CAUSAL MECHANISMS: A POTENTIAL TOOL FOR ECONOMIC POLICY?

A CASE STUDY APPROACH

Research Master Thesis

Felix Joachim Matthias den Ottolander
(342598)

Supervisor: Prof. Dr. J. J. Vromen
Advisor: Dr. J. W. Hengstmengel

Date of completion: 01.07.2016
Number of ECTS: 30
Word count: 29.235

MA, Research Master Philosophy and Economics
Erasmus Institute for Philosophy and Economics (EIPE)
Faculty of Philosophy
Erasmus University Rotterdam

A run on the bank (Berlin, 13 July 1931).

PREFACE

"The curious task of economics is to demonstrate to men how little they really know about what they imagine they can design."

F. A. von Hayek, *The Fatal Conceit: The Errors of Socialism (1988)*

Bank runs are an economist's worst nightmare. Liquidity dries up, banks – literally – close their doors, the economy stagnates, unemployment rises, public order deteriorates, and politicians are ousted from office. All this sounds painfully familiar. The past decade has witnessed the biggest bank run in modern history, as well as several near-bank runs, with devastating effects. Despite the fact that the general causes are well understood, economists have a hard time preventing history to repeat itself. Yes, they have tried to minimize the costs for taxpayers by moving from bailing out banks to bailing in shareholders. Moreover, there has been a recent effort in setting up a banking union in Europe with the aim of eradicating bank runs once and for all.

Although we may understand how bank runs evolve, it remains very difficult to effectively control the factors that play a role in this process. Causal mechanisms, showing the entities and processes that lead up to certain events, have proven useful for economists, sociologists and historians in the explanation of phenomena. This thesis, as the title suggests, asks whether causal mechanisms can also be useful within the domain of economic policy. Since the connotation of 'mechanism' in economics differs significantly from its philosophical conceptualisation, a great deal of attention will be devoted to bridge this gap. More importantly, the invocation of causal mechanisms for policy purposes raises several methodological issues, to which this thesis aims to contribute.

As with a sizable project like this, some words of gratitude are in order. First of all, I would like to sincerely thank Jack Vromen, my supervisor, for taking the time to provide sharp and insightful comments on earlier drafts. Thanks also to my advisor Joost Hengstmengel who managed to read and comment on my thesis in just one day, which significantly improved its structure, clarity and readability. Though only involved in the very beginning of the supervision process, Attilia Ruzzene, my ex-supervisor and personal tutor, has been a great source of inspiration. In the same vein, many thanks to Philippe Verreault-Julien and Julian Reiss who were so generous to have a look at my initial thesis proposal. I would also like to thank the organizers of the OZSW Graduate Conference in Theoretical Philosophy (April, 2016), in particular Koray Karaca, who granted the opportunity to present my thesis proposal and provided useful comments as well.

On a different note, thanks to my good friend and fellow campus trotter Luc van de Ven with whom I enjoyed numerous lunches and dinners the past few years. Lastly, a special shout-out to Kees Kraaijeveld and Marlot van der Stoel – my roommate and girlfriend, respectively – who, as non-philosophers, had to endure the inevitable consequences of someone doing a two-year Master's programme in analytical philosophy.

TABLE OF CONTENTS

INTRODUCTION

The economic and financial turmoil in Greece has been unfolding for over six years now. Despite the relatively calmness of the past year, Greece's systemic problems of low economic growth, rising public debt and a weak banking sector have recently resurfaced. Finance ministers and other policy makers from the Eurozone are currently reviewing the progress made by Greece with respect to its third bailout programme, which was agreed upon in July 2015 after weeks of uncertainty and speculation. A similar scenario now seems to have been set in motion, for most of the elements that played a role in Greece's "hot summer"[1] have unfortunately remained in place.

Clearly, this poses a severe threat for European policy makers. Their primary aim is to bring back economic and financial stability to Greece: sustainable public finances, renewed growth in Gross Domestic Product (GDP) and a well-functioning banking sector. Most important of all, policy makers want to prevent Greek banks from collapsing due to uncertainty about their ability to meet depositors' short term demands. Over the past six years, there have been numerous institutions and events involved in the build-up of Greece's problems. A key set of events took place at the end of June and beginning of July 2015, when Greece experienced a run on some of its largest banks. If policy makers want to succeed in their attempt to keep Greek banks operational in times of crisis, then it is imperative they understand the causal mechanism of a bank run.

A. GETTING STARTED: THE SELF-FULFILLING PROPHECY

The phenomenon of a bank run has been thoroughly studied by both economists and social scientists. Bank runs are commonly understood to occur when initial beliefs about the insolvency of a particular bank or group of banks ultimately lead to a large amount of panicked depositors trying to take up their funds. The fact that some depositors are making substantial withdrawals prompt others to do so as well out of fear that they might be left empty-handed when the bank eventually runs of out of money. Crucially, the financial position of the bank is further weakened by the continuous flow of withdrawals – whether the initial worries were justified or not. This downward spiral culminates in a literal 'run on the bank', where anxious depositors queue in front of bank offices and ATMs trying to salvage what they can (see page 2).

Robert Merton (1948) has been one of the first sociologists to formulate an explanation for the existence (and persistence) of bank runs. According to his influential account, a bank run

[1] Silvia Merler (2015), an Affiliate Fellow at the Bruegel think-tank in Brussels, was one of the first who had anticipated further financial problems in Greece last summer.

is essentially an expression of a self-fulfilling prophecy. By drawing on a basic theorem in social science – "If men define situations as real, they are real in their consequences." (ibid.: 193) – Merton claims that people do not only respond to the objective features of a given situation, but also to the subjective meaning they ascribe to that situation. That is, depositors might be prompted to believe that a particular bank has serious solvency issues due to some general rumour or previous withdrawals by others. What matters for them is not whether the rumour is actually true, but what the possible consequences could be in case people *believe* the rumour to be true.

In fact, a self-fulfilling prophecy is always false initially, in the sense that the anticipated consequences would not come about if people did not act on the basis of their beliefs. In other words, the initial *false* beliefs concerning a situation trigger actions by people that in turn make those beliefs come *true*. This shows us that certain false beliefs – in the form of rumours, speculations or what have you – can "become an integral part of the situation and thus affect subsequent developments" (ibid.: 195). Since a self-fulfilling prophecy often involves a range of different causal variables within a temporal structure, is has been described as a proper causal mechanism.

The characterisation of the self-fulfilling prophecy as a causal mechanism has become evident in Greece over the course of last summer. The story started with Alexis Tsipras, Greece's prime minster, announcing a referendum to be held on July 5 about an additional bailout proposal by the Troika[2]. This announcement was followed by the decision of the Troika to suspend its negotiations with the Greek government and let the existing bailout programme expire. Moreover, the ECB decided not to increase its emergency funding to the Greek banking sector. Due to severe uncertainty about the future of the euro as a viable currency for Greece, depositors increased their withdrawals and began to hoard large amounts of cash.[3] In order to avoid widespread bankruptcies, the ECB initiated a couple of last-resort measures: capital controls, a national bank holiday and a temporary closure of Greek stock markets. The bank run finally subsided when the Greek government and the Troika agreed upon a third bailout programme on July 12.

Now that another hot summer is looming for Greece, the causal mechanism of a self-fulfilling prophecy is becoming relevant again for policy makers. Assuming they do not want to repeat last year's scenario, policy makers are concerned with the question *when and where to intervene* in a potentially similar chain of events. In case of a bank run, it is important for policy makers to know at what stage they can stop the destructive process from unfolding further. For instance, given

[2] This term refers to the three main creditor institutions – the European Commission (EC), the European Central Bank (ECB) and the International Monetary Fund (IMF) – that jointly conduct the negotiations with Greece with regard to its conditional bailout payments.
[3] For more details about the deposit outflows from Greek banks, see Harari (2015: 17).

that the decision by the ECB not to increase its emergency funding to the Greek banking sector contributed to increased uncertainty and withdrawals, the ECB could decide to increase its emergency funding when it finds itself in a similar situation in the future.

More importantly, policy makers want to know *how to intervene* on a particular development effectively. Knowing at what stage of a bank run to intervene is often not enough; in addition, policy makers need to have evidence about what kind of intervention is likely to be effective. These kinds of considerations are often context-dependent, which means that the effectiveness of potential interventions largely depends on the specific conditions of the situation at hand. When considering the option to increase emergency funding in case of stagnating bailout negotiations, policy makers of the ECB will have to assess whether this option will lead to the desired outcome – i.e. stopping the formation of a bank run. This particular intervention may be effective for the aim of stopping a bank run, but it might be less effective for different purposes. Indeed, the decision of the ECB not to increase its emergency funding was aimed at forcing the Greek government to rejoin the bailout negotiations.

This thesis aims to defend the claim that mechanisms[4] are a useful tool for economic policy-making. The example of the 2015 bank run in Greece illustrates in what sense mechanisms could be useful for economic policy makers. However, the concept of mechanisms has received relatively little attention within the realm of economic policy. Instead, most philosophers of science and social scientists have focused on whether mechanisms can contribute to the explanation of social and economic phenomena. They have been mainly concerned with the explanatory power of mechanisms, which has resulted in a view of mechanisms as an explicit tool for causal inquiry.[5]

Although explanation of economic phenomena is an important aim for the academic community of economists, philosophers of science and social scientists in general, it is less relevant for policy makers. With respect to the earlier bank run example, policy makers can clearly benefit from the causal knowledge that mechanisms are able to provide: the self-fulfilling prophecy tells them that queuing depositors are likely to act on the basis of some initial false belief about bank insolvency. Yet what policy makers care about even more is whether they can control a certain situation with the tools at their disposal. *Prima facie*, mechanisms such as the self-fulfilling prophecy appear to be relevant for policy makers in this sense because they provide

[4] From this point onwards, I will refer to 'mechanisms' without mentioning the 'causal' component explicitly. As will be explained in chapter one, the reason is that the interpretation of mechanisms with respect to explanation is less important for the purposes of policy makers, which are more concerned with controlling economic phenomena instead.

[5] Peter Hedström and Petri Ylikoski (2010) review some of the most important contributions to causal mechanisms in the social sciences. This literature will be elaborated upon in the early sections of chapter one.

them with knowledge about how to intervene. If, for instance, policy makers could somehow prevent rumours about bank insolvency from spreading, then the formation of a bank run might be effectively avoided.

B. MAIN ARGUMENTS

The conceptualisation of mechanisms with respect to policy-making will be the central theme of the first chapter. By drawing on examples from sociology and economics, it will become clear what mechanisms generally look like, and what role mechanisms play in the different desiderata of science: explanation, prediction and control. While explanation and prediction are important aims for economics in general, they are less relevant from a policy point of view.

After sorting out the different functions of mechanisms, the two main arguments in defence of using mechanisms for policy purposes will be introduced. The first refers to the methodological problem of external validity, which reflects the difficulty of exporting causal relationships outside their artificial environments. According to some accounts[6], mechanisms are able to resolve the problem of external validity by specifying the similarity in background conditions between the artificial and target environments. These background conditions are then incorporated into the interventions of policy makers so as to make them more effective. The claim that mechanisms are actually able to support the effectiveness of interventions – also referred to as mechanism-based extrapolation – is, however, rather controversial: interventions are likely to alter the causal structure they wish to exploit. Here I will make two claims: one, mechanisms play an important role in the design and implementation of policy interventions; second, mechanism design, though susceptible to structure-altering interventions, can be a suitable extension of mechanism-based extrapolation, and should thus be taken seriously by the economic policy-making community.

The second argument relates to evidential relevance. From this alternative perspective, mechanisms have the ability to provide a preliminary understanding of the evidence that could be relevant for the effectiveness of policy interventions. Here the interpretation of mechanisms as causal scenarios will be introduced, where each causal scenario starts with the proposed intervention and ends with the desired outcome, and is considered plausible according to some basic theory, general principle, or widely-held public opinion. Specifically, I will make two claims with respect to mechanisms and evidential relevance: one, policy makers adopt the perspective of evidential relevance since they take the aims of interventions as their point of departure; second, they proceed in establishing evidential relevance by constructing and evaluating different

[6] Most notably that of Francesco Guala (2005, 2011) and Daniel Steel (2008).

causal scenarios. The scenarios themselves do not act as evidence for the effectiveness of interventions; rather, they are a tool for policy makers to being their search for relevant evidence in a preliminary manner.

C. OUTLOOK ON CASE STUDIES

As mentioned already, to what extent mechanisms can actually be useful for policy makers differs per case. For this reason, the potential of mechanisms for policy interventions will be assessed with the help of two case studies. Chapter two deals with mechanisms in the domain of auction policy. Here I will discuss one prominent case in which game theory has been successfully applied to the design of real auctions, namely the 1994 Federal Communications Commission (FCC) spectrum auctions in the United States. The choice for this particular case study rests on the fact that the auctions involved a multibillion-dollar business, which not only had a significant impact on public finances but also affected the economic performance of the Clinton administration. More importantly, the FCC auctions can be characterised as a kind of *economic engineering* where knowledge of mechanisms was used to design a new mechanism that worked well in the actual context of the spectrum auctions. In addition, several philosophers of science[7] have extensively engaged in the debate around auction design and its policy implications over the past few decades, which provides ample material to advance our assessment of mechanisms for policy-making.

Chapter three is concerned with mechanisms in the domain of behavioural policy. Here the case study consists of the *Save More Tomorrow*™ (SMT) pension plan as described by Richard Thaler and Shlomo Benartzi (2004, 2013). This particular intervention intended to increase savings contribution rates amongst employees by drawing on insights from behavioural economics. Although the SMT plan was satisfactory in terms of raising contribution rates, it was not clear *how* the intervention actually made a difference – i.e. through what mechanism the results were obtained. Since policies that use behavioural insights – also referred to as 'nudging'[8] – are quickly gaining popularity, investigating the role of mechanisms will prove helpful in this regard. According to Till Grüne-Yanoff (2015), behavioural policy (as the SMT pension plan) needs mechanistic evidence because, without specifying the operating mechanism, policy makers cannot sufficiently justify their interventions. The question then is, of course, which mechanism will count as sufficient to the justification of interventions. One option would be for policy makers to justify interventions according to the welfare effects inferred by mechanisms.

[7] Most notably, Francesco Guala (2005, 2011), Anna Alexandrova (2006), Anna Alexandrova and Robert Northcott (2009), and Alvin Roth (2002).
[8] For a seminal contribution to the concept of nudging, see Thaler and Sunstein (2008).

CHAPTER ONE

THE MEANING AND FUNCTION OF MECHANISMS

Over the past few decades, philosophers of science and social scientists have debated about the role of mechanisms[9] as a method of inquiry. Many early contributions have focused on delineating a proper definition of a mechanism, which has resulted in an extensive and rather confusing collection of definitions. For instance, Hedström and Ylikoski (2010: 51) list no less than nine different definitions of a mechanism put forward in the social sciences.[10] Mechanisms, it seems, come in a variety of shapes and sizes.

To make some sense of this conceptual mess, the many definitions of a mechanism can be separated into two groups: horizontal mechanisms and vertical mechanisms. The horizontal interpretation of mechanisms describes the intermediate causal processes that take place between a certain cause and effect. One of the most prominent and relatively straightforward horizontal definitions has been put forward by Daniel Little, who describes a mechanism as a "series of events governed by law-like regularities that lead from the explanans to the explanandum" (1991: 56). According to his account, a causal analysis of a certain phenomenon is constituted by a mechanism that identifies all (probabilistic) conditions to be causally relevant to the occurrence of an effect. In this sense, a mechanism presents the pathway through which a set of different causes consecutively lead to the observed effect. An example of a horizontal mechanism is presented in figure 1.

The vertical interpretation of mechanisms, on the other hand, is often more complex because it also explicitly deals with causal processes that operate on different levels of reality. Besides listing the causally relevant events in the order they supposedly occur, a vertical mechanism also describes how phenomena can be explained by, so to say, zooming in or out on certain causal processes. This reductionist approach is supposed to explain the occurrence of phenomena more accurately because it shows how micro-level variables are responsible for macro-level variables. What a vertical mechanism looks like and how this explanatory process goes about will be discussed in the next section.

---

[9] Throughout this chapter and the following chapters, I use the terms 'mechanisms' and 'mechanistic models' interchangeably. Mechanisms are, essentially, a type of model that uses assumptions to describe, explain, predict and control phenomena.

[10] Some of these definitions specifically refer to 'social mechanisms' (Hedström & Swedberg, 1998), i.e. mechanisms dealing with the explanation of social phenomena, such as revolutions, wars, etc. Since this thesis is less concerned with explanation and focuses more on economic phenomena, I will not elaborate on this particular terminology.

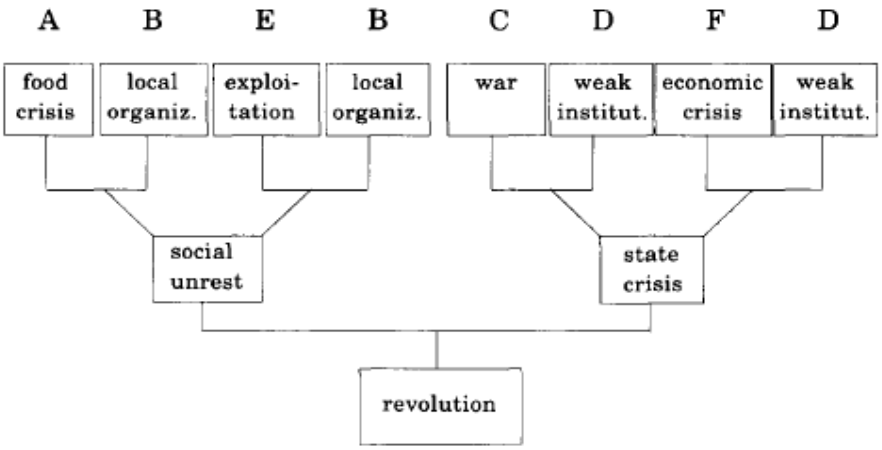

Figure 1: A horizontal mechanism (Little, 1995).

For now, it is important to emphasize that both groups of mechanisms are relevant to the topic of this thesis. In what follows, I will argue that policy makers can use mechanisms – whether they are horizontal or vertical, or some combination of the two – to better control economic phenomena. What matters here is not whether one group of mechanisms is more or less useful than the other in terms of explanation, but if mechanisms *in general* can be useful for policy purposes. In this sense, I do not wish to endorse one explicit definition (or group of definitions) of mechanisms, for the aim of this thesis is not to engage in a conceptual discussion as summarized by Hedström and Ylikoski (2010). Rather, its approach is methodological: to what extent, if at all, can mechanisms be useful within the economic policy-making process?

1.1. COLEMAN'S BOAT

To illustrate the general structure and operationalisation of mechanisms, it is useful to introduce a classic example by the famous sociologist Max Weber. In his principal thesis, *The Protestant Ethic and the Spirit of Capitalism* (1905/1930), Weber claims that the religious doctrine of Protestantism is chiefly responsible for the rise of our modern capitalist economic system. The causal relationship is in this case of a macro-to-macro type: one macro variable (Protestantism) causes another macro variable (capitalism) to occur. Figure 2 shows this particular causal relationship in a simple mechanistic model.

Presented in this way, a mechanism does not seem to be any different from the covering-law model of explanation advocated by Carl Hempel (1942/2011). In his account, to explain the occurrence of a phenomenon entails referring to a general causal law. For an explanation to be satisfactory, it must specify both the general law and the conditions that make the law applicable

in the particular case. With respect to Weber's thesis, the occurrence of capitalism could be explained by the general law that purportedly exists between religion and economic systems.

Figure 2: Macro-to-macro causal relationship between Protestantism and capitalism.

Yet Hempel admits that the existence of deterministic laws is highly unlikely in the social sciences. Instead, what at most can be invoked are laws of a probabilistic nature, i.e. laws that state the probability with which a particular phenomenon will come about given certain background conditions. Often there is merely a statistical association between two phenomena of interest, in which case "the specific explanation will offer no more insights than the law itself and will usually only suggest *that* a relationship is likely to exist, but it will give no clue as to *why* this is likely to be the case" (Hedström & Swedberg, 1998: 8, original italics). For this reason, mechanisms need to go beyond the macro-to-macro type of causal explanation.

One viable extension is for mechanisms to include causal relationships on the micro level as well. This approach is broadly defined as methodological individualism, which aims to explain macro-level phenomena in terms of micro-level entities such as individual behaviour. The causal relationship is now of a macro-micro-macro type: a macro variable (Protestantism) is reduced to a micro variable (individual values), that causes another micro variable (orientations to economic behaviour), which in turn is transformed back into a macro variable (capitalism). Figure 3 captures this extended mechanistic model of Weber's thesis.

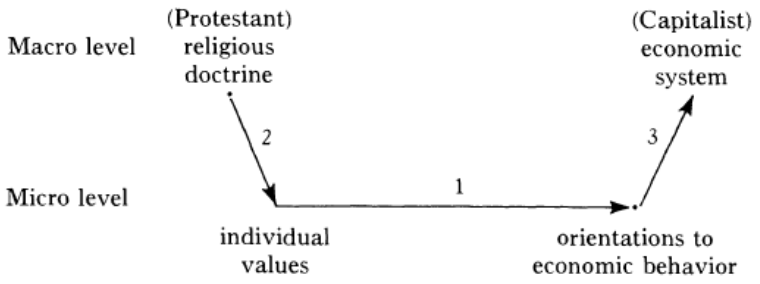

Figure 3: Macro-micro-macro causal relationship between Protestantism and capitalism. This general causal structure is referred to as Coleman's Boat, after its originator James Coleman (1986).

This type of explanation, based on methodological individualism[11], has been proposed by the sociologist James Coleman (1986) and has since then aptly been called 'Coleman's Boat'.[12] Although there are arguably many different theoretical mechanisms available for the purpose of explaining social and economic phenomena, Coleman's Boat is a good place to start our discussion of mechanisms in general. Before the analysis turns to the potential of mechanisms with respect to policy-making, it is important to pay a bit more attention to how mechanisms can be used as an explanatory tool. The reason is that most mechanistic accounts developed in the social sciences so far have aimed at acquiring a better explanation of the phenomena under investigation. If policy makers want to gain a general understanding of what mechanisms are and what they can do for them, then they at least have to see one of the functions of mechanisms – i.e. the desideratum of explanation – in action.

According to Hedström & Swedberg (2010), the idea of mechanistic explanations has partly arisen due to the shortcomings of Hempel's covering-law model. Granting this is true, how is a mechanism like Coleman's Boat able to determine why it is likely that a causal relationship between two or more variables exists? Essentially, a mechanism provides an explanation of *how* a certain effect could have come about. This implies that the effect in question can be produced by a number of different mechanisms, known or unknown to the investigator. To increase the plausibility of one mechanism over another, empirical evidence about the assumed entities, activities and relationships must be collected and systematically analysed. Thus, mechanistic explanations try to describe causal processes as accurately as possible, while being selective about the different aspects of those processes and by disregarding irrelevant details.

This is precisely what Coleman tries to achieve in his analysis of Weber's thesis. By selecting two causally relevant variables on a micro level – individual values and orientations to economic behaviour – the causal relationship between Protestantism and capitalism is presumably explained more accurately. In short, Protestant religious doctrine places the values of worldly possessions and soberness upon individuals in society; these values then lead to a positive attitude towards private enterprises and hard work; which eventually manifests itself in a capitalist mode of production and consumption within society. So there are three processes active in this particular mechanism: macro-to-micro (arrow 2 in figure 3), micro-to-micro (arrow 1) and micro-to-macro (arrow 3).

[11] Although Daniel Little thinks that Coleman's Boat presupposes some form of methodological individualism, Petri Ylikoski (forthcoming) argues this view is mistaken. Anyway, this point is not relevant to our discussion since Coleman's Boat is merely used as an illustration of what mechanisms in economics may look like and how they generally operate.

[12] Coleman's Boat, also referred to as the Coleman's diagram and Coleman's bathtub, has become one of the most famous theoretical mechanisms in sociology. After introducing it in his 1986 paper, Coleman used it extensively in his magnum opus *Foundations of Social Theory* (1990).

There are, of course, many other macro-level variables besides Protestant religious doctrine that could, in theory, explain the occurrence of a capitalist economic system, such as an abundance of natural resources and labour.[13] Nevertheless, the inclusion of micro-level variables gives strength to the causal claim because it elucidates how a certain cause leads to the observed effect. Another alleged advantage of mechanistic explanations is the fact that its assumptions can be empirically tested. The micro-to-micro causal relationship that individual religious values lead to orientations to capitalist behaviour, for instance, can be validated by sociological or anthropological research that studies the professions and general economic behaviour of Protestant communities. Since a range of psychological variables operate at the micro level, more precise methodological tools – such as experiments – will be available in order to test causal hypotheses. This is not to say that studying these kinds of micro-level hypotheses will be easy and will always bring about satisfactory results, but they are generally considered to be more reliable than mere macro-level explanations.

Although including micro-level explanations might be a reliable method in supporting macro-level explanations, they also introduce new problems. One key problem is that once micro-level explanations are established, it is difficult to transform their effects back to the macro level. In other words, the micro-to-macro process of a mechanism (arrow 3) often proves to be insufficient in explaining how some individual causal relationship results in a collective phenomenon. This is also the case with Weber's account, in which he fails to show "how these individual orientations combined to produce the structure of economic organisation that we call capitalism (if in fact they did in combination produce this effect)" (Coleman, 1986: 1323). Thus, whether to include micro variables when explaining macro-level phenomena not only depends on the strength of the micro-level explanation itself, but also on the ability of transforming the latter back into the macro phenomenon under investigation.

1.2. ALTERNATIVE FUNCTIONS OF MECHANISMS

Since the introduction of mechanisms in the social sciences, including economics, their aim has mainly been to enhance the explanation of phenomena. For example, the mechanism of a self-fulfilling prophecy has helped to explain how bank runs develop: false beliefs about bank insolvency trigger withdrawals by depositors, which eventually justify those initial beliefs.

[13] For example, the rise of China as an economic power could be seen as a counterexample to Weber's thesis because it has developed a largely capitalist economic system despite the almost complete lack of Protestant religious doctrine, or any other religious doctrine for that matter. A more plausible mechanistic explanation for the development of Chinese capitalism would feature, among other variables, an abundant supply of natural resources and cheap labour.

Likewise, the causal structure of Coleman's Boat has increased our understanding of the causal relationship between religion and capitalism by incorporating macro- and micro levels of explanation.

Despite the fact that explanation is an important aim – or function, as I will refer to it – of mechanisms, it is not the only one that is relevant to the use of mechanisms. In fact, there are two other functions that are distinct from, and at least as equally important as, explanation. These non-explanatory functions need to be distinguished in order to assess the usefulness of mechanisms for policy makers. In this section, some of the conceptual differences between the functions of explanation, prediction and control with respect to mechanisms will be discussed. Most attention will be devoted to the function of control because policy makers, as illustrated in the introduction, are mostly interested in how they can effectively intervene upon economic phenomena.[14]

1.2.1. PREDICTION

The first alternative function of mechanisms that will be discussed is prediction. Within economics, the aim of predicting the future values of certain variables is of great significance: correctly predicting next month's stock market index hugely benefits traders and investors; entrepreneurs will only stay in business if they can accurately predict future consumer demand for their products and services; during election periods, politicians heavily rely on GDP forecasts that show how their policy proposals will impact the economy. Admittedly, correct predictions are notoriously difficult to come by in economics, but it is clearly worthwhile for economists and policy makers to pursue this aim nonetheless.

From a methodological point of view, prediction has been the dominant desideratum of economic theory since at least the latter half of the 20th century. In his seminal paper *The Methodology of Positive Economics* (1953/2008), Milton Friedman argues that the "ultimate goal of a positive science is the development of a 'theory' or 'hypothesis' that yields valid and meaningful (i.e. not truistic) predictions about phenomena not yet observed" (ibid: 148). So what makes an economic model successful is whether it can provide accurate predictions about future phenomena, not whether it sufficiently explains the occurrence of observed phenomena. In this sense, economists are mostly concerned *if* their models can predict some phenomenon of interest, while remaining agnostic about *how* their models generally do so.

[14] Besides the effectiveness of interventions, policy makers also care about a variety of other considerations, such as costs and whether interventions are morally acceptable. This thesis focuses on the methodological issues related to the function of control and therefore remains agnostic about these other kinds of considerations that policy makers might have.

This implies that the assumptions incorporated in economic models do not necessarily have to be realistic in order to obtain accurate predictions. In fact,

> "Truly important and significant hypotheses will be found to have 'assumptions' that are wildly inaccurate descriptive representations of reality, and, in general, the more significant the theory, the more unrealistic the assumptions. … To be important, therefore, a hypothesis must be descriptively false in its assumptions; it takes account of, and accounts for, none of the many other attendant circumstances, since its very success shows them to be irrelevant for the phenomena to be explained." (ibid.: 153).

According to Friedman, the question whether the assumptions of a model are realistic is completely irrelevant. Instead, what matters is to what extent the assumptions result in accurate predictions. To use a simple example: the assumption of unlimited human rationality is obviously unrealistic, but it is allowed to be included in the model because it enables (more) accurate predictions of human behaviour. As a result, the appropriateness of a theory can only be judged according to its predictive capabilities.

The claim that assumptions can be descriptively false for models to predict well turns out to be rather problematic in the case of mechanistic models. Recall that mechanisms primarily aim to describe *how* phenomena come about, i.e. through what causal processes. Coleman's Boat became influential precisely because it could show in what way micro-level entities determined macro-level phenomena. The often complex structure of mechanistic models[15] is a result of their aim to specify the most salient causal variables that lead to some particular effect. The presumed advantage of mechanisms in terms of explanatory power is largely based on their ability to exhibit an empirically verifiable chain of causal claims.

Now if we accept Friedman's position, which a substantial part of the economics community has done, then the main advantage of mechanisms becomes largely undone. More specifically, by embracing the idea that predictions can be successful when the underlying assumptions are unrealistic, it makes little sense to empirically test whether the causal relationships postulated by mechanisms are actually realistic. In fact, inquiring about mechanisms would not be useful at all because what matters is *that* predictions are accurate, not *why* they are.

To illustrate this point, let us extend our simple example of unlimited human rationality as an accurate predictor of economic behaviour. Besides perfect rationality, let's assume there are two more unrealistic assumptions that support the successful prediction of economic behaviour,

[15] Daniel Little's (1995: 33) extension of the causal mechanism responsible for revolution is a good example of how mechanisms can quickly adopt a complex structure, with many different probabilistic relationships operating at the same time.

namely perfect information and fully flexible market prices. One could, in principle, construct and test a hypothetical mechanism that somehow links the three assumptions to the predicted economic behaviour, such as three separate causal relationships that equally influence the effect (see figure 4).

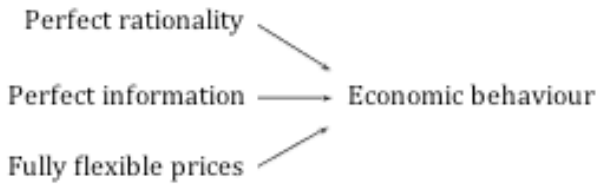

Figure 4: Simple example of a nonsensical mechanism that shows how a set of unrealistic assumptions predict a phenomenon (economic behaviour).

Or one could rearrange the order of the assumptions in the mechanism as to show that, for instance, perfect information causes perfect rationality, which together with the assumption of fully flexible market prices predicts economic behaviour. The point is that it is simply not important to know how predictions actually came about. Rather, it is only relevant to know that a set of (unrealistic) assumptions contributed to the prediction of a particular phenomenon. We could try to empirically verify these assumptions as postulated by some kind of mechanism but this would not benefit the prediction in any meaningful way.

Thus, it seems that mechanisms are not very useful when it comes to predicting economic phenomena. If successful prediction requires unrealistic assumptions, as Friedman suggests, then mechanistic models stating the order and interactions of these assumptions possess little added value. Mechanisms might perform better at explaining phenomena, with their potential to empirically test the assumptions of causal relationships, but it seems this ability cannot be exploited for predictive purposes.

At this point, however, one could object to this conclusion by stating that explanation and prediction are the two sides of the same coin. The so-called Symmetry Thesis (Hempel, 1958) holds that once a phenomenon is adequately explained, its occurrence can also be potentially predicted; similarly, the adequate prediction of a phenomenon implies its observation can be potentially explained. So Hempel believed every adequate explanation to be a potential prediction, and *vice versa*. In this sense, explanatory mechanisms automatically have predictive powers as well: assuming Coleman's Boat adequately explains the rise of capitalism, it is also able to predict future capitalist economic systems based on the same analysis.

Unfortunately, the Symmetry Thesis has been charged with numerous counter-examples[16] and has subsequently gone out of fashion. The argument that mechanisms have predictive power because they are capable of providing adequate explanations is simply false, since this kind of reasoning is subject to the same criticism raised against the Symmetry Thesis. The fact that Coleman's Boat can adequately explain the rise of capitalism does not mean it can at the same time predict all instances of capitalist economic systems, as the development of capitalism in contemporary China nicely illustrates.[17] Therefore, the objection towards mechanisms' lack of predictive power based on the Symmetry Thesis is unwarranted.

So far, two conclusions with respect to the function of prediction for mechanisms can be drawn. First, knowledge of mechanisms does not benefit the prediction of phenomena in any significant way because economists and policy makers are not particularly interested in *why* their predictions will be accurate.[18] Instead, they are concerned whether certain predictions will hold, while disregarding the possibility that the unrealistic assumptions could be characterized by some mechanistic model. Secondly, even when mechanistic models adequately explain phenomena this does not imply they have predictive power as well, for the criticised Symmetry Thesis does not hold in the case of mechanisms either.

1.2.2. CONTROL

Now that the functions of explanation and prediction have been discussed we can move on to the third function of mechanisms, namely control. The previous discussion about prediction will help clarify how mechanisms relate to the control of phenomena. This particular function is the central theme by which the potential of mechanisms for policy makers will be assessed. To emphasize, policy makers are mainly concerned with questions about how they can effectively intervene upon causal processes in order to bring about some preferred outcome. More specifically, they want to know how to manipulate causal variables so that a particular target variable takes on a specific value.

This means interventions by policy makers are often tied to some specific aim, which is not only being able to control a target variable in itself but also to control it in some preferred way. For example, in the case of a bank run, policy makers not only want to know whether they can influence the outflow of bank deposits. Rather, they want to be sure that their intervention in

[16] See, for instance, Scriven (1962).
[17] See footnote 13.
[18] Of course, it would be nice for policy makers to know about the exact composition of the causal variables that have contributed to their correct prediction. This is certainly useful information since the same composition can be used in order to derive similar correct predictions in the future. Yet in most cases it is more important for policy makers to trust their predictions in terms of accuracy.

terms of daily limits on withdrawals indeed reduces the outflow of capital. Thus, the function of control often takes on a highly concrete, context-specific form.

Explanation obviously plays a role in this process but it must, like prediction, be distinguished from control. Basically, to explain a certain phenomenon by way of a mechanism is to look back and determine what causal variables were jointly responsible for the occurrence of the phenomenon. Once this is done, the policy maker has presumably, to some extent, increased his causal knowledge of the phenomenon of interest: the fact that a bank run is set in motion by the mechanism of a self-fulfilling prophecy is, in and of itself, very important for policy makers to be aware of. Causal knowledge of this mechanistic type then acts as the starting point for further inquiries upon where and how to intervene upon the causal processes in order to control their outcome in some desirable way. Mechanistic explanation is thus an important but preliminary step in the aim of policy makers to control phenomena.

Whereas prediction is a function that involves a two-way relationship between a set of causal variables and a set of target variables, the function of control is essentially a three-way relationship between a set of (candidate) causal variables, a set of feasible interventions and a set of target variables. Control can therefore be seen as a special kind of prediction, where policy makers do not passively observe the outcome of some set of causal variables but in addition try to actively – i.e. by some intervention or sequence of interventions – bring about some value of the target variable. While predicting the future value of target variables is highly significant for economists and policy makers, being able to *control* the value of target variables is even more significant for quite obvious reasons. With respect to macroeconomics, policy makers are continually trying to control a range of different variables, such as interest rates, inflation rates and unemployment rates. Moreover, our bank run example has shown that controlling the amount of deposit withdrawals is crucial for maintaining financial stability within the economy.

The function of control is also very important in microeconomics, where the consequences for individuals and institutions are of primary concern. Here policy makers do not only want to explain or predict certain behaviour, but rather want to control behavioural processes so as to guarantee a desired outcome. These outcomes may be expressed in terms of fair and efficient behaviour by individuals, or some sufficient amount of government revenue.[19] Alternatively, policy makers might wish to optimise people's decisions with regard to saving for retirement.[20] There are clearly numerous instances, in both macro- and microeconomics, where policy makers engage in the design of interventions by which they aim to bring about some desirable outcome.

[19] Chapter two will discuss one specific case in which the aim of policy makers was to design a fair and efficient set of auctions that would generate sufficient government revenue.
[20] Likewise, chapter three will discuss one specific case in which the aim of policy makers was to influence employees to increase their savings contribution rates.

1.2.3. MECHANISMS AND CONTROL

The basic intuition behind the function of mechanisms for control is that they can help policy makers identify where and how to intervene upon causal processes in order to bring about some preferred outcome. In this thesis, I will defend the claim that mechanisms are a useful tool for policy makers by way of two main arguments. The first argument has to do with the methodological problem of external validity. This problem basically refers to the fact that it is often very difficult, if not impossible, to extrapolate causal relationships obtained in one context to another. Mechanisms are – or so it is claimed by some methodologists – able to overcome this problem because they show how the causal relationships actually operate in different contexts. Supposedly, the mechanistic solution to the problem of external validity lies in its ability to specify the necessary background conditions that enable the extrapolation of some causal relationship to the target environment.

Largely based on the principle of mechanism-based extrapolation, there is one particularly interesting approach: mechanism design. Here policy makers do not only focus on extrapolating a particular theoretical or experimental result by invoking mechanisms, but instead try to design a new kind of mechanism that best fits the conditions of the target environment. During this procedure, policy makers study many different mechanisms and integrate the most useful aspects – be it bits of theory, some experimental result or a salient background condition – into their final design. This way, the use of mechanisms for policy purposes can be seen as a kind of *economic engineering*, in which the need for extrapolation is combined with more inductive elements. The next section will further elaborate on the procedure of designing mechanisms in this regard.

The second argument is more general and comes down to the ability of mechanisms to provide a preliminary understanding of the evidence that could be relevant for the effectiveness of policy interventions. Although evidence for efficacy is important, its relevance for the target context needs to be determined first.[21] Essentially, drawing on the efficacy of causal variables in some artificial environment is only one piece of evidence for policy hypotheses; what is also needed is knowledge of the relevance of these efficacies for the policy (target) environment. While a theorist or experimenter asks in what context his obtained results might be relevant, a policy maker is more interested in what kind of evidence might be relevant for his particular policy aim. Unfortunately, the latter perspective has been relatively ignored in the debate about mechanisms within the realm of policy-making.

One way to account for this perspective is to interpret mechanisms as causal scenarios, which policy makers set up in order to gather evidence that is potentially relevant for the policy

[21] Section 1.4 will elaborate on the difference between evidence for efficacy and evidence for effectiveness, and what role this difference plays in establishing evidential relevance.

aim at hand. Each scenario – starting with a proposed intervention and ending with the desired outcome – has to be plausible according to some basic theory or public opinion. To be clear, the mechanisms (causal scenarios) themselves do not act as evidence for the effectiveness of interventions; rather, they are a tool for policy makers to search for evidence and evaluate its relevance in a preliminary manner. Once the evidential relevance has been established, policy makers can then empirically test the credibility of this evidence in the target context.

The remainder of this chapter will introduce the general features of external validity and evidential relevance with respect to mechanisms, after which the case studies in the subsequent chapters will further substantiate the arguments.

1.3. THE PROBLEM OF EXTERNAL VALIDITY

This section will deal with some of the main aspects of mechanisms with respect to external validity, focusing on why external validity is viewed as a serious methodological problem and how mechanisms are supposedly able to resolve this issue. Most importantly, the procedure of mechanism design will be given special attention as it appears to be a promising approach for policy makers.[22] The next section will introduce the notion of evidential relevance and will show in what sense mechanisms indicate which evidence could be relevant to the aims of policy makers.

External validity – or 'extrapolation' as Daniel Steel (2008) frames it – has been a central methodological problem for experimental economics. To understand the problematic aspects of external validity it is useful to contrast it with *internal* validity, which indicates that the causal relationships found within an experimental setting (such as in a laboratory) are indeed valid. That is, internal validity is established when the causality of an experimental result is properly understood by the investigator. Somewhat more formally: the effect E of an experiment X can be considered internally valid if the production of E is attributed to a cause (or set of causes) C, and C really is responsible for E in X. There are numerous problems related to internal validity too, but these are not relevant for our current discussion of mechanisms.[23] For the problem of external validity to occur, one must assume an experimental result to be internally valid: it does not make much sense to investigate whether a result is valid outside the experimental setting unless we have reason to believe that it is therein.

[22] For one thing, two Nobel prizes in economics have been awarded to works on mechanism design: one to Leonid Hurwicz, Eric Maskin and Roger Myerson (in 2007); and the other to Alvin Roth and Lloyd Shapley (in 2012).

[23] One key problem in experimental settings is that of confounding factors, where it is not clear if a causal factor leads to the observed effect or if it influences the effect via some other (unobserved) causal factor.

What is problematic about the validity of experimental results in economics, then, is that they are hardly ever obtained in contexts outside the laboratory. In his book *Across the Boundaries* (2008), Steel defines extrapolation as "the challenge of transferring causal generalizations from one context to another when homogeneity cannot be presumed" (ibid.: 3). The absence of homogeneity means that the population to which the experimental result is extrapolated differs with respect to characteristics that affect the originally obtained causal relationships. In other words, transferring causal knowledge from one situation to another often does not work because the complexities of the latter make it unsure whether the causal relationships will hold. The fact that C causes E in one experimental setting, which may include highly stringent background conditions upon the agents and materials involved, does not necessarily mean that C will also cause E in another setting where the background conditions might be very different.

A straightforward way to illustrates the problem of external validity is the example of the Bangladesh Integrated Nutrition Policy (BINP).[24] Based on the previously successful Tamil Nadu Integrated Nutrition Project (TINP), the aim of the BINP was to decrease the amount of malnourished children in Bangladesh. A focal point of both policies was to educate pregnant women and young mothers to better nourish their children as well as themselves. Although the two policies were very similar in their set up, the environments in which they were implemented differed in several important respects. One aspect proved to be particularly significant in the failure of the BINP to decrease malnutrition among children: it is not the mother but the paternal grandmother who usually takes care of children in Bangladesh. While mothers were effectively targeted by the TINP, in the case of the BINP the intervention did not achieve the same results. The education of mothers may lead to a decrease in malnutrition among children in one case, but it may not do so when similar educational programmes are introduced elsewhere. Thus, the example shows that even though knowledge of a causal relationship may be valid in one particular environment, it may cease to exist when applied to a different environment.

1.3.1. MECHANISM-BASED EXTRAPOLATION

Given the increasing reliance of economic policies on experimental results – generally referred to as 'evidence-based policy' by scholars such as Nancy Cartwright and Julian Reiss – it is imperative for policy makers to be aware of the problem of external validity and, even more importantly, to know how to mitigate its main implications. One prominent proposal to overcome this problem makes use of mechanisms, or what Steel calls 'mechanism-based extrapolation' (Steel, 2008: 85). While Steel is quite sceptical about the ability of mechanisms to enable successful extrapolation

[24] This example is borrowed from Cartwright and Hardie (2012).

in economics, our discussion will first present the main argument in favour of mechanism-based extrapolation.

Over the past decade, Francesco Guala (2005, 2011) has been one of the key proponents of mechanisms in applied economics. In general, he argues that mechanisms are able to bridge the gap between experimental results on the one hand, and their application in real world situations on the other. This is done by way of specifying the similarities between the experimental and target populations. Mechanisms establish these similarities in terms of *background conditions*, which together indicate the likeliness of a causal relationships to hold in the target environment. For example, knowledge of a mechanism that included the social norms within Bangladesh's society would have increased the chances of success for the BINP. If policy makers would have been aware of the fact that the similarities between the two populations (i.e. the experimental population of the TINP and the target population of the BINP) were limited with respect to family structure and children's care, then they would have been able to adapt their intervention accordingly.

To clarify this point, assume there was only some piece of abstract theory – without any empirical evidence whatsoever – that described the positive causal relationship between the education of mothers and nutrition of children in developing countries. In this situation, policy makers must first know whether this causal relationship will be confirmed empirically at all. They can begin their inquiry by devising an experiment in which the supposed causal relationship is put to the test. From these experimental results, which turned out overwhelmingly positive in the case of the TINP, policy makers then have some reason to believe that the causal relationship will also hold in the context of their interest (i.e. Bangladesh).

However, it is here that policy makers have to pay attention to the dissimilarities between the experimental and the target contexts. In this case, for instance, background knowledge such as social norms played an important role in extrapolating the results of the TINP to the context of Bangladesh. These kinds of background conditions about the target context are crucial for policy makers, as they determine whether their interventions will prove successful or not. Sometimes, knowledge of background conditions has to be obtained from a wide range of sources, such as other scientific disciplines.[25]

For Guala, the main advantage of using mechanisms for economic policy is the *integration* of theoretical, experimental and background knowledge into a unique, new mechanism that is able to perform effectively in the target environment. He describes the design of mechanisms as "an enterprise between theoretical and applied economics, which requires first stating clearly the

[25] In case of the BINP, "identification of the 'mother-in-law' effect came from reading anthropological literature" (White, 2009; cited by Cartwright 2012: 988).

goals to be achieved by the mechanism, and then finding the best means to achieve them given the circumstances" (Guala, 2005: 164). So, once the aim of a particular intervention has been determined, policy makers need to study different mechanisms that could potentially be useful to achieve this aim. Essentially, the procedure of mechanism design makes use of many different kinds of knowledge, including hypotheses about mechanisms that purportedly exist in the real world and could potentially be used for extrapolation.

Yet where it differs from traditional mechanism-based extrapolation is in its creative aspect, i.e. in the construction (or engineering) of a wholly new mechanism that did not exist before. According to Guala, policy makers adapt the mechanisms that presumably exist in the real world so as to make them operational in the target environment; they take what they need from each purported mechanism, so to say, and integrate it into their mechanism design. What this procedure looks like and what role different kinds of knowledge – theory, experiments, background conditions – play will become clear in the discussion of the two case studies.

1.3.2. STRUCTURE-ALTERING INTERVENTIONS

At this point, let us shortly discuss one of the main criticisms that have been raised against mechanism-based extrapolation. As mentioned before, Daniel Steel is, on average, rather pessimistic about the potential of mechanisms in extrapolating experimental results. His main point of critique has to do with the structure-altering nature of interventions, which is closely related to the famous Lucas Critique (1983) in economic methodology. Lucas basically argues that many policy interventions will prove ineffective because they tend to change the institutional structure upon which the initial causal relationships are based. With respect to econometric policy interventions, he claims that:

> "Given that the structure of an econometric model consists of optimal decision rules of economic agents, and that optimal decision rules vary systematically with changes in the structure of series relevant to the decision maker, it follows that any change in policy will systematically alter the structure of econometric models." (ibid.: 279).

According to Steel, since mechanisms are causal structures that operate on different levels of abstraction, including the macro or institutional level (see figure 3, for instance), policy interventions that draw upon knowledge of mechanisms often result in changing the initial causal relationships. In other words, interventions by way of mechanism-based extrapolation will lead to a significant change in the conditions that had contributed to the (internal) validity of the causal

relationships in the first place. Yet what exactly makes these causal relationships change in response to interventions by policy makers?

This question can be answered by looking at the concept of modularity, which requires interventions upon one component of the mechanism to leave the other components unaltered. That is, if an intervention is modular then it only affects that part of the mechanism for which it was intended; an intervention is said to violate modularity if it alters other (unintended) parts of the mechanism as well. This interpretation by Steel of what interventions are supposed to do and what not, sounds rather idealistic. Indeed, this type of intervention has been aptly referred to as 'ideal interventions' by Steel, as well as by other scholars using similar connotations.[26]

Originally developed by James Woodward (2003: 98–99), an ideal intervention is formally defined according to the following five conditions:

1. [Intervention] *I* causes [variable] *X*;
2. *I* acts as a switch for all other variables that cause *X*;
3. Any directed path from *I* to [the effect] *Y* goes through *X*;
4. *I* is (statistically) independent of any variable *Z* that causes *Y* and that is on a directed path that does not go through *X*;
5. *I* does not alter the relationship between *Y* and any of its causes *Z* that are not on any directed path (should such a path exist) from *X* to *Y*.

Taken together, the conditions for ideal interventions are very demanding, and for good reason: policy makers can only know whether their interventions are actually effective if they do not alter the causal relationships they wish to exploit. Most of all, policy makers need a stable connection between the variable they manipulate on the one hand, and the variable they intend to influence on the other. Since the stability of this connection depends on the background conditions that are present, interventions that change these conditions tend to make the causal relationship undone. Therefore, mechanism-based extrapolation becomes somewhat self-defeating and, consequently, the problem of external validity remains.

Given that modularity is a serious issue for policy makers, it is hard to see how mechanisms can be used for the task of extrapolation. It seems that the only way to solve the problem of external validity is to conduct ideal interventions, i.e. interventions that do not violate modularity. The question is then: how can policy makers make sure their interventions are modular? More importantly for us would be the question what role, if at all, mechanisms play in this task. Without going into detail here, the issue depends on one's interpretation of what

[26] Julian Reiss, for instance, speaks of "hypothetical interventions" having a number of very idealised properties (2008: 162).

mechanisms are supposed to do. Viewed from the mechanism-design interpretation put forward by Guala, mechanism-based extrapolation is only one part of the story. In this sense, mechanisms are studied not merely for extrapolating some theoretical or experimental result but primarily to provide useful input for the design of a new, more appropriate mechanism. Here the effectiveness of an intervention largely relies on how this mechanism is designed, in which the issue of external validity – and thus the issue of modularity – might be less pressing.

1.4. ALTERNATIVE PERSPECTIVE: EVIDENTIAL RELEVANCE

So far, the potential of mechanisms for policy-making has been discussed in light of the problem of external validity. More precisely, the argument that mechanisms can support extrapolation has been critically assessed. Despite numerous publications that have contributed to this debate, whether mechanisms truly solve the problem of external validity remains a highly contested issue within the philosophy of social science. For this reason, it is fruitful to also look at the potential of mechanisms for economic policy from another perspective.

This section will introduce the second argument in defence of the usefulness of mechanisms for policy purposes. This argument has to do with the ability of mechanisms to provide a preliminary understanding of the evidence that could be relevant for the effectiveness of policy interventions. Framed in this way, the argument is somewhat more general in that it does not respond to one specific problem (that of external validity) directly. Though it should be seen as complementary to the previous argument about external validity, whereby this second argument addresses an important issue that has been relatively ignored in the debate about mechanisms within the realm of policy.

So what is this neglected issue and why is it important? Fundamentally, it has to do with the relevance of causal knowledge for the purposes of policy-making. In the process of establishing whether some causal relationship is relevant for a particular policy hypothesis, the economic theorist or experimenter adopts the perspective that takes his theoretical or experimental result as point of departure. Put differently, the theorist or experimenter tries to answer the question 'in which contexts are my obtained results relevant?'. While this perspective is of course completely legitimate, it differs from that of policy makers in a crucial way. They basically adopt the opposite perspective, answering the question 'what kind of evidence is relevant for my policy hypothesis?'. These two perspectives might appear rather similar at first sight, but it will become clear that their differences have important consequences for our assessment of mechanisms.

1.4.1. FFROM EFFICACY TO EFFECTIVENESS

To make sense of these different perspectives it helps to become familiar with an important methodological distinction. Nancy Cartwright has made many valuable contributions to the topic of evidence-based policy, where she has both advocated the usefulness of mechanisms for policy makers as well as expressing concern about the true potential of mechanisms in supporting evidential relevance.[27] Above all, she makes one particularly interesting conceptual distinction in this regard, namely *efficacy* as opposed to *effectiveness*, where efficacy is "the ability of a treatment to produce benefit if applied ideally" and effectiveness is "the benefit that actually occurs when a treatment is used in practice" (Andrews, 1999; cited in Cartwright 2009c: 187–188).

The first concept, efficacy, is similar to that of internal validity: it indicates whether a causal relationship obtained in an experimental setting is genuine. Or more precisely, efficacy is established when C causes E if the production of E is attributed to C among the population in X. One prominent method for gathering evidence for efficacy is a randomised controlled trial (RCT). Originally based on John Stuart Mill's method of difference (1843), RCTs engage in causal inference by dividing the experimental population into two groups: the treatment group and the control group. The treatment group is actually exposed to the causal variable that is presumed to make a significant difference to the phenomenon under investigation.

In contrast, the control group is, unknowingly, not exposed to the same causal influence. The reason for having two separate groups is that an experimental setup like this will mitigate the problem of confounders. Since composition of both groups is supposed to be random, this means that "the distribution of causal factors other than the one in question between the two groups is (near enough) identical" (Cartwright & Hardie, 2012: 33). That is, the randomization element implies that the influence of confounding variables on the causal relationship of interest is evenly distributed and therefore rendered insignificant. As a result, RCTs have been successfully conducted in a wide range of scientific domains, most notably medicine.[28] Disregarding the legitimate concern that RCTs might not be completely able to resolve the problem of confounders, they are generally considered to be a suitable method for establishing efficacy.

---

[27] For instance, in her book *Evidence-based Policy* (2012) Cartwright, together with Jeremy Hardie, extensively deals with the problem of external validity specifically for the purposes of policy makers. It is important to note, however, that she does not explicitly refer to 'mechanisms' in most of her work. Nevertheless, her discussion of issues related to external validity – such as efficacy versus effectiveness – can to a large extent be used for the assessment of mechanisms in different policy contexts.

[28] Clarke et al. (2014) discuss the many proclaimed successes, and consequent popularity, of RCTs within the field of evidence-based medicine, which is defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" (ibid.: 339). They challenge the high position of RCTs in the general hierarchy of evidence by claiming that mechanistic evidence should be treated as complementary to the evidence obtained by RCTs (and other types of statistical trials, such as cohort studies, case series, etc.).

RCTs have also been successfully conducted in economics, as the example of the Tennessee class-size reduction programme has shown.[29] In short, the 1985 STAR project, which was designed as an RCT, concluded that pupils in smaller classes performed better than those in relatively larger ones. The RCT was able to confirm the positive causal relationship between student performance and class size – albeit in only one particular context. This conclusion, especially with respect to minority children who benefitted most from the reduction in class size, fitted nicely with the opinions of politicians and other policy makers. Surely this evidence for efficacy could be used for similar policies elsewhere, or so they thought.

Unfortunately, this has proved not to be the case. The results that policy makers had hoped for stayed out when a similar class-size reduction programme was introduced some years later in California. Cartwright (2009b) lists two plausible explanations for this failure. First, the implementation of the programmes was different in each case. The programme in California was rolled out in a short period of time, which created a sudden demand for additional teachers and classrooms. Consequently, large numbers of poorly qualified teachers were hired and many classes were organized in rooms that were inappropriate as a learning environment. Secondly, the distribution of confounding factors is likely to have been different in California as compared to Tennessee. It is plausible that parents who actively engage with their children's educational development – by practicing reading comprehension at home, for instance – are inclined to send them to schools that have smaller classes. This way, a common cause existed between class size and reading performance in the case of California; a condition that was not taken up in the original RCT conducted in Tennessee.

This example shows, much like the case of the TINP and the BINP, that establishing efficacy is simply not enough if policy makers want to exploit its results in other contexts. What is required in addition is the extrapolation of the experimental results to seemingly similar situations, but with often very different populations and background conditions. Indeed, this is precisely the issue of external validity: how do policy makers make sure that efficacy (or internal validity) also holds outside the experimental setting? As the previous section (1.3) has already shown, mechanism-based extrapolation could be one useful tool for this aim, as mechanisms specify the necessary background conditions for the intervention to be effective. Understood in this sense, mechanisms can be used to move from efficacy to effectiveness; from 'it worked there' to 'it will work here'.

[29] Again, this example is borrowed from Cartwright and Hardie (2012).

1.4.2. CAUSAL SCENARIOS

However, the crucial aspect that is missing here has to do with the relevance of the evidence available to policy makers. Evidence for efficacy, as established by RCTs, can be valuable to the experimenter and other scholars in his particular field, but it is not necessarily relevant for policy makers who are often concerned with different hypotheses. According to Cartwright, besides the requirement of evidence for a policy hypothesis to be credible – i.e. evidence that is likely to be true – it also needs to be relevant. It is not sufficient for policy makers to know that evidence for a certain efficacy claim is true in some experimental setting; rather, in addition, they require that evidence to be relevant for the policy aim at hand.

Fair enough, but how do policy makers know whether a particular piece of evidence for efficacy is actually relevant? A more appropriate formulation of this question would be: what evidence (for efficacy) is relevant for the policy hypothesis? This way, the usual perspective of moving from efficacy to effectiveness is switched around: policy makers do not start with the extrapolation of experimental results but instead focus on the criteria for a policy's effectiveness. To do this, Cartwright proposes that policy makers "begin to construct a variety of different causal scenarios [about what will happen when an intervention is conducted], some more plausible or more probable than others" (Cartwright, 2009b: 135). These scenarios would start with the intervention and end with its preferred outcome, based on several intermediate causal steps. For example, when the outcome is improved reading performance by Californian students, one plausible scenario could indeed involve a class-size reduction programme that has reading performance as its ultimate effect (see figure 5). This scenario assumes class size to influence how much personal attention students receive from their teachers, which, together with the intrinsic ability of students, determine reading performance. In this case, reducing class size seems to be an effective intervention in order to improve students' reading performance.

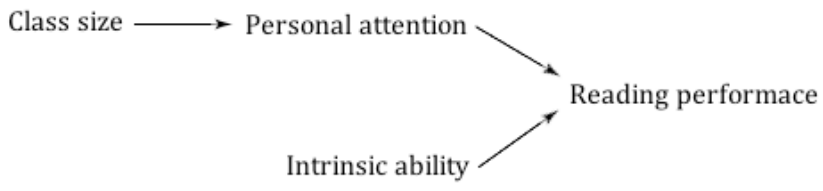

Figure 5: Sketch of one plausible mechanism between personal attention and intrinsic ability (as causes) on the one hand, and reading performance (as the effect) on the other. In this scenario, intervening on the average class size of students would presumably lead – via more personal attention – to an increase in reading performance.

There could be, of course, many other plausible scenarios that policy makers might take into account when they deliberate about potential interventions. For instance, introducing

mandatory reading assignments throughout the school year might enhance the intrinsic learning abilities of students, which could in turn lead to better reading performance (see figure 6). As mentioned already, the problem for policy makers is not the difficulty in constructing plausible mechanisms, but that there are often (too) many different plausible mechanisms available – all relevant, to some extent, to the particular policy aim. Evaluating the plausibility of mechanisms is not just about yes-or-no questions, but about *how* plausible certain mechanisms actually are. Answering these latter kinds of questions largely depends on the context in which policy makers operate, as the case studies in the following chapters will illustrate.

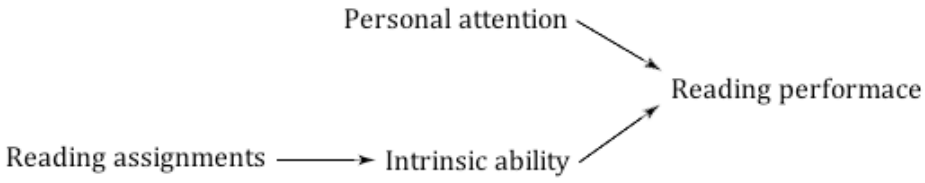

Figure 6: Sketch of an alternative mechanism between personal attention and intrinsic ability (as causes) on the one hand, and reading performance (as the effect) on the other. In this scenario, intervening on the amount of reading assignments might also lead – via enhanced intrinsic ability – to increased reading performance.

For now, suppose policy makers have identified a mechanism as in figure 5, where they have some reason to believe that reducing class size will have a positive effect on reading performance.[30] This belief could be reasonably based, for instance, on popular opinion among parents and politicians. As a next step, policy makers could use the evidence for efficacy between class-size reduction and reading performance obtained in the STAR project in order to substantiate this belief. Without blindly relying on the results of the RCT conducted in Tennessee, policy makers are now able to account for different causal variables, as well as background conditions, that may have an effect on the target variable. While it is plausible that in this case reducing class size will enhance reading performance based on the RCT results, it might be just as plausible that introducing (extra) mandatory reading assignments will significantly contribute to the same outcome. Or there could exist some other mechanism with an even greater plausibility that has yet to be identified.

Nonetheless, the main advantage is that policy makers can develop their interventions more carefully, and therefore more effectively, due to their preliminary understanding of the evidence that *could* be relevant for the effectiveness of policy interventions. In this sense,

[30] It is important to note that in this example policy makers have identified this particular mechanism independent of any evidence for efficacy between class-size reductions and reading performance. Although the evidence for efficacy between the two variables could play a role later on in the policy-making process, it does not have to play a leading role (such as in the Tennessee-California example).

mechanisms show policy makers which variables and background conditions are likely to be relevant. Whether the evidence actually *is* relevant has to determined empirically. Thus, within some specific context, mechanisms are able to inform policy makers about the relevance of different kinds of evidence – including, but not restricted to, evidence for efficacy.

1.4.3. A NOTE ON INTERPRETING MECHANISMS

To round up the discussion for now, let me emphasize the main differences in interpreting the use of mechanisms for policy-making. Roughly speaking, there are three such interpretations: mechanism-based extrapolation, mechanism design, and mechanisms as causal scenarios. The first two partly overlap while the third one should be seen as a stand-alone alternative.

With respect to mechanism-based extrapolation, it is claimed that mechanisms can support the extrapolation of experimental results to other contexts because they show *how* causal relationships operate across different contexts. More precisely, mechanisms are able to indicate the similarity in background conditions between the artificial and target environments, which favours extrapolation in case the similarities are strong. This way, mechanisms can help determine whether an intervention is likely to be successful or not.

Guala (2005) endorses mechanism-based extrapolation but adds another dimension to the use of mechanisms, namely the *design* of mechanisms. In this sense, policy makers still aim to extrapolate theoretical and experimental results to non-artificial environments by invoking mechanisms. Yet mechanism-based extrapolation is now used as a means for the design of a new mechanism, one that enhances the objectives of interventions as best as possible. In short, (purported) mechanisms are used to provide theoretical, experimental, and background knowledge that can be integrated into a new, specifically designed mechanism.

The interpretation of mechanisms as causal scenarios, however, is based on a completely different line of reasoning. Whereas mechanism-based extrapolation and mechanism design try to move from efficacy to effectiveness, mechanisms as causal scenarios adopt the opposite perspective. That is, policy makers take the aim of an intervention as point of departure and see what kind of evidence could be relevant to this aim. The evidence might include efficacy claims obtained in some other (non-target) context, but this is often only a small fraction of the total evidence available. To acquire a preliminary understanding of the evidence that could be relevant, policy makers construct different causal scenarios that plausibly connect the intervention with the preferred outcome. This way, their approach is not one of 'what worked *there* will also (hopefully) work *here*' but rather 'what will work *here*, given the policy objectives?'.

CHAPTER TWO

MECHANISMS AND AUCTION POLICY

When discussing the potential of mechanisms for economic policy, at some point one has to transfer the philosophical concepts, arguments, problems and solutions to more practical contexts. If we wish to understand *why* mechanisms can be a useful tool for policy-making, we need to see *how* mechanisms actually operate in different policy domains. In this chapter, the use of mechanisms in one such domain will be illustrated, namely the domain of auction policy. As mentioned in the introduction, this chapter will deal with one specific case study of real auctions. The reason for this particular case study is threefold: first, the auctions involved a multibillion-dollar business, which not only had a significant impact on public finances but also on the administration's economic performance. More importantly, the auctions can be characterised as a kind of *economic engineering* where knowledge of mechanisms is used to design a new mechanism that works well in the policy context. In addition, the philosophical literature has extensively engaged in the debate around auction design and its policy implications over the past few decades, which provides ample material to advance our assessment of mechanisms for policy-making.

Before we delve into the details of the case study related to auction policy, the approach to this case study (and of the next in chapter three) must be clear. The following case study will be analysed according to the two perspectives – external validity and evidential relevance – as introduced in the previous chapter. For instance, how did mechanism-based extrapolation in this case supposedly solve the problem of external validity? What kind of role did theories, experiments and background conditions play in the design of the auctions? Moreover, in what way can mechanisms establish evidential relevance for the effectiveness of the auctions? Which criteria were used in this process, and how were these determined? From these two different perspectives, the case study will be able to support the claim, or so I will argue, that mechanisms can be a useful tool for policy-making.

2.1. THE FCC AUCTIONS

One of the most celebrated applications of game theory and experimental economics to a real-world phenomenon was a set of auctions conducted by the Federal Communications Commission

(FCC), an agency of the United States government, in 1994.[31] The stage for this success was already set in the 1980's when a wave of decentralization hit the US economy. A new economic and political paradigm was established that replaced centralized, bureaucratic systems of allocation with more market-driven processes. The old way of allocating licenses to use radio spectrum was done via administrative hearings, in which each applicant had to convince the FCC of their (public) interest in obtaining the licenses. This process was slow, nontransparant and inefficient because it allocated the licenses for free instead of selling them for their market value. After trying out different alternatives, such as lotteries, the FCC turned to game theorists and experimental economists to help design an efficient set of auctions.

A licence for wireless Personal Communication Systems (PCS) provides the right to use some portion of the spectrum for radio communication (telephones, faxing machines, etc.). The primary aim of the FCC auctions was to achieve an efficient allocation of the licenses, which meant selling the licenses to companies who valued them the most, thereby generating substantial government revenue. Additionally, monopolies had to be prevented while relatively small companies, and minority-owned and women-owned companies were to be promoted. Although these additional aims were controversial, and their success therefore rather ambiguous[32], the FCC auctions clearly succeeded in achieving its main objective: a series of seven auctions between 1994 and 1996 attracted many bidders, allocated thousands of licences and raised more than $20 billion in government revenues (Cramton, 1998).

The designers of the FCC auctions tried to use scientific knowledge for a specific policy aim. The team included game theorists, experimentalists, software engineers, lawyers and policy makers who all worked together to create a procedure that could make the theory and experiments work well in their target context. One way they did this was to use 'testbed' experiments that identified important details about the behaviour of participants. Charles Plott, a prominent Caltech experimentalist, defines this kind of experiment as follows:

> "An experimental 'testbed' is a simple working prototype of a process that is going to be employed in a complex environment. The creation of the prototype and the study of its operation provides a joining of theory, observation, and the practical aspects of implementation, in order to create something that works." (1996: 1)

[31] Henceforth, I will refer to this particular set of auctions as the 'FCC auctions'.

[32] For instance, the initial aims to promote minority-owned and women-owned companies (positive discrimination) in the telecommunications industry were withdrawn by order of the Supreme Court in 1995 (McAfee & McMillan, 1996: 167). Nik-Khah (2008) argues that in fact the auctions' only success was large government revenue.

For instance, a group of trained students was hired to keep track of the software problems (bugs) that arose during small-scale experiments. They focused on detecting the effects of minor deviations from the rules that were included in the design of the auctions, such as investigating what happens if participants log in from multiple locations at the same time. It turned out that such apparently small variations in the auction design could have a significant impact on the actual outcome when implemented in the target environment.

The actual FCC auctions were therefore designed according to theoretical insights – especially from game theory – experimental evidence, and a substantial amount of background knowledge. The procedure of mechanism design effectively integrates these different epistemic sources in order to construct a new mechanism that is able to bring about the preferred outcome. In case of the FCC auctions, policy makers started with abstract game theoretical knowledge and proceeded to a more concrete design by incorporating the results of specifically conducted experiments and by studying some important background conditions of their target context. This so-called 'mechanism view' is quite unique because it differs from the more traditional view of scientific knowledge, in which spontaneously occurring phenomena are explained by developing some kind of theory. While economists normally move from the study of phenomena to the construction of theoretic models, policy makers are more inclined to proceed in the opposite direction (Guala, 2005).

2.2. THE NEED FOR A FULLER PICTURE

In order to substantiate the point that policy makers tend to study phenomena as a whole, this section will show how they did so in the case of the FCC auctions and in what sense this process differed from the more traditional view in economics. For this purpose, I will draw on the work of Anna Alexandrova (2006), who argues that policy makers do not just study causal relationships in isolation but instead inquire about phenomena taken as a whole. By engaging in this particular discussion, the subsequent arguments relating to external validity and evidential relevance will become more convincing.

So in what sense did the policy makers who were involved in the FCC auctions study the phenomenon of spectrum auctions 'as a whole'? When criticising the traditional view of causality employed in economics – referred to as the method of isolation or the analytic method, which originates in John Stuart Mill's account of causal tendencies (1843) – Alexandrova distinguishes between *derivation facilitators* and *situation definers*. Both types of assumptions are necessary to establish some general causal relationship. On the one hand, derivation facilitators are

assumptions that enable a certain deduction to go through. They represent the abstract aspects of a theoretic model that facilitate (mathematical) derivation.

One example of a derivation facilitator Alexandrova mentions is the assumption that utility functions should be twice-differentiable in the game theoretic model of a private value auction with two bidders. Without this assumption the model cannot predict the valuations of the bids in equilibrium, where both bidders should bid half their valuation.[33] The general result that bids tend to be lower under a first-price rule (i.e. each bidder only submits once) and sealed bidding (i.e. each bidder does not know the other's valuation) proved to be useful for policy makers. Most importantly, it showed them that a first-price sealed-bid auction would lead to bids below true valuation, which would mean lower government revenues if this type of auction was implemented. Since government revenue was one of the primary aims of the FCC, policy makers subsequently rejected this type of auction framework.

Situation definers, according to Alexandrova, "set the structure of the situation under consideration" (2006: 181). That is, they depict the context of the phenomenon under investigation so as to render the actual operation of causal relationships intelligible. These empirical features tell the policy maker where and when certain causal relationships will hold. With respect to the game theoretical model of first-price sealed-bid auctions, situation definers specify the bidding rules, the number of bidders, and the nature of their reasoning. These assumptions need not be critical to obtain the result – since the same result will be obtained when there are more than two bidders, for instance – but they may nevertheless be important for policy makers to make sense of a derivation. For without situation definers, the derivation facilitators in a game theoretical model "do not tell us how to translate the statements about entities in a model constrained by mathematics into statements about naturally occurring entities and properties in the world" (ibid.: 183).

Despite the fact that both derivation facilitators and situation definers are necessary when studying causal relationships in isolation, the presence of the former becomes problematic in policy contexts. The reason is that derivation facilitators are often highly abstract assumptions that simply do not apply to the target environments in which policy makers are interested. Even though some derivation facilitators might sound like entities that exist in the real world (like utility functions), policy makers cannot treat them as such because they do not know whether the assumptions actually hold in practice. For example, due to the mere presence of derivation facilitators, models "can fail to describe even one context in which [a] first-price auction leads to bids below true valuation" (ibid.: 182). Thus, game theoretical models, which often include a high degree of derivation facilitators, only played a limited role in the design of the FCC auctions.

[33] For a more detailed exposition of this particular game theoretic model that was used in the design of the FCC auctions, see Alexandrova (2006: 174–175).

So, yes, policy makers made use of game theory when designing the FCC auctions but only as a body of knowledge suggesting general issues to be aware of. They did not study game theory to detect any causal relationships that could be used directly in their auction design, since no such theories existed for the specific context of the spectrum auctions. Rather, they treated theory as a good way to start their auction design, without hoping to find one kind of theoretical framework that would apply perfectly. The challenge for policy makers was to figure out which theories could be useful for the task at hand. This mainly depended on the aim of the FCC auctions, namely to generate a substantial amount of government revenue. From this perspective, the theoretical result of a first-price sealed-bid auction leading to relatively low bidding appeared to be relevant.

However, the relevance of this particular theory was limited due to its ambiguous results with regard to one empirical feature of the FCC auctions: the existence of complementarities. The problem was that the spectrum auctions featured strong complementarities – i.e. the value of one license depended on what other licenses a participant owned – which made the theoretical result obsolete. Put differently, given the existence of complementarities, the causal relationship between first-price sealed-bid auctions and low bidding behaviour could no longer be relied on. One could assume, for instance, that bids for licenses with complementary value will be higher than for those without. However, open auctions in the presence of complementarities may also reintroduce the so called 'winner's curse', which is the tendency for the winning bid in a sealed-bid auction to be the one that most overestimates the true value of an object. The point was that policy makers simply did not know how bidders would behave when an auction involves complementarities.

In response to these problems, policy makers did no longer adhere to the method of isolation; instead, they studied the phenomenon of spectrum auctions taken as a whole. This alternative method is more sensitive to the empirical features of a given context. According to Alexandrova:

> "The method [in the FCC case] was different from the method of isolation because rather than hunting after knowledge about the behaviour of tendencies in isolation from others, the auction designers instead sought to find out facts about one material system as a whole. They remained agnostic about all but the most general features of the tendencies at work. 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." (2006: 186)

By studying the details of a given phenomenon, policy makers tried to identify at least one concrete environment where the design would satisfy their aims. Besides using game theory in a preliminary sense, they mainly relied on local, empirical knowledge of the target context. Experimental testbeds functioned as a particularly important source of knowledge in this process because these kinds of experiments showed to what extent causal relationships derived from theory could actually be realized. For instance, the presence of complementarities proved to be a significant hurdle in the design of the auctions and could only be solved by running a set of experiments that incorporated this specific condition. With the help of these results, the actual auctions conducted by the FCC included rules that forced participants to continue bidding in order to maintain their eligibility. These rules effectively prevented a process of slow and cautious bidding, which enhanced the efficiency of the auctions.

2.3. EXTERNAL VALIDITY

After having clarified that policy makers studied the phenomenon of spectrum auctions as a whole, let us turn to the problem of external validity. This section will be concerned with two aspects of external validity with respect to the FCC auctions: first, I will argue that mechanisms played an important role in the design and implementation of the auctions; and second, I will argue that mechanism design, though susceptible to structure-altering interventions, can be a suitable extension of mechanism-based extrapolation, and should thus be taken seriously by the economic policy-making community.

Recall that Guala considers the design of a mechanism as "an enterprise between theoretical and applied economics, which requires first stating clearly the goals to be achieved by the mechanism, and then finding the best means to achieve them given the circumstances" (2005: 164). The first and most important step in the process of designing mechanisms is determining the aim of the intervention. That is, policy makers need to be clear about what the intervention is supposed to achieve. As for the FCC auctions, the primary aim was to raise substantial government revenue, as well as some additional objectives such as the prevention of monopolies and the promotion of minority- and women-owned companies.

The next step is to construct the most effective mechanism for the task at hand, which is where the real work begins. Here policy makers make use of three distinct sources of knowledge: theory, experiments and background conditions.[34] The value of mechanism design is that policy

[34] Admittedly, in some cases these sources of knowledge overlap: an experiment might make use of certain background conditions, or a piece of theory may be tested by way of experiments. Nevertheless, to see the value of mechanism design for policy-making it is important to distinguish between these different kinds of knowledge.

makers are able to *integrate* these different kinds of knowledge into concrete, effective interventions. Mechanism design is therefore a complex, context-specific procedure that outlines how an intervention affects the phenomenon under investigation. Essentially, mechanism design makes use of many different kinds of knowledge, including hypotheses about mechanisms that purportedly exist in the real world and could potentially be used for extrapolation. These hypotheses form the initial building blocks of the mechanism design, in which some parts of the (purported) mechanisms are used while others are not: bits of theory, experimental results and background conditions are carefully selected and modified to fit the target environment. Eventually, a new mechanism is designed – based on the study of many different mechanisms – to make the intervention work effectively. Let us now see how this procedure actually took place in case of the FCC auctions.

2.3.1. MECHANISM DESIGN

First up is the role of theory in the design of an appropriate mechanism. In highly applied cases such as the FCC auctions, the previous section has already shown that theory played a relatively minor role because it often proved to be incomplete in the target context. In other words, theoretical knowledge about the causal relationships operating in the specific context of the spectrum auctions was only partially available, at best. Lacking a comprehensive theory of these causal processes, policy makers had to rely on "piecemeal theoretical insights" (Guala, 2005: 169) when designing the FCC auctions. The example of the first-price sealed-bid auctions proves this point, since this particular theoretical model could only be used as a rough approximation of what actually happens in real-world situations. Thus, to put it philosophically, theory is a necessary condition for the design of mechanisms but it is hardly a sufficient one.

Apart from drawing on abstract theory, policy makers also gained a great deal of understanding from conducting experiments. In fact, the upshot of Guala's discussion of the FCC auctions is that experiments were in large part responsible for its success. In particular, the introduction of testbed experiments had led to several significant results, ranging from the efficiency of different auction designs to the likely consequences of software bugs. Testbed experiments were used in several ways throughout the FCC's auction design procedure.

First, they were instrumental in choosing one auction design over another with respect to their fundamental properties. As with the comparative test between a simultaneous ascending auction and a combinatorial sealed-bid auction, the former came out on top because the experimenters noticed that the implementation of the latter raised too many practical

complications. These initial tests provided policy makers with the first operational details and potential problems of the different auction designs.

Secondly, testbed experiments "were used to transform the abstract design into a real process that could perform the required task reliably in the environment in which the auction was to be implemented" (ibid.: 175). In this sense, the experiments can be seen as some sort of *economic engineering*: rather than testing some existing theoretical model of auctions, the testbed experiments instead supported the development of completely new types of auctions. The issue was not about discarding one theory or another, but whether the concrete properties of a given auction design would actually contribute to the aims of policy makers. In order to determine which auction design performed best, policy makers had to discover new causal processes or sometimes construct them from scratch – hence the notion of engineering. For example, since no game theoretical model of a combinatorial sealed-bid auction existed, policy makers had to rely on experiments in order to see whether this type of auction would serve their aims. Even if there would have been some model that (in theory) produced desirable results, it could have been very difficult to make it operational due to practical constraints such as time, money, etc. In fact, lack of time was one of the reasons why policy makers did not choose to implement a combinatorial sealed-bid auction design. Therefore, it is clear that policy makers benefitted from using testbed experiments in designing the FCC auctions.

Thirdly, the experience gained during the experimental testbeds proved valuable for extrapolating the results to real-world situations. In October 1994, the team of experimentalists monitored a real auction to infer whether the favourable test results of the simultaneous ascending auction would also obtain outside the laboratory. A key element in this successful case of extrapolation was the strong similarity between the artificial setting of the testbed experiments and the actual environment of the FCC auctions. The testbed experiments had been constructed (or engineered) so as to mimic the empirical features of the real auctions as closely as possible: the same number of bidders and licenses, the existence of complementarities, an equal number of rounds, etc. This way, both situations (artificial and real) could be easily compared, and policy makers were able to conclude that a simultaneous ascending auction design best served their aims.

The last kind of knowledge that policy makers made use of in the design of the FCC auctions were background conditions. According to Guala, "the external validity step was based on a comparison between laboratory and real-world evidence, and stood on a stock of 'background knowledge' aimed at making the inference as strong as possible" (ibid.: 183). What made the extrapolation of the experimental testbed results to the actual auctions so successful was the amount of contextual information that was incorporated into the experiments. Knowledge of background conditions added a great deal of flesh to otherwise bare-boned auction designs, so to speak.

To emphasize the importance of background conditions in the case of the FCC auctions, let us turn to one of the details of the continuous ascending auction design: eligibility. A bidder is said to be eligible if she (1) has made an initial deposit based on the total number of licenses she wants to compete for and (2) her bid is either the standing high bid or is higher than the minimum bid allowed. The idea of eligibility was introduced to prevent the participation of bidders that were not genuinely interested in obtaining any spectrum licenses, and to regulate the duration of the auctions. Since the FCC was concerned about the costs that prolonged auctions would incur, it wanted to find a way to speed up the auction process. This particular background condition, along with many others, guided policy makers in the auction design and eventually contributed – through the introduction of eligibility rules – to its successful implementation.

The above discussion of the three kinds of knowledge used by policy makers has shown how mechanisms played an important role in the design and implementation of the FCC auctions. Before moving to a common critique against mechanism-based extrapolation, one feature of mechanism design needs to be emphasized at this point. Mechanism design relates to the problem of external validity in an intricate way: policy makers extrapolate certain theoretical and experimental results to a target context by *integrating* the most important parts of the theory, experiments and background conditions that are obtained through the study of mechanisms. Here extrapolation does not proceed by taking one particular piece of theory or experimental result and applying it to a target context (as in the original interpretation of mechanism-based extrapolation). Rather, it combines some basic theoretical insights with a set of specific experimental results and the most salient background conditions in order to construct an effective intervention. The point is that mechanism design, for it to be successful, is an inherently complex and context-dependent procedure. The careful and detailed design of the FCC auctions is a foremost example of how bits of game theory, experimental bidding behaviour, and background conditions of the spectrum auctions were used to serve the aims of policy makers.

2.3.2. MECHANISM DESIGN AND STRUCTURE-ALTERING INTERVENTIONS

The FCC auctions are frequently hailed as a success story of mechanism design, but can one really draw any general conclusions from this particular case? Since the procedure of mechanism design partly relies on mechanism-based extrapolation, we have to take the structure-altering nature of interventions into account here, too. As introduced in chapter one, this point of critique is raised by Steel (2008) who is rather sceptical about the idea of mechanism-based extrapolation within economics. More specifically, he argues that:

> "Causal relationships typically depend upon background conditions too numerous and complex to fully and explicitly incorporate into a model. Consequently, interventions that change such background conditions may be structure-altering. Even if the generalizations accurately described the causal relationship under the original set of background conditions, it might be an inaccurate representation of that relationship in the new circumstances brought about by the intervention." (ibid.: 155)

So, to say that an intervention is structure-altering means it changes the background conditions that support the causal relationship exhibited by the mechanism. As a result, the causal relationship policy makers wish to extrapolate may no longer operate due to the intervention. Modularity is a helpful concept here since it requires an intervention to only change that part of the mechanism for which it was intended. For if an intervention changes other parts as well, then policy makers cannot know for sure whether it was actually their intervention that had an effect (if at all) on the target variable. Therefore, the violation of modularity often results in the breakdown of causal relationships and renders interventions ineffective.

Unfortunately, the issue of structure-altering interventions cannot be ignored when discussing the potential of mechanism design within policy-making contexts. In fact, the design of the FCC auctions suffered from quite some interventions with significant side effects. One example relates to the eligibility condition in the continuous ascending auction design. The supposed problem here was that introducing eligibility did not just neatly affect the duration of the auction without having other unintended consequences, too. Whereas there was no theory that could inform policy makers about how long a continuous ascending auction would go on for, they turned to experiments that tested the effect of different background conditions on the duration of the auctions. In addition to the eligibility condition, policy makers also imposed large bid increments so as to identify the winning bidder quickly. Yet the experimenters "observed that big increments sometimes eliminated bidders too quickly, causing their eligibility to drop and therefore creating a 'demand killing' effect" (Plott, 1997: 633). It turned out that intervening upon a continuous ascending auction by imposing large bid increments indeed shortened its duration, but it also impeded the eligibility condition from operating properly.

Interestingly, this example illustrates a severe weakness as well was a potential strength in the mechanism-design account put forward by Guala (2005, 2011). On the one hand, interventions upon background conditions often lead to ambiguous results because of interaction effects: the eligibility condition and the bid increment condition, operating at the same time, had a different effect on the efficiency of the auctions than when applied separately. More worryingly, the introduction of large bid increments nullified the positive effects of the eligibility condition. These kinds of interventions may raise more questions than answers: given the interaction effects,

which condition is more important for increasing government revenue – eligibility or large bid increments? More generally, are there any other background conditions that increase government revenue, but do so unambiguously?

On the other hand, the question whether interventions are susceptible to interaction effects can only be answered by conducting (more) testbed experiments. For instance, the interaction effect between the eligibility and increment conditions could, according to Plott, only be made sense of by further empirical testing: "The complex ways the rules [background conditions] interact, and the presence of ambiguities, do not become evident until one tries to actually implement the rules in an operational environment" (1997: 628). In response to this particular interaction effect, the experimenters made some additional adjustments to the continuous ascending auction design, such as longer or more frequent bidding rounds. Eventually, they concluded that more frequent rounds did not significantly affect the efficiency of the auction.

Of course, there will always be a risk that structure-altering interventions will do more harm than good to the design of a mechanism. This point seems hard to deny. Interaction effects might be resolved by engaging in (even) more sophisticated experiments but this is by no means guaranteed. Besides, this strategy is likely to be a lengthy and costly endeavour for policy makers.

Nevertheless, the FCC auctions have proved the advantage of mechanism design over mechanism-based extrapolation within the domain of auction policy. On a charitable interpretation, mechanism design is able to account for the most problematic consequences of structure-altering interventions due of its clever use of experiments. To be fair, this interpretation is not too far off from Steel's general position on mechanism-based extrapolation, in which he sees "no reason *in principle* that mechanism-based extrapolation cannot be successfully utilized in social science" (2008: 150, italics added). Perhaps the procedure of mechanism design is a suitable extension of mechanism-based extrapolation, where the background conditions of the target environment are first imported into the experimental setup before the (experimental) results are extrapolated back to the real world. Guala's advice on economic engineering – "if you cannot export some laboratory conditions into the real world, you had better make sure that the relevant aspects of the real world are imported into the lab" (2005: 189) – may prove useful for policy makers in other policy domains, too.

2.4. EVIDENTIAL RELEVANCE

Until now, the FCC auctions have been discussed with respect to the problem of external validity. Essentially, this problem can be seen as policy makers trying to move from efficacy to effectiveness: given the fact that a causal relationship holds in one place, how do we make sure it also works where we want it to? In this regard, I have argued that, first, mechanisms played an important role in the design and implementation of the FCC auctions; and second, that mechanism design, though susceptible to structure-altering interventions, can be a suitable extension of mechanism-based extrapolation, and should thus be taken seriously by the economic policy-making community.

However, moving from efficacy in artificial environments to effectiveness in policy contexts is to adopt the wrong perspective. As was already remarked in chapter one, the crucial aspect that is missing here has to do with evidential relevance. That is, by taking some efficacious result as point of departure in their inquiry, policy makers tend to overlook other kinds of evidence that could be relevant to their aims. For this reason, policy makers should be less concerned with whether a particular efficacious result is also operative in the target context (basically the issue of external validity). Instead, they should focus on what kind of evidence – which may include but is not restricted to evidence for efficacy – is actually relevant to their policy hypotheses. This alternative perspective involves a wholly different use of mechanisms, namely as causal scenarios showing the intermediate processes between an intervention and its desired outcome.

In this section, I will argue for three things. First, I will show that policy makers involved in the FCC auctions adopted the right perspective as they focused on establishing evidential relevance. Second, they proceeded in this task by constructing causal scenarios that provided a preliminary understanding of the evidence that could be relevant for the effectiveness of policy interventions. Third, since the move from efficacy to effectiveness is reversed, the issues related to external validity appear to be less pressing from a policy point of view.

2.4.1. A FOCUS ON RELEVANCE

As discussed earlier in this chapter (section 2.2.), policy makers involved in the FCC auctions did not merely study causal tendencies in isolation, but instead studied the phenomenon of spectrum auctions as a whole. More specifically, they did not only use insights from game theory but also drew upon a range of experimental results and background knowledge for their auction design. For instance, the model of a first-price sealed-bid auction did not fit well with the existence of complementarities in the real-world spectrum auctions. The precise effect of complementarities

could only be learned through the use of experimental testbeds, which also tested the effects of other background conditions and their interactions.

In this sense, policy makers treated evidence for efficacy as only one of the (many) kinds of evidence that might be relevant to their interventions. Their primary aim was to bring about an efficient allocation of spectrum licenses and to raise substantial government revenue, which featured as their point of departure. That is, every piece of evidence – either a game theoretical model, some experimental result or a salient background condition – was evaluated in light of whether it contributed to the efficiency of the auctions and the US treasury. So what mattered most was the relevance of the evidence with respect to the specific aims of policy makers; the issue of whether the evidence was also credible had to be taken up after its relevance was established.

Right from the start, the FCC informed the group of game theorists, experimentalists, software engineers and lawyers about its objectives. These objectives acted as constraints for the kinds of evidence that were to be gathered during the auction design procedure. In principle, any evidence that was produced in this procedure was allowed *as long as it supported the aims of the FCC*. For example, the first-bid sealed-price auction model was brought in as evidence for its contribution to the winner's curse. This particular piece of theoretical evidence was only relevant to the FCC because it suggested that closed auctions would reduce government revenue. Due to this evidence, policy makers knew a closed auction would not serve their aims well and opted for an open framework instead.[35] From a purely scientific point of view, it is clearly valuable to know that a causal relationship between closed auctions and the winner's curse exists. For policy makers, though, this knowledge also needs be relevant to their specific policy aims, which in this case it was.

2.4.2. CAUSAL SCENARIOS

How then, could policy makers inquire about the relevance of different kinds of evidence? This is where the interpretation of mechanisms as causal scenarios comes in: mechanism provide a preliminary understanding of the causal relationships that could be relevant to the policy aims at hand. Policy makers draw on different kinds of theoretical, experimental, and background knowledge in constructing these causal scenarios. At least, the scenarios need to be plausible

[35] Anna Alexandrova and Robert Northcott (2009) note that the decision by policy makers to choose an open auction framework for their final design required "a judgment call rather than a neat theoretical demonstration" (ibid.: 317). This does not only show that the ultimate auction design included many different kinds of evidence, but also that mechanisms can sometimes merely indicate that some piece of evidence is *not* likely to be relevant. Nevertheless, this limited role of mechanisms might still produce useful insights for policy makers since they are able to eliminate irrelevant evidence.

according to some basic theory, general principle, or widely-held public opinion. Each scenario starts with the proposed intervention and ends with its effect on the preferred outcome.

In the FCC case, we have seen that policy makers took the aims of the spectrum auctions – i.e. enhancing the efficiency of the spectrum auctions and increasing government revenue – as their point of departure. This formed the outcome of the different causal scenarios; the end, so to say, of the intermediate causal processes that are assumed to operate between the intervention and its effect upon the target variable. In this sense, causal scenarios address the question 'what happens to outcome *Y* if we do intervention *I* on variable *X* in target context *T*?' The possible answers to this question can be numerous and complex, but they are guided by the specific aims of the intervention.

To illustrate how mechanisms were used as causal scenarios during the design of the FCC auctions, let us return to the example of open and closed auctions in the presence of complementarities. As explained earlier, policy makers could not rely on theory alone to guide their decision on the matter. To proceed, they conducted a range of experiments that tested the effects of complementarities upon the bidding behaviour of participants. The problem, however, was that the experimenters had no initial idea about how this should be done: "How this [winner's curse] might work out when there [are] complementarities … is simply unknown. No experiments have been conducted that provide an assessment of what the dimensions of the problem might be" (Plott, 1997: 626).

Without any leads on what to look for, experimenters set up different causal scenarios that showed how the presence of complementarities might affect bidding behaviour and, consequently, government revenue. One such scenario is presented in figure 7, which shows how complementarities could lead to a reduction in the winner's curse via a two-rounds procedure: licenses are initially auctioned in packages and later on an individual basis. This would, presumably, result in an efficient auction design because licenses are allocated to the highest bidder in either the first or the second round, depending on whether the license is worth more to the buyer on an individual basis or as part of a package. Despite the fact that no theoretical or experimental evidence existed for this scenario involving complementarities, policy makers were still able to acquire a rudimentary understanding of how the introduction of complementarities could affect their aims by constructing these kinds of causal scenarios. This particular scenario showed, for instance, that government revenue will be higher when bidders value the aggregation of licenses more than licenses sold individually. Whether these causal relationships actually existed in the context of the FCC auctions had to be studied empirically, but policy makers at least had a rough idea about the operation of auctions that included complementarities.

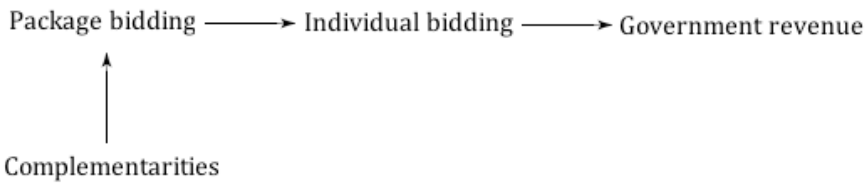

Figure 7: A causal scenario between the introduction of complementarities and government revenue: policy makers intervene by incorporating the existence of complementarities within the auction design. Presumably, auctioning the licenses as packages first would result in higher government revenue because bidders are supposed to value the aggregation of licenses more than licenses sold individually.

Arguably, policy makers could have come up with a whole list of different scenarios, each portraying some plausible chain of causal processes between the introduction of complementarities and government revenue. Interpreted as a causal scenario, a mechanism should be seen as a rudimentary representation of the variables and background conditions that *could* be relevant to the effectiveness of an intervention; it does not determine which of these actually *are* relevant. The actual relevance of evidence for an intervention's effectiveness will have to be established empirically. Thus, the usefulness of mechanisms in establishing evidential relevance is limited in the sense that they primarily play a role in the early stages of the policy-making process, where policy makers know little to nothing about the phenomena of interest.

2.4.3. CAUSAL SCENARIOS AND MODULARITY

The use of mechanisms in policy contexts by way of constructing causal scenarios differs from its use in extrapolating some piece of evidence from an artificial environment to the real world. Mechanism-based extrapolation and mechanism design, as put forward by Guala (2005) and criticised by Steel (2008), concerns knowledge about causal relationships that purportedly exist in the real world. By invoking mechanism, policy makers try to specify the necessary background conditions so as to make the extrapolation succeed. In addition, policy makers modify the mechanisms – i.e. the theory, experimental results and background conditions exhibited by the different mechanisms – in order to design a new mechanism that works best in a specific policy context. This particular mechanism is likely to not have existed before since it was created to suit the specific purposes of policy makers, as the final design of the FCC auctions aptly shows.

Mechanisms as causal scenarios, on the other hand, are not necessarily based on real-world entities with supposed causal tendencies. The mechanisms do need to be plausible to some extent, according to a basic theory or widely-held opinion among politicians and the public at large. For example, the scenario in which package bidding would precede individual bidding (see figure 7), which would result in higher government revenue, is based on the principle of 'the whole being

greater than the sum of its parts': the aggregation of licenses will be more advantageous to participants than the acquisition of individual licenses. This principle, it turned out, proved to be a useful starting point for policy makers in making sense of the effect of complementarities upon bidding behaviour.

So, what does this alternative use of mechanisms mean for the problem of external validity? Since the move from efficacy to effectiveness is reversed, as Cartwright (2009b) advocates, the issue whether mechanisms are able to support extrapolation appears to be less pressing. The reason is that some other kind of evidence, which might not be efficacious at all, could be more relevant to the effectiveness of an intervention. For example, it has been proved, both theoretically and empirically, that open auctions tend to reduce the winner's curse. Yet this particular efficacious result by itself is not very relevant for the FCC because it cannot account for the presence of complementarities. Indeed, evidence that describes the effect of complementarities upon government revenue, at least in some plausible way, is presumably more useful to policy makers than a confirmed causal tendency in isolation.[36] Although evidence for efficacy can be important, "efficacy is only one small piece of one kind of evidence (ibid.: 133).

Now, focusing on the effectiveness of interventions instead, let us turn to the issue of modularity. According to Cartwright:

> "Modularity thus secures a sure connection between causality and predictability under manipulation [intervention]. But how satisfying is this connection? In fact, it will show us to use a given causal relation for very few policy manipulations. That is because the kind of manipulations under which it guarantees invariance – and hence predictability from the laws of the system – are very special. They are just the kinds of 'surgical incisions' that we would demand in a controlled experiment, and these are very unlike real policy changes." (2009a: 414)

The notion of a "surgical incision" relates to our earlier discussion about the structure-altering nature of interventions, which remains problematic in many policy contexts. A major worry is that interventions will eliminate the causal relationship policy makers wish to exploit; due to the intervention, the relationships are likely to become spurious or merely reflect correlations. In fact, "when we do manipulate some factor [which is questionable in itself] we generally find ourselves changing far more than that single factor and its direct consequences. We usually end up changing

[36] Of course, evidence for efficacy may become useful in more advanced stages of the policy-making process. Supposing there exists a genuine causal relationship between complementarities and government revenue, policy makers can then clearly use this evidence to support their auction design. When this kind of evidence for efficacy is lacking, as is often the case in policy contexts, they can, initially, only rely on the causal scenarios that link complementarities with government revenue in some plausible manner.

a number of other factors relevant to the effect and very often *change the very principles by which these factors operate* as well" (ibid.: 415, italics added). As a result, policy makers often cannot be sure whether their interventions will be effective.

2.4.4. IT'S ALL ABOUT STABILITY

Interpreting mechanisms as causal scenarios cannot completely resolve the issue of modularity, for no matter what kind of evidence policy makers intend to use, there has to be a stable connection between the policy and target variable. This does not mean, however, that every stable connection has to be of a causal nature. Put differently, the supposition that causal knowledge is inherently better than knowledge of spurious relationships or correlations is dubious from a policy point of view. To see why spurious relationships or correlations might also be useful for policy makers, consider an abstract mechanism like that of figure 8. If we assume a relationship to be causal, then, according to the modularity requirement, there is always some intervention that leaves the relationship unaltered. This is then equally true for a spurious relationship between the joint effects of common cause, for an intervention upon the one will leave the other unchanged: intervening on α will affect x via z, while also affecting γ and leaving the spurious relationship between z and γ intact (Cartwright, 2009a). Therefore, interventions based on spurious relationships (or correlations) may be just as effective as those based on genuinely causal relationships, if not more.

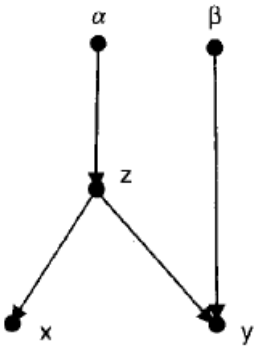

Figure 8: An abstract mechanism that shows how an intervention upon a spurious relationship (αaffecting x via z) will also affect y and leave this last relationship intact. Effective interventions, therefore, need not be based on causal relationships per se.

According to Julian Reiss, "what matters for policy is the stable association between the policy variable and the target, not the reason *why* the correlation is stable" (2008: 163–164, original italics). There is nothing wrong with the fact that the common cause z contributes to γ from a policy point of view. Of course, it would be a mistake to say that z explains the occurrence

of γ but, as discussed in chapter one, policy makers do not care too much for explanation anyway. What they do care about is how policy variables can be arranged so that they affect a target variable in a stable and reliable manner. So, whether the effectiveness of an intervention is brought about by causal or non-causal (spurious relationships or correlations) relationships is not too important. Both types of relationships can be seen as relevant evidence for policy hypotheses *as long as they are stable*. Following Cartwright and Reiss, it seems that the use of mechanisms as causal scenarios is not necessarily constrained to knowledge of causal relationships, but may also include non-causal knowledge.

With respect to the FCC auctions, a number of interventions were actually based on mechanisms that partly exhibited non-causal relationships. The introduction of complementarities, for example, lead to the breakdown of the causal relationship between open auctions and the winner's curse. Since policy makers had no evidence as to how open auctions would affect government revenue in the presence of complementarities, they initially had to rely on scenarios that showed how complementarities might affect government revenue (such as in figure 7). This particular mechanism was not constructed by way of causal relationships *per se*, for whether causality really existed had to be determined empirically. Without the benefit of hindsight, policy makers reasoned that complementarities could plausibly increase government revenue according to the principle of synergy. As a result, this mechanism did not include genuine causal relationships; in fact, this would have been unlikely due to its hypothetical nature.[37]

To close our discussion of evidential relevance with respect to the FCC auctions, let me reformulate the use of mechanisms as causal scenarios: the construction of (causal) scenarios, based on some plausible theory or principle, provided a preliminary understanding of the (causal) relationships relevant to the aim of the spectrum auctions. Again, the scenarios were used as a tool to indicate which evidence *could* be relevant; the actual credibility of the evidence within the context of the spectrum auctions had to be checked afterwards using empirical methods. While some of the scenarios may have included genuine causal relationships, others clearly did not (figure 7). Since the effectiveness of an intervention in a specific environment – as opposed to the evidence for efficacy elsewhere – counts as the most important criteria for policy makers, a scenario that reflects spurious relationships or mere correlations between variables could be just as useful as a purely causal scenario.

Although mechanisms, interpreted as (causal) scenarios, included non-causal features, these spurious relationships and correlations still had to be stable in order to be of real use to

[37] Other plausible scenarios with respect to the relationship between open auctions and government revenue may include causal elements but, in general, they do not have to. Generally speaking, the notion of 'scenario' implies that the variables need not be causally connected because their causality is a possibility to be confirmed empirically, not a given *a priori*.

policy makers. In short: while the relationships incorporated into the scenarios need not be causal, stability is nonetheless a necessary condition for the effectiveness of interventions. It is imperative, then, that policy makers are aware of this condition and, whenever possible, try to account for violations of modularity further on in the policy-making process.

CHAPTER THREE

MECHANISMS AND BEHAVIOURAL POLICY

The previous chapter discussed the use of mechanisms in the domain of auction policy. To further substantiate the arguments in favour of using mechanisms for policy-making this chapter will deal with another policy domain, namely the domain of behavioural policy. More specifically, the potential of mechanisms will be assessed with respect to interventions that are based on insights from behavioural economics. As with auction policy, this will be done by way of a case study: the Save More Tomorrow™ (SMT) pension plan developed by Richard Thaler and Shlomo Benartzi (2004, 2013). This particular intervention should be seen as an example of *nudging*, which is a relatively recent movement within the behavioural policy domain that tries to induce people to make better choices by changing the choice architecture they face. The justification of behavioural policies, including the SMT pension plan, remains rather controversial. Therefore, investigating the role of mechanisms will prove helpful in this regard.

Similar to the FCC auctions, the usefulness of mechanisms with respect to the SMT pension plan will be discussed from the perspectives of external validity and evidential relevance. For this purpose, I will largely draw on the work of Till Grüne-Yanoff (2015) who advocates, amongst other things, that knowledge of mechanisms is necessary for the justification of interventions. With regard to external validity, I will argue that whether interventions are judged efficient depends on which mechanism is used by policy makers. While using one mechanism to explain how an intervention operates may result in the violation of modularity – which would render the intervention ineffective – using another mechanism might result in a reliable extrapolation. As for evidential relevance, the evaluation of different mechanisms, interpreted as causal scenarios, supports the identification of the relevant evidence for policy hypotheses. Although there are often many mechanisms available, policy makers nonetheless have to search for them in order to justify their interventions.

3.1. SAVE MORE TOMORROW™

Since the 1980's, principles from psychology have started to penetrate economics. Apart from developing normative theories, such as rational choice theory (RCT)[38], economics now also deals

[38] Not to be confused with a randomised controlled trial, which is also abbreviated as 'RCT'. Throughout this chapter, the abbreviation 'RCT' will refer to rational choice theory – if not indicated otherwise.

with descriptive and prescriptive theories.[39] Descriptive theories try to model the actual behaviour of people, emphasizing their frequent departure from normative theories. For example, the contributions to prospect theory (Kahneman & Tversky, 1979) and regret theory (Loomes & Sugden, 1982) have exposed the anomalies of people's behaviour with regard to RCT. Prescriptive theories, then, are attempts to improve individual's decision-making and bring their behaviour closer to the normative ideal. The SMT pension plan is a foremost example of the application of prescriptive theory, where policy makers and policy-oriented economists have tried to increase savings contribution rates amongst employees. Their basic idea was "to give workers the option of committing themselves now to increasing their savings rate later, each time they get a raise" (Thaler & Benartzi, 2004: 166).

The raison d'être of the SMT pension plan was the troubling observation that households typically fail to save enough for retirement. More precisely, households tend to save less than the life cycle theory of saving would suggest, which assumes people to optimize their expected consumption level – i.e. they intend to smooth consumption over the course of their lives. Admittedly, this task is a difficult one, even for those who have an economics background. Many people also lack the self-control to commit to a higher savings contribution rate. A third issue is procrastination, which results in a status-quo bias: once people are enrolled into a certain pension plan they are likely to remain in it for a considerable period of time due to inertia or naivety.

To overcome these problems, Thaler and Benartzi (2004, 2013) included four rules in their design of the SMT pension plan:

1. *Early sign-up call*. Employees are approached by company personnel about signing up for the SMT plan way ahead of their first scheduled pay rise. Since employees tend to value increased saving in the future more than higher savings contributions rates now (hyperbolic discounting), the time lag between the sign-up date (in case they decide to sign up) and the first pay rise should be as long as possible. This way, employees will perceive the plan to be more attractive because its consequences are not immediate.

2. *Automatic escalation of the savings contribution rate*. The savings contribution rate of employees is increased after each pay rise. According to the behavioural mechanism of loss aversion, people have the tendency to weigh losses more heavily than gains. Consequently, employees are reluctant to increase their savings contribution rate because they perceive this as a loss in disposable income. Increasing the savings contribution rate

[39] The distinction between normative, descriptive, and prescriptive theories has been proposed by Howard Raiffa (1982).

just after every pay rise mitigates the effect of loss aversion since employees' higher income compensates for their increased amount of savings.

3. *Cap on maximum contribution rate*. The savings contribution rate of employees is gradually increased until it reaches a pre-set maximum. This way, policy makers exploit the states-quo bias of employees, which acts as a pull factor in keeping participants enrolled in the plan.

4. *Possibility to opt-out*. Knowing that they can always quit the SMT plan, employees will be more comfortable about singing up. In effect, this option increases the chance that employees will actually enrol and start accumulating more savings.

So, did the SMT pension plan actually succeed in its objective of increasing total retirement savings amongst employees? The first implementation at a US mid-size manufacturing company generated promising results: 78% of the employees that were offered the SMT plan decided to sign up, and subsequently increased their savings contribution rate from 3.5% to 13.6% in about four years (Thaler & Benartzi, 2004). Since then, the adoption of the SMT pension plan – along with similar programmes using automatic enrolment[40] and automatic escalation rules – has grown rapidly. Thaler and Benartzi (2013) report that 56% of employers automatically enrol their employees in a pension plan and, more significantly, 51% of them do so using an automatic escalation scheme. Based on conservative assumptions, estimations show that the SMT type of pension plan boosted annual savings by $7.4 billion since its inception.

Despite its resounding success, the SMT pension plan was not designed on theoretical grounds alone. Although important insights from behavioural economics were available, these did not indicate the ideal level of the parameters to be included in the design. For instance, while loss aversion suggested that an increase in the savings contribution rate should shortly follow a pay rise, it did not tell policy makers exactly how high the increase must be. In addition, then, the group of behavioural economists, financial consultants and policy makers had to rely on the specific background conditions of the environment in which the plan was to be implemented. With respect to the US manufacturing company, these included the preferences of blue collar workers in receiving advice from a financial consultant. When they were confronted with the consultant's recommendation straight away they would become unresponsive; only after having discussed their financial situation could an appropriate rate increase be determined. The initial SMT plan,

[40] Based on the original SMT plan, some pension plans include automatic enrolment where employees are automatically signed up unless they deliberately decide to opt out. This further increases participation and, in turn, contribution rates.

based on behavioural theory and particular background conditions, was then tested empirically by way of subsequent implementations.[41] Thus, the success of the SMT pension plan largely rested on the intricate combination of theoretical, experimental, and background knowledge – three sources of knowledge that are, as we have seen in the previous chapter, crucial to the design of any effective intervention.

3.2. EXTERNAL VALIDITY

The problem of external validity is a recurring issue within applied economics that cannot be understood properly by referring to just one case study. This section, therefore, has the task of analysing how the SMT pension plan relates to mechanism-based extrapolation. That is, to what extent did the intervention of the SMT plan made use of mechanisms in order to improve its effectiveness? With respect to this case, I will argue that its effectiveness depended on which mechanisms were used by policy makers. While mechanisms are, in principle, able to identify the necessary background conditions for extrapolation to work well, they may not always succeed in doing so. Moreover, whereas an intervention might violate modularity according to one mechanism, it may not do so according to another.

3.2.1. BACKGROUND CONDITIONS

According to Steel (2008), mechanism-based extrapolation has two main requirements. First, the background conditions between the artificial (experimental or otherwise) and the target environment need to be as similar as possible. Policy makers can only conduct effective interventions if they possess sufficient background knowledge about their target environment, which is provided by inquiring upon different mechanisms. This way, each mechanism partly shows how an intervention that worked in one context will also be operative in the context of interest. In other words, mechanisms indicate which background conditions are necessary for some piece of evidence (for efficacy) to travel to other policy contexts.

The identification of the necessary background conditions can be, of course, a difficult and time-consuming task for policy makers. More often than not there are numerous possible mechanisms that could be used to support the effectiveness of interventions. Although Thaler and Benartzi (2004) mention a couple of mechanisms that could play a role in the success of the SMT

[41] Thaler and Benartzi (2004) describe two other implementations of the SMT pension plan. Each implementation generated new insights, such as the fact that "linking savings increases to pay increases, while desirable, may not be essential" (ibid.: 179).

plan – specifically procrastination and loss-aversion – there are arguably many more potential mechanisms to choose from. In this sense, "the problem is not that there are no mechanistic models associated with these policy proposals [like the SMT pension plan], but rather that there are too many – and that there is hardly ever any evidence provided to choose between them" (Grüne-Yanoff, 2015: 5).

However, following Guala (2005), the evaluation of mechanisms is guided by way of the explicit aims of interventions. Recall that before policy makers set out to study the mechanisms, whereby they wish to acquire the necessary background conditions, they first specify the aim of the intervention in concrete terms. This helps to focus their attention upon mechanisms that would potentially contribute to the policy objectives, while disregarding those who presumably do not (directly) enhance the aim at hand or merely produce ambiguous results.

As for the SMT pension plan, the objective was straightforward: to increase total savings of employees. Here some mechanisms, *prima facie*, appeared to be more promising than others. Those including behavioural insights such as procrastination and loss-aversion were clearly worth investigating because they were already presented by the academic community as likely causes for the observed behaviour of low savings. A mechanism involving procrastination, for instance, results in a status-quo bias, which positively affects employee enrolment. In turn, higher (and longer) enrolment may significantly contribute to the total savings by employees. Now, whether this effect will actually materialise depends on the background conditions that the mechanism assumes to be in place. The status-quo bias of employees will only be obtained if the they really are inert when it comes to thinking about saving for retirement. Without the condition of inertia, the intervention will be less effective because employees may be more inclined to quit the SMT plan at any given moment. Therefore, the inertia of employees turned out to be a necessary background condition for the success of the SMT pension plan (see rule 3 of the SMT plan in section 3.1).[42]

In contrast, a mechanism involving the recommendation effect – the causal process where a policy maker recommends a pension plan to employees through automatic enrolment – would not result in a status-quo bias, and thus does not assume employees to be inert. For if they were, then the recommendation effect would become obsolete. According to this mechanism, employees do not enrol in a pension plan because they are simply ignorant about the whole idea, but instead they judge it positively due to the policy maker's implicit recommendation. In this sense, the

[42] Admittedly, the distinction between a mechanism and its background conditions can be quite vague and open-ended. Here, the background condition of inertia is supposed to be a part of the procrastination-mechanism in the sense that it enables the development of status-quo bias. Yet inertia could also be a separate mechanism including similar psychological processes but in a somewhat different order. Even if we would consider inertia to be a mechanism itself, Grüne-Yanoff notes not everybody would agree on this – see, for instance, Berg and Gigerenzer (2010).

background condition of inertia does not appear to be necessary for the SMT pension plan's effectiveness.

This example illustrates how mechanisms *can* identify the necessary background conditions, but it far from proves that mechanisms will always do so. Whereas the procrastination-mechanism is able to show why the condition of inertia is important for a pension plan aimed at increasing total savings, the recommendation-mechanism gives policy makers a reason to believe the opposite. How, then, do policy makers decide on whether inertia is a necessary background condition or not? Analogously, which mechanism actually operates here, procrastination or recommendation?

These kinds of questions go back to a central concern with respect to mechanism-based extrapolation, namely that knowing which mechanism operates in the target environment would no longer require extrapolation. Steel aptly calls this concern the *extrapolator's circle* (2008: 94–96). He argues that the appropriateness of a mechanism can only be determined if one already knows that the extrapolation will be successful; since it is unclear whether extrapolation is possible at all, justifying the operation of a mechanism remains problematic.

If policy makers really only care about the effectiveness of an intervention in a specific environment, then they might as well leave extrapolation for what it is and focus on simple induction. When designing the SMT pension plan, for example, policy makers could put the insights from behavioural economics aside and inquire about the causal processes with respect to the saving behaviour of employees empirically. Especially when mechanism-based extrapolation leads to contradictory results – as the example of the procrastination and recommendation mechanisms shows – would it not be better to refrain from the enterprise altogether?

Unsurprisingly, it would not. Despite the call for more inductive inference within economics[43], the problem of external validity, and thus the need for extrapolation, would not go away completely. Since there is often a considerable time lag between field experiments and the implementation the results derived from these experiments, the target environment is likely to undergo significant changes in terms of demographic, technological, and political conditions. For policy makers, this problem is particularly severe because they often deal with rapidly-changing environments (consider the high turnover rates of employees at large companies, for instance).

Perhaps, then, it is wrong for policy makers to look for that single, unique mechanism 'somewhere out there' which happens to satisfactorily extrapolate a theoretical/experimental result to the environment of interest. Instead, I claim that policy makers have more to gain from mechanism design, which can be seen as a procedure that combines the need for extrapolation

[43] See, for instance, Levitt and List (2009).

with more inductive methods. In effect, policy makers study an array of mechanisms and take what they need from each of them in order to further develop their intervention. Some of these mechanisms may prove more useful than others but for an intervention to work well, many different mechanisms must be studied. Eventually, this procedure results in the design of a new mechanism, one that has integrated the most useful aspects – including theory, experimental results, and background conditions – of all the mechanisms that have been evaluated by policy makers.[44]

3.2.2. MODULARITY: BITING THE BULLET?

The second requirement of mechanism-based extrapolation is modularity: an intervention should only affect those parts of the mechanism for which it was intended. According to Steel (2008), interventions will only be effective if they do not alter the causal structure of a mechanism used for extrapolation. In general, structure-altering interventions can be problematic because they lead to a breakdown in the causal relationships policy makers wish to exploit. The discussion of the FCC auctions in the previous chapter has already shown us that modularity remains a thorny issue. Is there another way out?

Grüne-Yanoff (2015) claims that interventions may or may not violate modularity, depending on which mechanism policy makers use. Some interventions can be considered structure-altering because they do not switch off all other causal tendencies on the variable intervened upon.[45] In this situation, the intervention is but one of the causal tendencies active upon the policy variable. Lacking the assumption of *ceteris paribus*, interventions might change the causal relationships that purportedly exist between the policy and target variable. As a result, policy makers cannot know for sure whether their intervention actually contributes to the preferred outcome.

To see if interventions are indeed structure-altering in the sense described above, let us return to the SMT pension plan. The first rule of the SMT plan was based on the behavioural concept of hyperbolic discounting, where people tend to value lower consumption in the future (due to an increasing savings contribution rate) more than lower consumption now. According to

[44] Although the approach of economic engineering looks as if it begs the question – using mechanisms to construct mechanisms – but in fact it does not. For the approach does not use hypotheses about the existence of mechanisms as premise in its argument that mechanisms exist. Instead, policy makers take those aspects of the purported mechanisms (be it some piece of theory, a certain experimental result, or a salient background condition) and modify them to fit the target environment. These aspects, then, together form a new mechanism, especially designed for the aims of the intervention at hand.

[45] This critique corresponds to the second condition of an ideal intervention in Woodward's (2003) framework. See section 1.3.2.

Grüne-Yanoff, there are two potential mechanisms that could indicate how introducing a temporal difference between signing up to the pension plan and the first increase in the savings contribution rate is deemed effective. The first mechanism is related to visceral factors – such as hunger, thirst or other kinds of physical cravings – which increase in strength as their expected satisfaction nears the present. As influences on behaviour, they contrast with people's rational deliberations in terms of self-interest. While employees might want to save more for retirement because they believe it would a rational thing to do, the prospect of giving up current consumption is amplified through visceral factors. Consequently, according to this mechanism, employees are more likely to increase their savings contribution rate only if presented with the option of doing so in the future.

The second mechanism proceeds from the distinction between certain and uncertain outcomes. In general, people tend to prefer the former over the latter due to the possibility of not getting a reward that is acquired in the future. Employees that choose to give up consumption now for more savings later are trading a certain outcome for an uncertain one, which is a choice that not many people would find attractive. The SMT pension plan, however, effectively mitigated the influence of this distinction by changing the choice architecture faced by employees. As opposed to increasing one's savings contribution rate now, doing so in, say, six months' time means employees are trading one uncertain outcome for another. This would make the prospect of a higher savings contribution rate in the future (and thus lower consumption) relatively more attractive because it does not compete with a certain outcome.

Now let's see how these two mechanisms would fare with respect to the requirement of modularity in the case of the SMT pension plan. Judging from the uncertainty-mechanism, the first rule of the plan may violate modularity because it does not block all other causal influences on uncertainty. For example,

> "a clumsy communication and implementation of the *Save More Tomorrow*™ plan might also create the impression that retirement plans are changed haphazardly, letting employees revise their uncertainty judgments about their ability to retrieve funds. Such an increase in uncertainty, caused by the implementation, might well erase any positive effects that the policy would otherwise have had on contributions rates." (Grüne-Yanoff, 2015: 11)

Although the SMT plan is likely to increase total savings by employees through the clever use of differential uncertainty judgments, its effectiveness may be undone if the remaining influences on uncertainty – such as communicative flaws – are significant. If this is indeed the case, then modularity would be violated and the usefulness of this mechanism could be questioned.

However, the same intervention does not violate modularity when judged with the help of a different mechanism. Using the mechanism of visceral factors as line of reasoning instead, flaws in the communication and implementation of the SMT plan will, *ceteris paribus*, have no complicating effect on the success of the intervention. Here the temporal difference between sign-up date and start date of the SMT plan is based on the idea that visceral factors will be weaker (and employees therefore more rational) when employees are able to defer increases in their savings contribution rate. Given the mechanism of visceral factors, whether the SMT pension plan is communicated and implemented properly does not matter much, for differential uncertainty judgments are not assumed to play a significant role.

When discussing an intervention's potential violation of modularity, Grüne-Yanoff (2015) adds two more complications to the argument that it depends on which mechanism policy makers use. Namely, even if an intervention initially does not violate modularity, it may do so in a later stage or after repeated implementation. With regard to the mechanism of visceral factors, for instance, the positive effect of temporal difference might wear off when the SMT plan is well under way: employees anticipate the benefits of the plan as they near retirement age, which may enhance their desire for higher consumption now (and thus a lower savings contribution rate). At some point, employees' positive attitude towards saving in the future as opposed to saving now will become blurred because their perceptions of 'the future' and 'now' gradually merge into one.[46]

Similarly, an intervention may violate modularity only after it has been repeatedly implemented. For example, Thaler and Benartzi (2004) report that the link between contribution rate increases and pay rises – based on the mechanism of loss-aversion (see rule 3 of the SMT plan) – appeared not to be "essential" (ibid.: 179) after three consecutive implementations. That is, the SMT plan became less effective because its repeated implementation had structurally altered the causal relationship related to people's perception of gains and losses. The intervention based on the mechanism of loss-aversion, therefore, had resulted in breaking down the causal relationship it intended to exploit.

These complications are valid and reveal further shortcomings of mechanism-based extrapolation. No intervention by policy makers will be effective indefinitely. The environments in which interventions take place are often highly complex, which increases their chances of violating modularity. Depending on which mechanism is used by policy makers, interventions can

[46] In their analysis, Thaler and Benartzi (2004) find that the replacement ratio (the percentage of an employee's income that is paid out by a pension plan upon retirement) of employees who commit to a SMT-type of pension plan early on are significantly higher than for those joining at an older age. Quite obviously, increasing one's savings contribution rate is more attractive for younger employees. Once employees become older, they may decide to drop out of the plan because its relative impact on their total savings becomes smaller. Hence, the urge of having a higher present consumption level becomes stronger.

indeed be considered modular. Yet even (initial) modular interventions may become ineffective over time or after repeated implementation. The problem is that policy makers often have very little evidence to choose between different mechanisms on the basis of modularity. That is, it is very difficult for policy makers to justify an intervention by referring to a particular mechanism because they cannot know for sure that modularity will not be violated. The upshot, then, is that policy makers have to study many different mechanisms, and regularly check whether their interventions do not violate modularity.

3.3. EVIDENTIAL RELEVANCE

Although mechanism-based extrapolation can, in principle, be used to overcome the problem of external validity, the issue of finding the right mechanism (or mechanisms) for the job remains pertinent. Grüne-Yanoff (2015) argues that interventions can be considered effective – i.e. they identify the necessary background conditions and do not violate modularity – depending on which mechanisms policy makers use. Given that there are often many potential mechanisms to choose from, how do policy makers decide upon which mechanisms to use for their interventions? In this section, I will discuss one possible approach to this problem, namely that of evidential relevance. More specifically, I will defend the following two claims. First, policy makers in the case of the SMT pension plan adopted the perspective of evidential relevance – what evidence is relevant to the policy hypothesis? Secondly, they proceeded in this task by using mechanisms as causal scenarios. These scenarios indicated which evidence was likely to be relevant for the aims of the SMT plan, and which was not. In effect, mechanisms functioned as a filter for welfare judgments about the consequences of the SMT plan.

3.3.1. WELFARE

In his discussion of the SMT pension plan, Grüne-Yanoff endorses the distinction between efficacy and effectiveness put forward by Cartwright (2009c). The policy makers involved in the SMT plan were mostly interested in the effectiveness, or, more precisely, the *benefit* that would actually occur when the plan was implemented. In particular, they wanted to know how the SMT plan would impact the welfare of employees with respect to savings. In this sense, a subjective welfare criterion in terms of welfare-promoting decisions was used to evaluate the effectiveness of different pension plans. This criterion was in line with the general aim of nudging, which is to improve people's decision-making by changing the choice architecture they face.

The aim of the SMT plan was to increase total savings by employees. As with the FCC auctions, this aim featured as the point of departure for policy makers. This meant that all gathered evidence – a behavioural model, some experimental result, or a certain background condition – was evaluated according to its potential contribution to increasing total savings. While some evidence only contributed indirectly, such as the background condition of inertia, this kind of evidence was highly relevant nonetheless. Other evidence, however, proved to be far less relevant to this aim, like the life cycle theory of saving. Either way, it was crucial for policy makers to determine the relevance of the evidence with respect to increasing retirement savings early on in the policy-making process. Whether the relevant evidence also turned out to be credible had to be determined at a later stage.

Recall that evidence for efficacy, though important, "is only one small piece of one kind of evidence" (Cartwright, 2009b: 133). From a policy point of view, knowing that a variable is efficacious in some experimental (ideal) environment is nice, but it is not particularly helpful. As discussed in section 3.2.1, being able to extrapolate evidence for efficacy to the target environment requires the specification of all the necessary background conditions as well as checking the modularity of interventions. It was shown that mechanism-based extrapolation can be done but that its success is by no means guaranteed. By focusing less on evidence for efficacy and more on evidential relevance, policy makers are able to develop more effective interventions because each piece of evidence that they plan to use is supposed to be relevant, though in different degrees, to the aim at hand.

Of course, evidence for efficacy might be part of the relevant evidence for a particular policy aim, too. The behavioural concepts of status-quo bias and loss aversion, for example, were both relevant to and efficacious in the target environment. Yet it is only possible for policy makers to know that an efficacy claim is applicable to the target environment if it is also relevant. In short: while evidence for efficacy may also be relevant evidence, a piece of relevant evidence need not be efficacious. That is, it is not only evidence for efficacy that can be relevant to a policy hypothesis; more often than not we need additional evidence, which does not have to be efficacious *per se*. In the case of the SMT pension plan, policy makers did not restrict themselves to only using evidence for efficacy. Rather, they included many different kinds of evidence in their plan – such as theory, experimental results, and background knowledge about the target environment. Thus, what made the SMT plan successful was not merely the study of evidence for efficacy, but the evaluation of all the relevant evidence.

3.3.2. CAUSAL SCENARIOS RECONSIDERED

Mechanisms, interpreted as causal scenarios, can be useful for policy makers in order to determine the relevance of different kinds of evidence.[47] Grüne Yanoff (2015) distinguishes three aspects of an intervention's effectiveness: robustness, persistence and welfare effects. All three aspects are clearly related to the problem of external validity, but they can also – as will be done below – be taken as a starting point for the development of interventions. Since the welfare effects of interventions matter greatly for any policy, these function as the main criterion for establishing evidential relevance. Given this welfare criterion, policy makers then construct a variety of causal scenarios, where each scenario indicates how certain variables and background conditions supposedly contribute to improving welfare. In this process, policy makers may base the causal scenarios on some basic theory, general principle, or widely-held public opinion. While the list of plausible scenarios is potentially endless, it is at least guided by the welfare-improvement that the intervention intends to bring about.

To illustrate how mechanisms were used as causal scenarios during the development of the SMT pension plan, let us return to the example of introducing a temporal difference between the sign-up date and the first savings contribution rate increase (rule 1 in the SMT plan). As discussed in section 3.2.2, there are two mechanisms that could potentially be used to safeguard modularity: the mechanism related to visceral factors and the uncertainty-mechanism. It was shown that, in this case, the mechanism related to visceral factors made sure the intervention did not violate modularity.

Now let's shift our attention away from issues of robustness and persistence, and focus on the potential welfare effects of interventions. Nudges like the SMT pension plan are supposed to improve the welfare of employees by helping them avoid manipulative forces. Using the visceral factors mechanism, the temporal difference induces people to behave according to their self-interests. According to Grüne-Yanoff, "it helps free people's choices from visceral influences, and instead allows people to choose so as to satisfy their consistent and well-informed preferences" (2015: 17). More concretely, employees are more likely to make rational decisions concerning saving for retirement when they are not exposed to visceral (irrational) factors. Their behaviour brought about by the SMT plan constitutes a welfare improvement because they are no longer manipulated through short-term thinking.

In contrast, the SMT pension plan does *not* constitute a welfare improvement according to the uncertainty-mechanism. When viewed from this perspective, one could argue that preferring

[47] As such, mechanisms themselves do not act as evidence for the effectiveness of interventions. Rather, they are a tool for policy makers to search for evidence and evaluate its relevance in a preliminary manner.

relatively less consumption now to more consumption in the future is compatible with the rationality of employees. The reason is that sooner consumption is judged as less uncertain than postponed consumption. Since it is perfectly rational to prefer a certain outcome over an uncertain one, the welfare of employees would not be enhanced by changing the option of a certain outcome into one of uncertainty. Consequently, the same behaviour brought about by the SMT plan would not constitute a welfare improvement because, under the uncertainty-mechanism, it is assumed there is no manipulative force present. Whereas this particular intervention is considered to be welfare-improving with respect to the mechanism of visceral factors, the same intervention cannot be judged as such when viewed from the uncertainty-mechanism.[48]

The two mechanism in the previous example can be seen as causal scenarios that showed whether and how an intervention actually affected welfare. The first step in constructing causal scenarios was for policy makers to specify the preferred outcome that every scenario must arrive at. This was done by referring to the subjective welfare criterion, which judged behaviour to be welfare-improving if based on rational preferences. In the example above, the causal scenarios were supposed to improve the decision-making of employees with respect to saving for retirement.

From this, policy makers proceeded by setting up different causal scenarios including variables and background conditions that would presumably contribute to improving welfare. These causal variables and background conditions were derived from theories loosely related to the policy context, such as RCT and other behavioural theories. With the help of some basic theoretical insights, policy makers were able to formulate concrete interventions. One of these interventions was the introduction of a temporal difference between the sign-up date and the first increase in the savings contribution rate, where its general idea was based on the behavioural concept of hyperbolic discounting.

Once the preferred outcome and the causal scenarios had been specified, policy makers evaluated whether the scenarios actually had the desired effect. In this case, only the causal scenario involving visceral factors would constitute a welfare improvement. The causal scenario of uncertainty, though plausible, was not considered to improve welfare because it did not avoid a behavioural manipulation. Given the specific aim of policy makers, the construction and evaluation of causal scenarios enabled them to determine the relevance of the evidence in a preliminary manner: the rule of having a temporal difference in the SMT plan was relevant for improving welfare when viewed from the causal scenario involving visceral factors, but it was

[48] It is important to emphasize here that, when interpreting mechanisms as causal scenarios, policy makers are not looking for that one mechanism actually operating in the target environment (as with mechanism-based extrapolation). Rather, they evaluate several different mechanisms (scenarios) and then decide which one to use for the justification of the intervention. More on this below.

much less relevant when judged by the causal scenario based on uncertainty. Thus, although the SMT plan appeared to be welfare-improving in one sense (visceral factors), its supposed relevance to welfare was weakened upon further inquiry (uncertainty).

Ultimately, then, policy makers had to make a judgment call whether introducing a temporal difference actually constituted a welfare-improvement or not. Their final judgment was supported by causal scenarios that indicated which kinds of evidence *could* be relevant to improving welfare. These scenarios informed policy makers about the preliminary relevance of the evidence – be it theory, experimental results or background conditions. Yet, whether this evidence actually *was* relevant remained an open question and had to be determined empirically. In case of the SMT pension plan, policy makers decided that introducing a temporal difference would constitute a welfare improvement for employees. In a sense, they used causal scenarios as a filter for judging the welfare effects of their proposed intervention: the first rule of the SMT plan could be justified according to the causal scenario involving visceral factors. Using another filter, however, would have led to the rejection of the intervention since, under the causal scenario with regard to uncertainty, the temporal difference would not have been justified.

3.3.3. JUSTIFYING INTERVENTIONS

This brings us to the question of when an intervention is sufficiently justified. Grüne-Yanoff tries to make sense of this difficult issue by proposing the following sufficiency principle:

> "A policy [intervention] is based on sufficient mechanistic evidence if it takes all available mechanistic evidence[49] into account, where availability is constrained by current theoretical and technological feasibility. If information of this sort does not enter the discussion at all, these policies cannot and should not be described as 'evidence-based'." (2015: 18)

Mechanisms – interpreted as causal scenarios – can, in general, be used to justify an intervention.[50] If there is one particular mechanism that indicates an intervention to be welfare-improving, and

[49] In his paper, Grüne-Yanoff refers to 'mechanistic evidence' when he discusses the role of mechanisms in behavioural policy. Here, I interpret the use of mechanistic evidence by policy makers as being similar to the construction and evaluation of different causal scenarios: just like there can be many potential mechanisms that policy makers can use for extrapolating some result from an artificial environment to the one of interest, there can also be many causal scenarios that could plausibly indicate what evidence could be relevant to the policy aim.

[50] Discussing another interesting case study, Michiru Nagatsu (2015) convincingly defends the justification of nudge policies by referring to mechanisms. Although he does not explicitly interpret mechanisms as causal scenarios, an attempt could be made to integrate his arguments with the interpretation presented in this thesis. For the sake of focus (and space), I will not do so here.

there are no other mechanisms that provide relevant evidence in the opposite direction, then policy makers can be confident to invoke this mechanism for the justification of their intervention. At the very least then, every intervention should be based on evidence put forward by some kind of mechanism.

Yet, as he admits in the same passage, it is often very difficult to obtain sufficient evidence that fully justifies an intervention by invoking mechanisms. Unfortunately, policy makers have to do with partial evidence in favour of a particular intervention, where mechanisms can only provide "qualitative information regarding factors upon which the policy's effectiveness and welfare-properties is likely to depend" (ibid: 18). Even in situations where mechanisms indicate evidence *not* to be relevant to the policy aim, it will be useful for policy makers to be aware of the reasons why an intervention might not be justified. For these reasons will, in themselves, contribute to the justification of the intervention (albeit in a negative way).

In order to enhance the justification of interventions, mechanisms need to be further differentiated, accounting for different theories, empirical data and background conditions. According to Cartwright and Hardie (2012), "to decide about [the] effectiveness [of interventions] requires an open-ended process of thinking that is inevitably contextual and cannot be reduced to rules" (ibid.: 11). This induces a certain modesty amongst policy makers given that they can hardly ever completely justify their interventions, as the SMT pension plan aptly illustrates. If nothing else, mechanisms caution policy makers against the premature implementation of proposed interventions. Thus, the process of establishing evidential relevance by invoking mechanisms might yield no initial result, may be very costly, or could take considerable time. But search one must.

CONCLUSION

In this thesis, I have defended the claim that mechanisms are a useful tool for economic policy-making. More specifically, the potential of mechanisms for the function of control has been assessed according to two main arguments: external validity and evidential relevance. To conclude this thesis, I will restate my claims with respect to these two arguments and briefly summarize the most important findings obtained through the discussion of the previous two case studies. Finally, some suggestions for future research will be provided.

The problem of external validity is a pressing and recurring issue within applied economics. More often than not, theoretical and experimental results do not hold outside the artificial environments in which they were originally obtained. This is problematic for policy makers because they need to know whether certain causal relationships found in one context will also operate in a target context. Put differently, if policy makers want to conduct effective interventions, then they have to be able to rely on causal relationships that can to travel to other environments, too. One supposed solution to this problem is mechanism-based extrapolation, where a mechanism is used to specify the similarity in background conditions between the artificial and the target environment in order to successfully extrapolate a causal relationship from the former to the latter. This way, policy makers will know that a result obtained 'there' will also hold 'here', for the mechanism indicates whether all the necessary background conditions are in place.

With respect to mechanism-based extrapolation, I have argued that mechanisms play an important role in the design and implementation of policy interventions. Following Guala (2005), the main advantage of using mechanisms for policy purposes is the ability to integrate theoretical, experimental, and background knowledge into concrete, effective interventions. That is why he speaks of 'mechanism design': policy makers try to design a mechanism that works best in the target context by first specifying the aim of the intervention, and then studying many different mechanisms which are able to contribute to this aim. In this procedure, only the most useful parts of the theory, experiments and background conditions are included into the final design. The FCC auctions have aptly illustrated this procedure, in which policy makers were able to integrate insights from game theory and experimental (testbed) results with the specific background conditions of the spectrum auctions.

Using mechanism-based extrapolation for policy-making should be seen a kind of *economic engineering*: policy makers modify mechanisms that purportedly exist in the real world so as to make their interventions work effectively. In this sense, policy makers do not try to

extrapolate one particular causal relationship by invoking its underlying mechanism, but instead combine knowledge of many different mechanisms related to the phenomenon under investigation. They study an array of mechanisms and take what they need from them, so to say, in order to further develop their intervention.

The testbed experiments used in the FCC auctions are a good example of what this procedure looks like in practice. Here the experiments mimicked the background conditions of the actual spectrum auctions as closely as possible, which enabled policy makers to compare the results obtained in the artificial environment with their operation in the target environment in a straightforward manner. In effect, policy makers first move from the real world to the laboratory and then back to the field again. Since extrapolating some causal relationship found under artificial conditions to the real world is often very difficult, policy makers can benefit from importing the most important empirical features of the phenomenon of interest into the experimental setup. Eventually, this procedure leads to the development of a new type of mechanism, which is specifically created to enhance the aims of policy makers.

There is, however, one particularly important objection against mechanism-based extrapolation to take into account. Namely, interventions that draw on knowledge of mechanisms can alter the structure of the causal relationships policy makers wish to exploit. On this, I have argued that mechanism design, though susceptible to structure-altering interventions, can be a suitable extension of mechanism-based extrapolation, and should thus be taken seriously by the economic policy-making community. Although interventions that adhere to the demands of modularity are very hard to come by in policy contexts, the FCC auctions have proven it is possible: the extensive use of experiments played a crucial role in this regard.

Admittedly, the procedure of mechanism design may be very costly and time-consuming. A more fruitful way to deal with the issue of modularity is proposed by Grüne-Yanoff (2015), who claims interventions may or may not violate modularity depending on which mechanism policy makers use. The SMT pension plan, for instance, can be considered structure-altering according to the uncertainty-mechanism, while the same conclusion does not hold when using the mechanism of visceral factors. If policy makers manage to identify a mechanism through which their intervention does not violate modularity, then the justification of that intervention can be enhanced. Yet to complicate matters even more, interventions might also violate modularity in a later stage or after repeated implementation. Therefore, the only way to make sure that an intervention does not violate modularity is for policy makers to evaluate many different mechanisms and to do so regularly.

The potential of mechanisms for policy purposes can be understood differently, and more appropriately, in terms of evidential relevance. That is, mechanisms have the ability to provide a preliminary understanding of the evidence that could be relevant for the effectiveness of policy interventions. Whereas the theorist or experimenter tries to answer the question 'where are my obtained results relevant?', the policy maker is primarily concerned with the question 'what kind of evidence is relevant to my policy hypothesis?'. So the conventional perspective of moving from efficacy to effectiveness is reversed within the domains of evidence-based policy.

With respect to evidential relevance, I have argued that policy makers adopt this alternative perspective since they take the aims of interventions as their point of departure. Evidence for efficacy, established by experimental methods such as RCTs, can be valuable to scientists but this kind of evidence is not necessarily relevant to policy makers. In principle, any kind of evidence can be useful for policy makers as long as it is relevant to the aims of their interventions; before evidence is deemed credible (i.e. efficacious) in the target environment it must be considered relevant first. Both in the case of the FCC auctions as with the SMT pension plan did policy makers focus on the relevance of the evidence for the interventions' objectives, where government revenue was a focal point in the former and total retirement savings of employees in the latter. While some evidence proved to be highly relevant to the aims of these particular interventions, other evidence turned out be far less relevant. Either way, establishing the relevance of evidence was a necessary first step in the development of effective interventions.

As I have subsequently argued, policy makers can inquire upon evidential relevance through the construction and evaluation of causal scenarios. Interpreted in this sense, mechanisms provide a preliminary understanding of the causal relationships that are likely to be relevant for a given policy hypothesis. Each scenario starts with a proposed intervention and ends with the desired outcome, showing the intermediate causal processes along the way. Policy makers draw on different kinds of theoretical, experimental, and background knowledge in constructing these causal scenarios. Every scenario has to be plausible according to some basic theory, general principle, or commonly held opinion. For instance, the package bidding procedure during the FCC auctions was initially based on the principle of synergy. Similarly, behavioural insights such as hyperbolic discounting were used as a foundation for the construction of causal scenarios with respect to the SMT pension plan.

Interestingly, the interpretation of mechanisms as causal scenarios has important consequences for the problem of external validity. Given that the efficacy–effectiveness perspective is reversed, the task of extrapolating some efficacious result from one context to another becomes less pressing from a policy point of view. As for the issue of modularity, mechanisms do not necessarily have to reflect genuinely causal relationships; they may also reflect spurious relationships or mere correlations. However, interpreting mechanisms as causal

scenarios cannot completely resolve the issue of modularity, for no matter what kind of evidence policy makers intend to use, there has to be a stable connection between the policy and target variable. If there is no stability, then policy makers cannot know for sure whether their intervention is actually effective. Thus, the construction of scenarios requires stable relationships that need not be causal.

When constructing (causal) scenarios, policy makers are left with the question as to which evidence actually *is* relevant (and credible) for their policy objectives. There are often many plausible scenarios available that could justify an intervention. For this reason, it is very difficult for policy makers to sufficiently justify an intervention by invoking mechanisms. Nevertheless, even if a mechanism indicates a particular piece of evidence not to be relevant for the aim at hand, it will be useful for policy makers to be aware of reasons why their intervention might not be justified. This way, policy makers are cautioned against the premature implementation of their interventions. Ultimately, then, they have to make a judgment call with respect to the actual relevance of the evidence presented. This kind of judgement can be supported by the further differentiation of mechanisms, or so I have argued.

Finally, let me make two suggestions for future research with regard to the use of mechanisms for policy-making. First, the role of experiments – especially its 'testbed' version – in the design of mechanisms appears to be particularly interesting for policy makers. The success of the FCC auctions can largely be attributed to the clever use of experiments: the background conditions of the spectrum auctions were first imported into the laboratory, after which the results were used in the design of the auctions. This procedure combines extrapolation with inductive methods, which is a promising approach for policy makers given the complexity of the contexts in which they usually operate. One way to further improve this procedure would be to specify the type of background conditions that are to be incorporated into the experiments. Since the potential of experimental economics is naturally constrained by ethical as well as practical considerations, it needs to be clear when and in which contexts experiments are actually feasible.

Secondly, the interpretation of mechanisms as plausible causal scenarios needs to be conceptualised more rigourously. To illustrate, Cartwright refers to "causal scenarios" (2009b), "stories" (2009a) and "causal chains" (2012) in various pieces of her work. Despite these somewhat similar formulations, all but the general idea of causal scenarios is still pretty unclear. It is often mentioned as just one possible alternative to extrapolation, or merely as a supplement to randomised controlled trials. Instead, it would be fruitful to approach the idea of causal scenarios more explicitly by discussing, for instance, how these scenarios should be constructed

and in what phase of the policy-making process this can best be done.[51] To be fair, Cartwright and Hardie (2012: 175–178) do briefly discuss a couple of ways in which causal processes could be represented, each with its own advantages and drawbacks. The literature on mechanisms for policy-making would greatly benefit from more elaborate discussions in this regard.

[51] For example, Raoul Gervais and Erik Weber (2015) discuss the role of 'orientation experiments' in the discovery of mechanisms with respect to the natural sciences. They claim that "orientation experiments are a special type of intervention experiments used to provide evidence for or against a qualitative characterization of a mechanism" (ibid.: 47). In the so-called 'orientation-phase', "one or more mechanism sketches about the qualitative character of the mechanism responsible for the explanandum-phenomenon are proposed" and scientists subsequently "gather evidence for or against these mechanism sketches" (ibid.: 49). A similar inquiry with respect to the social sciences (especially economics) could prove useful for policy makers.

BIBLIOGRAPHY

Alexandrova, A. (2006). Connecting Economic Models to the Real World: Game Theory and the FCC Spectrum Auctions. *Philosophy of the Social Sciences*, *36*(2), 173–192.

Alexandrova, A. & Northcott, R. (2009). Progress in Economics: Lessons from the Spectrum Auctions. In H. Kincaid & D. Ross (Eds.), *The Oxford Handbook of Philosophy of Economics* (pp. 306–336). Oxford University Press.

Andrews, G. (1999). Efficacy, Effectiveness, and Efficiency in Mental Health Service Delivery. *Australian and New Zealand Journal of Psychiatry*, *33*, 316–322.

Berg, N. & Gigerenzer, G. (2010). As-if Behavioral Economics: Neoclassical Economics in Disguise? *History of Economic Ideas*, *18*, 133–166.

Cartwright, N. (2009a). Causality, Invariance, and Policy. In H. Kincaid & D. Ross (Eds.), *The Oxford Handbook of Philosophy of Economics* (pp. 410–423). Oxford University Press.

Cartwright, N. (2009b). Evidence-based Policy: What's to be Done About Relevance? *Philosophical Studies*, *143*(1), 127–136.

Cartwright, N. (2009c). What is This Thing Called "Efficacy"? In C. Mantzavinos (Ed.), *Philosophy of the Social Sciences* (pp. 185–206). Cambridge University Press.

Cartwright, N. (2012). Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps. *Philosophy of Science*, *79*(5), 973–989.

Cartwright, N. & Hardie, J. (2012). *Evidence-Based Policy*. Oxford University Press.

Clarke, B. et al. (2014). Mechanisms and the Evidence Hierarchy. *Topoi*, *33*(2), 339–360.

Coleman, J. S. (1986). Social Theory, Social Research, and a Theory of Action. *American Journal of Sociology*, *91*(6), 1309–1335.

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge: Belknap Press.

Cramton, P. C. (1998). The Efficiency of the FCC Spectrum Auctions. *Journal of Law and Economics*, *41*, 727–736.

Friedman, M. (1953/2008). The Methodology of Positive Economics. In D. Hausman (Ed.), *The Philosophy of Economics* (3rd ed., pp. 145–178). Cambridge University Press.

Gervais, R. & Weber, E. (2015). The Role of Orientation Experiments in Discovering Mechanisms. *Studies in History and Philosophy of Science*, *54*, 46–55.

Grüne-Yanoff, T. (2015). Why Behavioural Policy Needs Mechanistic Evidence. *Economics and Philosophy*, 1–21.

Guala, F. (2005). *The Methodology of Experimental Economics*. Cambridge University Press.

Guala, F. (2011). Building Economic Machines. In D. Steel & F. Guala (Eds.), *The Philosophy of Social Science Reader* (pp. 262–278). New York: Routledge.

Harari, D. (2015). *Greek Debt Crisis: Background and Developments in 2015*. *House of Commons Library*. Retrieved from http://www.parliament.uk/commons-library

Hayek, F. A. (1988). *The Fatal Conceit: The Errors of Socialism*. Routledge.

Hedström, P. & Swedberg, R. (1998). Social Mechanisms: An Introductory Essay. In P. Hedström & R. Swedberg (Eds.), *Social Mechanisms. An Analytical Approach to Social Theory* (pp. 1–30). Cambridge University Press.

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

Hempel, C. G. (1958). The Theoretician's Dilemma: A Study in the Logic of Theory Construction. *Minnesota Studies in the Philosophy of Science*, *2*, 173–226.

Hempel, C. G. (1942/2011). The Function of General Laws in History. In D. Steel & F. Guala (Eds.), *The Philosophy of Social Science Reader* (pp. 68–77). New York: Routledge.

Kahneman, D. & Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, *47*(2), 263–292.

Levitt, S. D. & List, J. A. (2009). Field Experiments in Economics: The Past, the Present, and the Future. *European Economic Review*, *53*(1), 1–18.

Little, D. (1991). *Varieties of Social Explanation*. Westview Press.

Little, D. (1995). Causal Explanation in the Social Sciences. *The Southern Journal of Philosophy*, *34*, 31–56.

Loomes, G. & Sugden, R. (1982). Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, *92*(368), 805–824.

Lucas, R. (1983). Econometric Policy Evaluation: A Critique. In K. Brunner & A. Meltzer (Eds.), *Theory, Policy, Institutions: Papers from the Carnegie-Rochester Conference Series on Public Policy* (pp. 257–284). Elsevier.

McAfee, P. & McMillan, J. (1996). Analyzing the Airwaves Auction. *Journal of Economic Perspectives*, *10*(1), 159–175.

Merler, S. (2015). *Greece's Hot Summer*. Brussels. Retrieved from http://bruegel.org/2015/05/greeces-hot-summer/

Merton, R. (1948). The Self-Fulfilling Prophecy. *The Antioch Review*, *8*(2), 193–210.

Mill, J. S. (1843). *A System of Logic*. London: Parker.

Nagatsu, M. (2015). Social Nudges: Their Mechanisms and Justification. *Review of Philosophy and Psychology*, *6*(3), 481–494.

Nik-Khah, E. (2008). A Tale of Two Auctions. *Journal of Institutional Economics*, *4*(1), 73–97.

Plott, C. R. (1996). *Laboratory Experimental Testbeds: Application to the PCS Auction* (No. 957).

Plott, C. R. (1997). Laboratory Experimental Testbeds: Application to the PCS Auction. *Journal of*

Economics and Management Strategy, *6*, 605–638.

Raiffa, H. (1982). *The Art and Science of Negotiation*. Cambridge, MA: Harvard University Press.

Reiss, J. (2008). *Error in Economics: Towards a more Evidence-based Methodology*. Routledge.

Roth, A. E. (2002). The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics. *Econometrica*, *70*(4), 1341–1378.

Scriven, M. (1962). Explanation, Predictions, and Laws. In H. Feigl & G. Maxwell (Eds.), *Scientific Explanation, Space, and Time: Minnesota Studies in the Philosophy of Science* (3rd ed., pp. 170–230). University of Minnesota Press.

Steel, D. (2008). *Across the Boundaries*. Oxford University Press.

Thaler, R. H. & Benartzi, S. (2004). Save More Tomorrow™: Using Bahavioral Economics to Increase Employee Saving. *Journal of Political Economy*, *112*, 164–187.

Thaler, R. H. & Benartzi, S. (2013). Behavioral Economics and the Retirement Savings Crisis. *Science Magazine*, *339*, 1152–3.

Thaler, R. H. & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth and Happiness*. Yale University Press.

Weber, M. (1905/1930). *The Protestant Ethic and the Spirit of Capitalism*. Routledge.

White, H. (2009). *Theory-Based Impact Evaluation: Principles and Practice* (No. 3). New Delhi.

Woodward, J. (2003). *Making Things Happen*. Oxford University Press.

Ylikoski, P. (forthcoming). Thinking with the Coleman Boat.