of International Economics*, 11 (3): 305-340.

Hitchcock, Christopher R. 1993. A generalized probabilistic theory of causal relevance. *Synthese*, 97 (3): 335-364.

Hitchcock, Christopher R. 1995. The mishap at Reichenbach fall: singular vs. general causation. *Philosophical Studies*, 78 (3): 257-291.

Hitchcock, Christopher R. 2001a. Causal generalizations and good advice. *The Monist*, 84 (2): 219-242.

Hitchcock, Christopher R. 2001b. A tale of two effects. *The Philosophical Review*, 110 (3): 361-396.

Hitchcock, Christopher R. 2003. Of Humean bondage. *British Journal for the Philosophy of Science*, 54 (1): 1-25.
Hitchcock, Christopher R. 2007a. Three concepts of causation. *Philosophy Compass*, 2/3 (2007): 508-516.
Hitchcock, Christopher R. 2007b. How to be a causal pluralist. In *Thinking about causes: from Greek philosophy to modern physics*, eds. Peter Machamer, and Gereon Wolters. Pittsburgh (PA): University of Pittsburgh Press, 276-302.
Hitchcock, Christopher R., and James Woodward. 2003. Explanatory generalizations, part II: plumbing explanatory depth. *Noûs*, 37 (2): 181-199.
Hodgson, Geoffrey M. 2009. The great crash of 2008 and the reform of economics. *Cambridge Journal of Economics*, 33 (6): 1205-1221.
Holland, Paul W. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81 (396): 945-960.
Hoover, Kevin D. 2001. *Causality in macroeconomics*. Cambridge: Cambridge University Press.
Hoover, Kevin D., and Stephen J. Perez. 2004. Truth and robustness in cross-country growth regressions. *Oxford Bulletin of Economics and Statistics*, 66 (5): 765-798.
Howick, Jeremy. 2011. *The philosophy of evidence-based medicine*. Chichester (UK): Wiley-Backwell.
Hume, David. 1975a [1739]. *A treatise of human nature*, ed. Lewis Amherst Selby-Bigge, revised by P. H. Nidditch. Oxford: Clarendon Press, Abstract.
Hume, David. 1975b [1777]. Enquiry concerning human understanding. In *Enquiries concerning human understanding and concerning the principles of morals*, ed. Lewis Amherst Selby-Bigge, revised by P. H. Nidditch. Oxford: Clarendon Press, II-VII.
Hummels, David, and Peter J. Klenow. 2005. The variety and quality of a nation's exports. *American Economic Review*, 95 (3): 704-723.
Humphreys, Paul. 1989. *The chances of explanation: causal explanation in the social, medical, and physical sciences*. Princeton (NJ): Princeton University Press.
ILO. 1982. Resolution concerning statistics of the economically active population, employment, unemployment and underemployment. The Thirteenth International Conference of Labour Statisticians. Genève: International Labour Organization.
Imbens, Gido W. and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47 (1): 5-86.
IMF. 1997. *World economic outlook*. Washington (DC): IMF.
Irwin, Douglas. 2009. *Free trade under fire*. Princeton (NJ): Princeton University Press.
Jakob, Christian. 2006. Hitchcock's (2001) treatment of singular and general causation. *Minds and Machines*, 16 (3): 277-287.
Keller, Wolfgang. 2004. International technology diffusion. *Journal of Economic Literature*, 42 (3): 752-782.
Keynes, John Neville. 1917 [1890]. *The scope and method of political economy*. Macmillan and Co.
Kincaid, Harold, John Dupré, and Alison Wylie (eds.). 2007. *Value-free science?: ideals and illusions*. Oxford: Oxford University Press.
Kirman, Alan. 2010. The economic crisis is a crisis for economic theory. *CESifo Economic Studies*, 56 (4): 498-535.
Kitcher, Philip. 1981. Explanatory unification. *Philosophy of Science*, 48 (4): 507-531.
Kitcher, Philip. 2001. *Science, truth, and democracy*. Oxford: Oxford University Press.
Klamer, Arjo, and Jennifer Meeham. 1999. The crowding out of academic economists: the case of NAFTA. In *What do economists know?*, ed. Robert F. Garnett Jr. London: Routledge, 65-85.

Klemperer, Paul. 2003. Using and abusing economic theory. *Journal of the European Economic Association*, 1 (2-3): 245-765.
Koertge, Noretta. 2000. Science, values, and the value of science. *Philosophy of Science*, 67 (Proceedings): S45-S57.
Kornai, János. 1982. *The health of nations: reflections on the analogy between the medical sciences and economics*. Memphis (TE): P. K. Seidman Foundation.
Krueger, Anne O. 1983. *Trade and employment in developing countries 3: synthesis and conclusions*. Chicago: NBER/University of Chicago Press.
Krugman, Paul R. 1979. Increasing returns, monopolistic competition, and international trade. *Journal of International Economics*, 9 (4): 469-479.
Krugman, Paul R. 1980. Scale economies, product differentiation, and the pattern of trade. *American Economic Review*, 70 (5): 950-959.
Krugman, Paul R. 1981. Intra-industry specialization and the gains from trade. *Journal of Political Economy*, 89 (5): 959-973.
Krugman, Paul R. 1993. The narrow and broad arguments for free trade. *The American Economic Review*, 83 (2): 362-366.
Krugman, Paul R., and Maurice Obstfeld. 2009. *International economics: theory and policy*. Boston (MA): Prentice Hall.
Lancaster, Kelvin. 1980. Intra-industry trade under perfect monopolistic competition. *Journal of International Economics*, 10 (2): 151-175.
Law, Averill M., and W. David Kelton. 2000. *Simulation modeling and analysis* [3rd ed.]. Boston: McGraw-Hill.
Lee, Jong-Wha. 1993. International trade, distortions, and long-run economic growth. *International Monetary Fund Staff Papers*, 40 (2): 299-328.
Lehtinen, Aki. 2016. Allocating confirmation with derivational robustness. *Philosophical Studies*, (forthcoming).
Leslie, Sarah-Jane. 2007. Generics and the structure of the mind. *Philosophical Perspectives*, 21 (1): 375-403.
Leuridan, Bert, Erik Weber, and Maarten Van Dyck. 2008. The practical value of spurious correlations: selective versus manipulative policy. *Analysis*, 68 (4): 298-303.
Levin, Morton L., Hyman Goldstein, and Paul R. Gerhardt. 1950. Cancer and tobacco smoking: a preliminary report. *Journal of the American Medical Association*, 143 (4): 336-338.
Levine, Ross, and David Renelt. 1992. A sensitivity analysis of cross-country growth regressions. *American Economic Review*, 82 (4): 942-963.
Lewis, David. 1973. Causation. *Journal of Philosophy*, 70 (17): 556-567.
List, John A. 2011. Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives*, 25 (3): 3-16.
Longworth, Francis. 2009. Cartwright's causal pluralism: a critique and alternative. *Analysis*, 70 (2): 310-318.
Lynch, Michael P. 2009. *Truth as one and many*. Oxford: Clarendon Press.
Machlup, Fritz. 1978. *Methodology of economics and other social sciences*. London: Academic Press.
Mackie, John. 1974. *The cement of the universe: a study of causation*. Oxford (UK): Clarendon.
Madsen, Jakob B. 2007. Technology spillover through trade and TFP convergence: 135 years of evidence for the OECD countries. *Journal of International Economics*, 72 (2): 464-480.
Mäki, Uskali. 1986. Rhetoric at the expense of coherence: a reinterpretation of Milton Friedman's methodology. *Research in the History of Economic Thought and Methodology*, 4: 127-143.

Mäki, Uskali. 1992a. On the method of isolation in economics. In *Idealization IV: intelligibility in science*, ed. Craig Dilworth, special issue of *Poznan Studies in the Philosophy of the Sciences and the Humanities*, 26: 319-354.

Mäki, Uskali. 1992b. The market as an isolated causal process: a metaphysical ground for realism. In *Austrian economics: tensions and new directions*, eds. Bruce J. Caldwell, and Stephan Boehm. Dordrecht: Kluwer Academic Publishers, 35-59.

Mäki, Uskali. 1994. Isolation, idealization, and truth in economics. In *Idealization in economics*, eds. Bert Hamminga, and Neil De Marchi, special issue of *Poznan Studies in the Philosophy of the Sciences and the Humanities*, 38: 147-168.

Mäki, Uskali. 1995. Diagnosing McCloskey. *Journal of Economic Literature*, 33 (3): 1300-1318.

Mäki, Uskali. 2004a. Some truths about truth for economists and their critics and clients. In *Economic policy under uncertainty*, eds. P. Mooslechner, H. Schuberth, and M. Schürz. Cheltenham: Edward Elgar, 9-39.

Mäki, Uskali. 2004b. Realism and the nature of theory: a lesson from J. H. von Thünen for economists and geographers. *Environment and Planning* A, 36: 1719-1736.

Mäki, Uskali. 2006. Remarks on models and their truth. *Storia del Pensiero Economico*, 1: 9-21.

Mäki, Uskali. 2009. Unrealistic assumptions and unnecessary confusions: rereading and rewriting F53 as a realist statement. In *The methodology of positive economics: reflections on the Milton Friedman legacy*, ed. Uskali Mäki. Cambridge (UK): Cambridge University Press, 90-116.

Mankiw, N. Gregory. 1998. *Principles of economics*. New York: The Dryden Press.

Martin, John P. 1996. Measures of replacement rates for the purpose of international comparisons: a note. *OECD Economic Studies*, 26: 99-115.

Martini, Carlo, and Marcel Boumans (eds.). 2014. *Experts and consensus in social science*. Dordrecht (NL): Springer.

McCloskey, Donald. 1983. The rhetoric of economics. *Journal of Economic Literature*, 31: 482-517.

Menzies, Peter. 1989. A unified account of causal relata. *Australasian Journal of Philosophy*, 67 (1): 59-83.

Menzies, Peter, and Huw Price. 1993. Causation as a secondary quality. *British Journal for the Philosophy of Science*, 44 (2): 187-203.

Milberg, William. 1996. The rhetoric of policy relevance in international economics. *Journal of Economic Methodology*, 3 (2): 237-259.

Mill, John Stuart. 1844 [1830]. On the definition of political economy and on the method of investigation proper to it. In *Essays on some unsettled questions of political economy*. London: John W. Parker, 86-114.

Mill, John Stuart. 1874 [1843]. *A system of logic, ratiocinative and inductive: being a connected view of the principles of evidence and the methods of scientific investigation*. New York: Harper.

Mill, John Stuart. 1909 [1848]. *Principles of political economy with some of their applications to social philosophy*, ed. William J. Ashley. London: Longmans, Green and Co.

Milner, Chris, and Peter Wright. 1998. Modelling labour market adjustment to trade liberalisation in an industrialising economy. *The Economic Journal*, 108 (447): 509-528.

Mitchell, Sandra D. 2000. Dimensions of scientific law. *Philosophy of Science*, 67 (2): 242-265.

Mitchell, Sandra D. 2009. *Unsimple truths: science, complexity, and policy*. Chicago and London: University of Chicago Press.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and causal inference: methods and principles for social research*. Cambridge (UK): Cambridge University Press.

Nelson, John S., Allan Megill, and Donald McCloskey. 1987. *The rhetoric of the human sciences: language and argument in scholarship and public affairs*. Madison: University of Wisconsin Press.

Neyman, Jerzy. 1935. Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2 (2): 107-180.

Nickell, Stephen, Luca Nunziata, and Wolfgang Ochel. 2005. Unemployment in the OECD since the 1960s. What do we know? *Economic Journal*, 115 (500): 1-27.

Norton, John D. 2003. A material theory of induction. *Philosophy of Science*, 70 (4): 647-670.

Ñopo, Hugo, and David Colander. 2007. The making of a Latin American global economist. *Middleburry College* Working paper.

OECD. 1994a. The OECD jobs study. *OECD economic outlook*, no. 55. Paris: OECD Publishing, 1-4.

OECD. 1994b. *OECD jobs study: evidence and explanations*. Paris: OECD Publishing.

OECD. 1994c. *OECD jobs study: facts, analysis, strategies*. Paris: OECD Publishing.

OECD. 1998. *Open markets matter: the benefits of trade and investment liberalisation*. Paris: OECD Publishing.

OECD. 2006. *Boosting jobs and incomes: policy lessons from reassessing the OECD jobs strategy*. Paris: OECD Publishing.

OECD. 2010. *OECD employment outlook: moving beyond the jobs crisis*. Paris: OECD Publishing.

Ohlin, Bertil G. 1933. *Interregional and international trade*. Cambridge: Harvard University Press.

Parkin, Michael. 2000. *Economics*. New York: Addison-Wesley.

Pearl, Judea. 2000. *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press.

Pearl, Judea. 2001. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, eds. John Breese, and Daphne Koller. San Francisco (CA): Morgan Kaufmann, 411-420.

Peregrin, Jaroslav. 2012. What is inferentialism? In *Inference, consequence, and meaning: perspectives on inferentialism*, ed. Lilia Gurova. Newcastle upon Tyne: Cambridge Scholars Publishing, 3-16.

Peto, Richard, Alan D. Lopez, Jillian Boreham, and Michael Thun. Mortality from smoking in developed countries 1950-2005 (or later). Clinical Trial Service and Epidemiological Studies Unit (CTSU), updated March 2012, Oxford, UK. http://www.ctsu.ox.ac.uk/~tobacco/index.htm (accessed January 2013).

Petty, Geoffrey. 2006. *Evidence-based teaching: a practical approach*. Cheltenham (UK): Nelson Thornes Ltd.

Psillos, Stathis. 2009. Causal pluralism. In *Worldviews, science and us: studies of analytical metaphysics: a selection of topics from a methodological perspective*, eds. Robrecht Vanderbeeken, and Bart D'Hooghe. Singapore: World Scientific Publishing Co., 131-151.

Putnam, Hilary. 2002. *The collapse of the fact/value dichotomy and other essays*. Cambridge: Harvard University Press.

Putnam, Hilary, and Vivian Walsh. 2011. *The end of value-free economics*. London: Routledge.

Rachels, James. 2003. *The elements of moral philosophy*. New York: McGraw-Hill.

Reiss, Julian. 2004. Evidence-based economics: issues and some preliminary answers. *Analyse & Kritik*, 26: 346-363.

Reiss, Julian. 2005. Causal instrumental variables and interventions. *Philosophy of Science*, 72 (5): 964-976.

Reiss, Julian. 2008. *Error in economics: towards a more evidence-based methodology*. London: Routledge.

Reiss, Julian. 2009a. Causation in the social sciences: evidence, inference, and purpose. *Philosophy of the Social Sciences*, 39 (1): 20-40.
Reiss, Julian. 2009b. Counterfactuals, thought experiments and singular causal analysis in history. *Philosophy of Science*, 76 (5): 712-723.
Reiss, Julian. 2011a. Third time's a charm: causation, science, and Wittgensteinian pluralism. In *Causality in the sciences*, eds. P. M. Illari, F. Russo, and J. Williamson. Oxford: Oxford University Press, 907-927.
Reiss, Julian. 2011b. Empirical evidence: its nature and sources. In *The SAGE handbook of the philosophy of social sciences*, eds. Ian C. Jarvie, and Jesús Zamora-Bonilla. London: SAGE, 551-576.
Reiss, Julian. 2012. Causation in the sciences: an inferentialist account. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 43 (4): 769-777.
Reiss, Julian. 2013. *Philosophy of economics: a contemporary introduction*. London: Routledge.
Reiss, Julian. 2015. A pragmatist theory of evidence. *Philosophy of Science*, 82 (3): 341-362.
Reiss, Julian, and Nancy D. Cartwright. 2004. Uncertainty in econometrics: evaluating policy counterfactuals. In *Economic policy under uncertainty*, eds. P. Mooslechner, H. Schuberth, M. Schürz. Cheltenham: Edward Elgar, 204-232.
Ricardo, David. 1821 [1817]. *On the principles of political economy and taxation*. London: John Murray.
Rodríguez, Francisco, and Dani Rodrik. 2001. Trade policy and economic growth: a skeptic's guide to the cross-national evidence. In *NBER macroeconomics annual 2000*, eds. Ben Bernanke, and Kenneth Rogoff. Cambridge (MA): MIT Press, 261-338.
Romer, Paul M., 1990. Endogenous technological change. *Journal of Political Economy*, 98 (5): S71-S102.
Roush, Sherrilyn. 2009. Randomized controlled trials and the flow of information: comment on Cartwright. *Philosophical Studies*, 143 (1): 137-145.
Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66 (5): 688-701.
Rubin, Donald B. 1977. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2 (1): 1-26.
Rubin, Donald B. 1978. Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6 (1): 34-58.
Rubin, Donald B. 1990. Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25 (3): 279-292.
Rubin, Donald B. 2005. Causal inference using potential outcomes: design, modelling, decisions. *Journal of the American Statistical Association*, 100 (469): 322-331.
Russell, Bertrand. 1912-1913. On the notion of a cause. *Proceedings of the Aristotelian Society*, 13: 1-26.
Russo, Federica, and Jon Williamson. 2007. Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21 (2): 157-170.
Ruzzene, Attilia. 2012. Drawing lessons from case studies by enhancing comparability. *Philosophy of the Social Sciences*, 42 (1): 99-120.
Rybczynski, Tadeusz M. 1955. Factor endowment and relative commodity prices. *Economica*, 22 (88): 336-341.
Sachs, Jeffrey D., and Andrew Warner. 1995. Economic reform and the process of global integration. *Brookings Papers on Economic Activity*, (1): 1-118.
Sackett, David L., William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, W. Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *British Medical Journal*, 312 (7023):71-72.

Sackett, David L., W. Scott Richardson, William M. C. Rosenberg, and R. Brian Haynes. 1997. *Evidence-based medicine: how to practice and teach EBM.* New York: Churchill Livingstone.

Salmon, Wesley C. 1980. Causality: production and propagation. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2: 49-69.

Samuelson, Paul A., and William D. Nordhaus. 1989. *Economics.* New York: McGraw-Hill.

Scheines, Richard. 1997. An introduction to causal inference. In *Causality in crisis?*, eds. Vaughn McKim, and Steven Turner. Notre Dame (IN): University of Notre Dame Press, 185-200.

Selinger, Evan, and Robert P. Crease (eds.). 2006. *The philosophy of expertise.* New York: Columbia University Press.

Sen, Amartya. 1981. Accounts, actions, and values: objectivity of social science. In *Social theory and political practice*, ed. C. Lloyd. Oxford: Clarendon Press.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin Company.

Silverman, David. 2001 [1993]. *Interpreting qualitative data: methods for analysing talk, text, and interaction.* London: SAGE Publications.

Simon, Herbert A. 1954. Spurious correlation: a causal interpretation. *Journal of American Statistical Association*, 49: 467-492.

Skyrms, Brian. 1977. Resiliency, propensities, and causal necessity. *The Journal of Philosophy*, 74 (11): 704-713.

Skyrms, Brian. 1980. *Causal necessity: a pragmatic investigation of the necessity of laws.* New Haven (CT): Yale University Press.

Smith, Alasdair, and Anthony J. Venables. 1988. Completing the internal market in the European Community: some industry simulations. *European Economic Review*, 32 (7): 1501-1525.

Sober, Elliott. 1985. Two concepts of cause. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2 (symposia and invited papers): 405-424.

Solomon, Miriam. 2011. Just a paradigm: evidence-based medicine in epistemological context. *European Journal for Philosophy of Science*, 1 (3): 451-466.

Sosa, Ernest, and Michael Tooley. 1993. *Causation.* Oxford: Oxford University Press.

Speaks, Jeff. 2011. Theories of meaning. *The Stanford Encyclopedia of Philosophy* (summer 2011 edition), ed. Edward N. Zalta.
http://plato.stanford.edu/archives/sum2011/entries/meaning/

Spirtes, Peter, and Richard Scheines. 2003. Causal inference of ambiguous manipulations. *Carnegie Mellon University Department of Philosophy Technical Report* 138.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 1993. *Causation, prediction, and search.* Berlin: Springer.

Spohn, Wolfgang. 2006. Causation: an alternative. *The British Journal for the Philosophy of Science*, 57 (1): 93-119.

Steel, Daniel. 2004. Social mechanisms and causal inference. *Philosophy of the Social Sciences*, 34 (1): 55-78.

Steel, Daniel. 2008. *Across the boundaries: extrapolation in biology and social science.* New York: Oxford University Press.

Stegenga, Jacob. 2011. Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 42 (4): 497-507.

Stegenga, Jacob. 2013. An impossibility theorem for amalgamating evidence. *Synthese*, 190 (12): 2391-2411.

Stevens, Alex. 2011. Telling policy stories: an ethnographic study of the use of evidence in policy-making in the UK. *Journal Social Policy*, 40 (2): 237-255.

Stolper, Wolfgang F., and Paul A. Samuelson. 1941. Protection and real wages. *Review of Economic Studies*, 9 (1): 58-73.
Strevens, Michael. 2004. The causal and unification approaches to explanation unified—causally. *Noûs*, 38 (1): 154-176.
Strevens, Michael. 2012. Ceteris paribus hedges: causal voodoo that works. *Journal of Philosophy*, 109 (11): 652-675.
Summers, Robert, Alan Heston. 1991. The Penn world table (mark 5): an expanded set of international comparisons 1950-88. *Quarterly Journal of Economics*, 106 (2): 327-368.
Suppes, Patrick. 1970. *A probabilistic theory of causality*. Amsterdam: North-Holland.
Suppes, Patrick. 1984. Philosophy of science and public policy. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2 (symposia and invited papers): 3-13.
Svetlova, Ekaterina. 2013. De-idealization by commentary: the case of financial valuation models. *Synthese*, 190 (2): 321-337.
Swain, Marshall. 1978. A counterfactual analysis of event causation. *Philosophical Studies*, 34 (1): 1-19.
Swann, G. M. Peter. 2006. *Putting econometrics in its place: a new direction in applied economics*. Cheltenham (UK): Edward Elgar.
Taylor, John. 2003. *Economics*. New York: Houghton Mifflin Co.
Trefler, Daniel. 2004. The long and short of the Canada-U. S. free trade agreement. *American Economic Review*, 94 (4): 870-895.
Vandenbroucke, Jan P. 2004. When are observational studies as credible as randomised trials? *Lancet*, 363 (9422):1728-1731.
Vandenbroucke, Jan P. 2008. Observational research, randomised trials, and two views of medical science. *PLoS Medicine*, 5 (3): e67.
Wacziarg, Romain. 2001. Measuring the dynamic gains from trade. *The World Bank Economic Review*, 15 (3): 393-429.
Wacziarg, Romain, and Karen Horn Welch. 2008. Trade liberalization and growth: new evidence. *The World Bank Economic Review*, 22 (2): 187-231.
White, Colin. 1990. Research on smoking and lung cancer: a landmark in the history of chronic disease epidemiology. *Yale Journal of Biology and Medicine*, 63 (1): 29-46.
Williamson, Jon. 2005. *Bayesian nets and causality: philosophical and computational foundations*. Oxford: Oxford University Press.
Williamson, Jon. 2008. Causal pluralism versus epistemic causality. *Philosophica*, 77 (1): 69-96.
Williamson, Jon, and Dov M. Gabbay 2004. Recursive causality in Bayesian networks and self-fibring networks. In *Laws and models in the sciences*, ed. Donald Gillies. London: King's College Publications, 173-221.
Winters, L. Alan, Neil McCulloch, Andrew McKay. 2004. Trade liberalization and poverty: the evidence so far. *Journal of Economic Literature*, 42 (1): 72-115.
Woodward, James. 1996. Explanation, invariance, and intervention. *Philosophy of Science*, 64 (supplement): S26-S41.
Woodward, James. 2003. *Making things happen: a theory of causal explanation*. Oxford: Oxford University Press.
Woodward, James. 2010. Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology and Philosophy*, 25 (3): 287-318.
Woodward, James, and Christopher R. Hitchcock. 2003. Explanatory generalizations, part I: a counterfactual account. *Noûs*, 37 (1): 1-24.
World Bank. 1993. *Report on adjustment lending III*. Washington (DC): The World Bank.
World Bank. 2008. *World trade indicators 2008: benchmarking policy and performance*. Washington (DC): The World Bank.

Worrall, John. 2002. 'What' evidence in evidence-based medicine? *Philosophy of Science*, 69 (S3): 316-330.

Worrall, John. 2007. Evidence in medicine and evidence-based medicine. *Philosophy Compass*, 2 (6): 981-1022.

SUMMARY

The idea that knowledge about causal relations can be exploited for the attainment of practical goals has been commonly taken for granted in the literature on causation. However, there is rarely any elaboration on what exactly the practical relevance of causal knowledge amounts to. The way causal knowledge can be exploited for practical purposes is typically understood as follows: if 'X *causes* G' is true, then bringing about X "will be an effective strategy for G in any situation" (Cartwright 1979, 432). Leuridan, Weber, and Van Dyck (2008) have labelled this position the "standard view on the practical value of causal knowledge" (p. 298).

While overall the standard view is taken as given, little attention has been paid to the details of what the practical relevance of causal knowledge actually amounts to. What kind of practical power is causal knowledge in fact capable of conferring? How exactly is one meant to make use of causal knowledge for the effective attainment of practical goals in specific situations? Is the same kind of practical power present in all forms of causal knowledge? How different evidence and evidential techniques used to test causal relations affect practical relevance. Are all the methods used to evaluate causal relations appropriate to support or evaluate practical recommendations as well? A serious philosophical effort to investigate this kind of questions is still missing.

The present dissertation is an attempt to begin providing such a philosophical account of the practical relevance of causal knowledge in economics. The focus is on the notions of 'causation' and 'evidence' in the process of using economic science for policy purposes. Overall, the result is a contribution to *the methodology of economic policy-making*.

In **chapter 1**, I argue that the twofold positive-normative distinction in economics has led methodologist to focus either on issues related to conceptual analysis of economic science and its methods or on the ethical and welfare consequences of economic approaches. Still, the ways in which economic science is used (or misused) to achieve policy goals have been rarely analysed and essentially ignored by most philosophers of economics. I explain this situation by referring to the classical threefold distinction in economics as proposed by J. S. Mill and J. N. Keynes. They recognised three distinct epistemic activities related to economics: scientific economics, normative economics, and the art of economic policy-making. After reassessing the classical and the modern accounts of this distinction, I argue that a revised version of the threefold classical distinction can offer significant benefits to the study and understand of the methodology of economics. The aims of scientific inquiry are very distinct from the aims of policy making, thus the criteria of success and the methods required also significantly differ. This chapter serves as an account of why a more detailed methodology of policy making would be beneficial, and

also makes explicit the questions that motivate and are analysed in the rest of the dissertation.

In **chapter 2**, I discuss the notions of 'practical relevance' and 'effectiveness' in relation to causal knowledge. In the standard view, the notion of 'practical relevance' is confined exclusively to one of its possible meanings, namely a specific form of intervening power, while in fact causal knowledge can contribute to the attainment of practical goals in many different and equally important ways. I discuss how predictive power and intervening power relate to each other and to causal knowledge. The complexity of the notion of practical relevance becomes clearer by distinguishing between the 'practical potential' of causal knowledge (which depends on its predictive and intervening power), and the notion of 'effectiveness' (which refers to the actual accomplishment in relation to concrete practical goals). The aim of this chapter is to argue that practical potential is not sufficient for the effectiveness of causal knowledge.

In **chapter 3**, I discuss the relation between causal pluralism and practical relevance. 'Causality' has many different meanings, some of which are mutually exclusive. The philosophical literature provides us with theories of causality put in terms of regularities, probability raising, counterfactual dependence, physical processes, and invariance under interventions. In addition, there are distinct causal concepts referring to dissimilar configurations of causal factors within a given causal structure. These include for instance net cause, component cause, average cause, preventer, and so forth. As a result, a generic causal claim of the form 'X causes Y' can have a variety of meanings and interpretations. A proper disambiguation of the meaning of causal relations should complement any attempt to use causal knowledge for practical purposes. The aim of this chapter, then, is to show that a proper philosophical account of the practical relevance of causal knowledge should make clear how different notions of a causal claim relate to different forms of practical relevance.

In **chapter 4**, a case study is used to illustrate the main points made in the previous chapter. The OECD research on unemployment in the 1990s was presented to the members of the OECD as an up-to-date scientific study including the available theoretical and empirical findings from economic science on the causes of, and the effective strategies against, the problem of unemployment. Most of the findings were stated in the form of causal generalisations. It is not clear, however, what the precise meaning of these causal claims is. The example shows (1) that the meaning of causal claims is not always clear when they are intended as guides for policy; (2) that the ambiguities in the meaning of a causal generalisation are of different types and come from distinct sources; and (3) that it is rather unclear which precise policy recommendations are to be implemented in particular targets, given that the results of the OECD study are mainly about average causal effects.

In **chapter 5**, I discuss how it is possible to infer effective strategies for particular country-cases from causal generalisations. I start by exploring the existing philosophical literature on the relation between general- and unit-level causation. I argue that most proposals fail to offer a clear and useful account. I then show how the potential-outcomes framework commonly used to establish causal generalisations in economics and other social sciences relies on a number of conceptual and methodological assumptions close to those used in probabilistic accounts, which at the same time enable and constrain the inferences that can be made from general results to the unit level. Causal generalisations are typically presented to non-academic audiences as scientific results without any definite provisos about how they are meant to apply to individual members of a population, or what the specific requirements of the background context of application are.

In **chapter 6**, I suggest that causal generalisations and their practical relevance are inferentially connected to certain features of the *evidential base* employed to established them, and then elaborate on such features. An evidential base is the set of all the evidence that supports and allows an epistemic agent or community to believe that a piece of scientific knowledge is true. The first part of this chapter is devoted to critically review the existing philosophical literature on the so-called evidence-based movement and Nancy Cartwright's "evidence for use" approach. I argue that the methodological worries related to the evidence-based movement and of Cartwright's account are not really useful to understand the evidential requirements relevant to most policy purposes in economics. In contrast to many other sciences, randomised controlled trials (RCTs) are not the gold standard of evidence in economics. Empirical evidence and evidential techniques used in economics mostly come from econometrics, which has a number of specific methodological consequences for economic policy-making. In the second part of the chapter, I discuss the basic problems of policy inference in relation to economics. Following Christopher Hitchcock, I identify different types of causal claims so as to characterise more precisely both: *causal generalisations* about large populations and *policy hypotheses* about particular units. I then propose an account of evidence to support policy hypotheses that is distinct from, and yet grounded on, the typical ways in which empirical economists obtain evidence for causal generalisations.

In **chapter 7**, I illustrate what I have said in the previous chapters by analysing in detail the empirical literature in international trade economics that studies the benefits of free trade. A huge amount of empirical research has been produced to test the hypothesis that (more) trade liberalisation causes (more) economic improvements (in terms of growth, investment, employment, and the like). I analyse the evidence and evidential methods employed. Three methodological issues emerge: (1) the lack of definite referents for the relevant causal relata, and specially for the posited causal variable “trade liberalisation”; (2) the use of econometric methods as a “gold standard” of scientific evidence, which always come with a built-in conception of causality; and (3) a tendency to seek generality over specificity in the results of the research. The pursuit of generality has been originally intended to enhance the confidence in the truth of purported policy-oriented causal generalisations, but it now seems to be backfiring by leading to scientific results in the form of broad causal generalisations with no clear ways to be applied in specific real cases. Even if the existing empirical evidence could be taken as acceptable to support that there is “some” causal relation between trade liberalisation and economic benefits, the existing evidence is not conclusive at all to support concrete policy recommendations for individual countries and their specific socio-economic conditions.

SAMENVATTING

De literatuur over oorzakelijkheid doet vermoeden dat kennis van causale verbanden zonder al te veel problemen gebruikt kan worden voor praktische doeleinden. Maar het is helemaal niet zo duidelijk welke epistemische status dergelijke praktische kennis precies heeft. De praktische relevantie van causale kennis wordt typisch als volgt begrepen: als de propositie 'X is de oorzaak van G' waar is, dan zal het doen plaatsvinden van X "een effectieve strategie zijn voor G in alle situaties" (Cartwright 1979, 432). Leuridan, Weber, and Van Dyck noemen deze positie "het standaardbeeld van de praktische waarde van causale kennis" (2008, 298).

Welk praktisch gewicht brengt oorzaakskennis nou precies met zich mee? Hoe moet zulke causale kennis gebruikt worden ten behoeve van praktische doelen in specifieke situaties? Is dezelfde graad van praktische relevantie verbonden met alle soorten causale kennis? Verschillen in het gebruik van empirisch bewijs en bewijsmiddelen bij het testen van oorzakelijke verbanden hebben gevolgen voor de praktische relevantie ervan. Zijn alle middelen van de evaluatie van causale relaties geschikt voor het ondersteunen van praktische aanbevelingen? We ontberen een degelijk filosofisch perspectief op dit soort vragen.

Dit proefschrift doet een poging om zo'n filosofische benadering van de praktische relevantie van causale kennis op het gebied van de economie te ontwikkelen. De aandacht gaat vooral uit naar de begrippen 'oorzaak' en 'bewijs' in de economie bij beleidskwesties. Zo ontstaat een bijdrage aan de *methodologie van economisch beleid*.

In **Hoofdstuk 1** laat ik zien dat methodologen door het tweevoudige feit-norm onderscheid in de economie hetzij in zijn gegaan op conceptuele kwesties in economie en zijn methode, hetzij op welvaartsvraagstukken en op de ethische consequenties van economisch denken. Tegelijk is de vraag naar het gebruik (of misbruik) van economische kennis voor beleidsdoeleinden genegeerd. Ik verklaar deze stand van zaken met behulp van een drievoudig onderscheid in soorten van epistemische activiteit zoals bedacht door J. S. Mill en J. M. Keynes: wetenschappelijke economie, normatieve economie en het ambacht van economische beleidsinterventies. Na een herbezinning op de klassieke en moderne versies van dit onderscheid kom ik met een herziene versie. Ik beweer dat deze interessante voordelen biedt voor het begrijpen van de aard van de economie. Het doel van wetenschappelijk onderzoek is zeer verschillend van het doel van beleid. Dan mag je verwachten dat de succescriteria en de methoden die ervoor nodig zijn ook uiteenlopen. Dit hoofdstuk verhaalt dus vooral wat een zekere detaillering in de methodologie van economisch beleid oplevert om daaruit de vraagstelling van dit proefschrift af te leiden en te motiveren.

Hoofdstuk 2 gaat in op de begrippen ‘praktische relevantie’ en ‘doelmatigheid’ in verband met causale kennis. In het standaardbeeld krijgt ‘praktische relevantie’ uitsluitend de betekenis van één *specifieke* interventiemogelijkheid. Maar causale kennis kan bijdragen aan praktische doelen op tal van manieren die alle even belangrijk mogen heten. Ik bespreek het verband tussen verklarende kracht en interventievermogen. De notie van praktische relevantie blijkt complex en de distinctie tussen de begrippen ‘praktische relevantie van causale kennis’ en ‘doelmatigheid’ helpt die complexiteit te analyseren. De eerste hangt af van verklarende kracht en interventievermogen, de tweede gaat over het feitelijke succes bij het bereiken van concrete beleidsdoelstellingen. Het doel van dit alles is om te adstrueren dat louter het vermogen om iets te bereiken in de praktijk onvoldoende is voor het vaststellen van de doelmatigheid van causale kennis.

In **hoofdstuk 3** bespreek ik de relatie tussen causaal pluralisme en praktische relevantie. Veroorzaking is een notie met meerdere betekenissen en sommige van die betekenissen sluiten elkaar uit. Zo kennen we theorieën van oorzakelijkheid in termen van regelmatigheden, van waarschijnlijkheidsgroei, van contrafactuele afhankelijkheid, van fysieke processen en van invariantie onder interventies. Bovendien verwijzen onderscheiden concepten van oorzakelijkheid naar diverse configuraties van causale factoren binnen één gegeven causale structuur. Denk aan netto-oorzaak, samengestelde oorzaak, gemiddelde oorzaak, preventie-oorzaak en dergelijke. Zo kan een causale bewering van de vorm ‘X veroorzaakt Y’ een hele reeks van interpretaties genereren. Een poging om causale kennis te gebruiken in de beleidspraktijk moet met een beduidende desambiguering gepaard gaan. Het doel van dit hoofdstuk is dan ook om te tonen dat een zinnige filosofische benadering van de praktische relevantie van onze kennis over de oorzakelijke structuur van de wereld vereist dat de diverse mogelijke interpretaties van een bepaalde causale uitspraak in samenhang gebracht worden met verschillende vormen van praktische relevantie die aan de orde kunnen zijn.

Hoofdstuk 4 illustreert het voorgaande met een *case study.* De OECD presenteerde een onderzoek over werkloosheid in de negentiger jaren aan OECD-leden als wat toen de meest recente theoretische en empirische economische inzichten waren over de oorzaken van werkloosheid en over maatregelen ertegen. De meeste conclusies waren geformuleerd in termen van causale generalisaties. Het is echter onduidelijk hoe deze conclusies precies begrepen moeten worden. Preciezer gezegd, het voorbeeld onthult (1) dat de betekenis van causale claims vaak niet helder is als deze bedoeld zijn om beleid te schragen, (2) dat de resulterende ambiguïteit van de generalisaties voortkomt uit meerdere bronnen en (3) dat – gegeven het feit dat steeds een concept van ‘gemiddelde oorzaak’ aan de orde is – mistig blijft welke beleidsaanbevelingen nou precies gedaan worden.

Ik kom in **hoofdstuk 5** toe aan de vraag hoe doelmatige beleidsstrategieën voor landen af te leiden zijn uit causale generalisaties, die verwijzen naar de veronderstelde werkzame principes. Eerst ga ik na wat de literatuur te zeggen heeft over de relatie tussen wetmatige en singuliere causaliteit. Ik kom tot de conclusie dat de meeste uitwerkingen niet voldoen als het gaat om helderheid en bruikbaarheid voor beleidskwesties. Vervolgens laat ik zien hoezeer het kader van 'potentiële uitkomsten' – dat veel gebruikt wordt om causale generalisaties in economie en andere sociale wetenschappen te formuleren – berust op conceptuele en methodologische veronderstellingen die in de buurt komen van wat we vooral in probabilistische benaderingen terugzien. Deze aannames scheppen wel de mogelijkheid om uit de werkzame principes conclusies te trekken op het niveau van een enkele casus, maar beperken die tegelijk ook. Het is typerend dat men de causale generalisatie presenteert aan niet-academisch publiek zonder enige voorbehouden met betrekking tot de toepassing ervan voor afzonderlijke elementen uit een populatie. Of er is geen indicatie van wat vereist is van de contextuele factoren van een interventie.

Hoofdstuk 6 presenteert het idee dat causale generalisaties en hun praktische relevantie logisch verbonden zijn met bepaalde eigenschappen van de *bewijsbasis* die wordt ingezet om die generalisaties vast te stellen. Dit is alles wat een epistemische actor of gemeenschap doet geloven dat bepaalde wetenschappelijke kennis waar is. Ik onderzoek de eigenschappen van die bewijsbasis in dit hoofdstuk nader. Eerst bespreek ik de filosofische literatuur rond de zogenaamde *evidence based* beweging; met name Nancy Cartwright's *evidence-for-use* benadering. Ik laat zien dat de methodologische kwesties die in deze benaderingen aan de orde gesteld worden te weinig te bieden hebben als het gaat om een beter begrip van de voorwaarden die aan deugdelijk bewijs gesteld mogen worden voor het voeren van economisch beleid. In de economie zijn *Randomised Controlled Trials* (RCTs) niet de Gouden Standaard; dit in tegenstelling tot veel andere wetenschappen. Economisch empirisch bewijs en bewijsvoeringtechnieken zijn vooral afkomstig van de econometrie en dit heeft nogal wat specifieke methodologische consequenties voor economisch beleid. In het tweede deel van dit hoofdstuk bespreek ik de fundamentele problemen die spelen bij het afleiden van de kennis die nodig is voor economisch beleid. Ik identificeer de soorten van oorzakelijke uitspraken in navolging van Christopher Hitchcock om zowel causale generalisaties over hele populaties als beleidshypothesen betreffende enkele eenheden preciezer te karakteriseren. Ten slotte volgt een verhandeling over bewijsvoering voor beleidshypothesen die verschilt van maar gebaseerd is op de typische manier waarop empirisch georiënteerde economen bewijs verzamelen voor causale generalisaties.

Hoofdstuk 7 illustreert wat in het voorgaande is gezegd met behulp van een detailanalyse van de empirische literatuur over de voordelen van vrijhandel. Zeer veel onderzoek toetst de hypothese dat handelsliberalisatie economische voordelen biedt in groei, investeringen, werkgelegenheid en dergelijke. Het bewijs voor deze hypothese en de gebruikte bewijsmethoden worden aan een analyse onderworpen. Er komen drie methodologische kwesties naar voren: (1) er is een tekort aan referenten voor de relevante causale relata, met name 'vrijhandel', (2) het gebruik van econometrische methoden als Gouden Standaard voor bewijs draagt onvermijdelijk reeds een concept van causaliteit in zich en (3) er is een neiging om algemene uitspraken de voorkeur te geven boven specifieke claims over de onderzoeksresultaten. De neiging te algemeniseren komt oorspronkelijk voort uit het streven de waarheidswaarde vast te kunnen stellen van de causale generalisaties die voor beleidskeuzen bedoeld zijn. Maar deze wens verkeert inmiddels in zijn tegendeel. Wetenschappelijke resultaten hebben de vorm gekregen van brede generalisaties die geen heldere clous opleveren voor toepassing in specifieke gevallen in de werkelijkheid. Zelfs als gegeven empirisch bewijs voldoet om te geloven dat er een of ander oorzakelijk verband bestaat tussen handelsliberalisatie en bepaalde nastrevenswaardige economische voordelen, dan nog geeft dat bewijs de regering in een bepaald land – immers met geheel eigen socio-economische condities – onvoldoende uitsluitsel over het te volgen beleid.

ABOUT THE AUTHOR

Before completing his PhD in philosophy, Luis Mireles-Flores received a MPhil in philosophy and economics at the Erasmus Institute for Philosophy and Economics (EIPE), Erasmus University Rotterdam, and a MSc in philosophy of the social sciences at the London School of Economics and Political Science (LSE). He obtained a BA in economics from the Instituto Tecnológico Autónomo de México (ITAM). He is co-founder and (from 2008 to 2015) was co-editor and webmaster of the *Erasmus Journal for Philosophy and Economics*. Since 2010, he is secretary of the Sociedad Iberoamericana de Metodología Económica (SIAME). He is currently a researcher at the Finnish Centre of Excellence in the Philosophy of the Social Sciences (TINT), Department of Social and Moral Philosophy, University of Helsinki.