A PHILOSOPHICAL ANALYSIS OF THE CONCEPT OF AN EXTERNALITY IN ECONOMIC THEORY AND POLICY

by

REBECCA LIVERNOIS

B.A., University of Guelph, 2011
M.A., University of British Columbia, 2013

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Philosophy)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

A Philosophical Analysis of the Concept of an Externality in Economic Theory and Policy

submitted by Rebecca Livernois in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in Philosophy

Examining Committee:

Margaret Schabas, Philosophy
Supervisor

John Beatty, Philosophy
Supervisory Committee Member

Alison Wylie, Philosophy
Supervisory Committee Member

Matt Bedke, Philosophy
University Examiner

Brian Copeland, Economics
University Examiner

David Schmidtz, Philosophy
External Examiner

ABSTRACT

Economists generally understand externalities as unpriced spillover effects. The paradigmatic case is pollution because it is unpriced and affects agents external to the market choices that lead to its production. One solution to an externality is to set a tax on the unpriced activity at the value of the externality in equilibrium. The concept of an externality, however, is notoriously difficult to precisely define and there is a notable absence of consensus among economists.

In this dissertation, I offer an analysis of the contemporary treatment of externalities in economic theory and arrive at the following definition: Externalities arise when unpriced activities generate untapped gains from exchange that are associated with untapped welfare gains. It is unclear, however, whether this concept could be instantiated in any concrete sense because gains from exchange often diverge from welfare gains. I suggest possible ways to interpret an externality given this problem, but argue that each interpretation falls short of an adequate account of externalities.

The ambiguity of the concept of an externality carries over to attempts to estimate the value of a specific externality. I suggest that this accounts for some of the controversy among both economists and philosophers over one approach to estimating the value of an externality, called the contingent valuation method. Furthermore, this ambiguity renders problematic certain policies, such as the carbon tax, that are intended to internalize an externality.

I then argue that the problem of climate change is not merely caused by the presence of externalities, as some economists have suggested. I argue that, even if all externalities were eliminated, a social planner might still bring about a regretful environmental state when they aim to maximize net benefit derived from polluting activities. This is a result of the peculiar cost structure of climate change in which the marginal costs are uninformative of the total costs of polluting. I suggest that, instead of aiming to balance the costs and benefits of polluting, we might need to forgo some of the potential net benefits in order to avoid reaching an irreversible and regretful state.

LAY SUMMARY

Economists generally understand externalities as unpriced spillover effects. The paradigmatic case is pollution because it is unpriced and affects agents external to the market choices that lead to its production. The concept of an externality, however, is notoriously difficult to precisely define. In this dissertation, I develop a novel definition of externalities that coheres with the contemporary treatment of externalities in economic theory. I then argue that there are several distinct interpretations of externalities in the world that are consistent with this definition. As a consequence of this ambiguity, I suggest that policies that aim to internalize an externality are not well supported. Finally, I argue that the problem of climate change is not merely the result of the presence of externalities. If all externalities were resolved, the peculiar cost structure associated with climate change could still create an incentive for a community to pollute beyond the socially optimal level.

PREFACE

This dissertation is the original and independent work of Rebecca Livernois. A slightly modified version of chapter five is originally published as Livernois, Rebecca. 2018. "Regretful Decisions and Climate Change." *Philosophy of the Social Sciences* 48, no. 2: 168-191. Copyright © 2017 (Rebecca Livernois) https://doi.org/10.1177/0048393117741335.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Margaret Schabas for the mentorship and support she generously offered over the past five years. Her feedback and suggestions helped to form my ideas and shape this dissertation. I would also like to thank Alison Wylie and John Beatty for their valued advice and insightful comments.

The feedback I received from audiences at several conferences helped me to refine my arguments, including the congresses of the Canadian Philosophical Association, the Canadian Society for History and Philosophy of Science, and the Philosophy of Science Association, as well as the CIRED Workshop, the University of Washington Graduate Conference, the INET Young Scholars Initiative, the Philosophy of Social Science Roundtable, and the Athena in Action Graduate Workshop.

I am grateful for the feedback offered by Chris Stephens and Daniel Steel on work that turned into chapter five, and the guidance offered by Murat Aydede on work that was incorporated into chapter three. Fellow graduate students at UBC also offered helpful feedback and commitment mechanisms, including Sina Fazelpour, Servaas van der Berg, Matthew Smithdeal, Jelena Markovic, Kousaku Yui, Madeleine Ransom, Richard Sandlin, Jasper Heaton, and Alastair Fraser.

I would also like to thank Mike Hoy and Louise Grogan for encouraging my interest in the philosophy of economics many years ago, and Stefan Linquist for creating a field course in Tofino, British Columbia, in which I discovered the problems I was most interested in trying to understand.

I am indebted to both my parents for spending many evenings taking seriously my early questions about economics. Listening to my mother Brenda Dyack's experiences using the contingent valuation method in her role as an economic policy advisor played a large role in the development of my ideas in chapter four of this dissertation. My father, John, instilled in me a love of nature and tackling complex problems, a fruitful combination. My sister, Ali, has consistently set a high bar for me to strive towards. My grandmother, Audrey, demonstrated the joy of building a meaningful career. My partner, Nouri Najjar, was an expert sounding board for my ideas in this dissertation. He kept me grounded and offered love and support through this process, for which I am forever grateful.

I am thankful for the financial support I received from the Social Sciences and Humanities Research Council of Canada, the Killam Foundation, and the University of British Columbia.

for Nouri

CHAPTER ONE

Introduction

A statement of consensus among economists on the ideal policy response to climate change was recently issued in the *Wall Street Journal* (Akerlof et al. 2019). This statement was signed by forty-five eminent economists, twenty-seven of whom are recipients of the Nobel Memorial Prize in Economic Sciences. They state that, "guided by sound economic principles," they are united in recommending a carbon tax to address the climate change problem. This is because it "offers the most cost-effective lever to reduce carbon emissions at the scale and speed that is necessary. By correcting a well-known market failure, a carbon tax will send a powerful price signal that harnesses the invisible hand of the marketplace to steer economic actors towards a low-carbon future" (Akerlof et al. 2019).

This statement of consensus summarizes the received view of this "well-known market failure," purportedly caused by the presence of an externality (Stavins and Hahn 1992; Mas Colell et al. 1995; Stavins 2011). An externality in economics is broadly understood as an unpriced spillover effect. Greenhouse gas emissions generate a negative externality because they are unpriced and have a harmful effect on individuals external to the market decisions that create the emissions. Such unpriced activities are produced

beyond their optimal level because these activities are treated as costless when there is, prima facie, good reason to suppose that they are costly. This is to say that climate change is partly caused by the overproduction of greenhouse gas emissions, and this overproduction is the result, first and foremost, of the absence of a market price on these emissions.

Economic theory stipulates that imposing a tax on the unpriced activity at the "correct" price induces economic actors to produce the activity at the optimal level. The correct price of the activity is the monetary value of the externality in equilibrium. Hence, setting a marginal tax rate on an unpriced activity at the value of the externality in equilibrium will resolve an externality problem. Therefore, when emitting greenhouse gases is a priced activity, because it is taxed at the correct level, producers are incentivized to generate the optimal quantity of these emissions. The ideal policy response to climate change thus appears to be remarkably simple. Climate change is an externality problem; therefore, imposing a carbon tax that eliminates the externality will adequately address this problem.

In this dissertation, I aim to show that this apparent simplicity is, in fact, illusory. In the following three chapters, I argue that the concept of an externality is ambiguous. I first develop a characterization of externalities called the gains view of externalities. I argue that this characterization is more consistent with the way externalities are treated in prominent economic models than the typical view of externalities in economics. I then show that there are several ways to interpret an externality in the world that are consistent with the gains view. On one interpretation, greenhouse gas emissions do not

generate an externality at all. On another, these emissions generate an externality that is unmeasurable. To assign a tax that optimally "internalizes" an externality, however, there must be a measurable externality that determines the tax rate. Therefore, the concept of an externality in economics is ambiguous, and this ambiguity is not merely theoretical; it impedes efforts by economists to accurately measure externalities. As a result, I mean to show that carbon taxes that are intended to internalize an externality, or to resolve a market failure, cannot be fully justified.

I then argue that climate change is not merely an externality problem, regardless of how externalities are interpreted. Even if all externalities were eliminated, there would still be an incentive for a social planner to over pollute. This incentive is caused by the inherent uncertainty of climate change coupled with its peculiar cost structure in which the marginal costs of polluting are uninformative of the total costs. Economists could thus better serve policymakers by addressing the ambiguity of their policy recommendations and by addressing the additional challenges the climate change problem presents for policymakers.

Although this dissertation has implications for prominent policy responses to environmental problems, it is primarily a philosophical analysis of the concept of an externality in economics. As we will see, an externality is an important concept in economics because it is a foundational concept that establishes why markets can fail to allocate goods efficiently; consequently, this form of market failure establishes the conditions that call for market-based policy interventions into markets. The way an externality is defined thus determines the reach of markets into social life, as

conceptualized in economics. If the concept of an externality were more sharply drawn and understood, it could serve to distinguish the line between market and non-market activities, and the sense in which a policy intervention that expands the inclusion of non-market activities into the market is well motivated.

In this dissertation, I use environmental problems as the paradigmatic example of externalities. Externalities are not limited to environment problems, however. The choice of one individual to smoke a cigarette in public generates a negative externality because this action imposes an unpriced cost on others; one component of this cost includes the discomfort and negative health consequences of others caused by the second-hand smoke. The use of a road by one person imposes an unpriced spillover effect of increased travel time on others, because their use of the road increases the traffic on the road. The creation of knowledge that circulates freely by one individual generates a positive externality because it creates positive effects on others who benefit from this knowledge despite not having to pay for it.

Environmental problems are particularly interesting instances of externality problems, however, for at least two reasons. First, climate change is arguably the most prominent contemporary policy issue that is treated as an externality problem by economists. Indeed, policy instruments such as a carbon tax or a cap-and-trade system, which are based on the theory of externalities in economics, are used by dozens of nations across the globe (World Bank Group 2018). For example, the European Union established a carbon trading scheme in 2005 (European Commission n.d.). A cap was set on the total quantity of some types of greenhouse gas emissions and this quantity has

reduced over time. Polluters can trade their allowances, and they are fined for exceeding the allowances they own. The Government of Canada, on the other hand, recently implemented a nation-wide carbon tax. In line with the recommendations made by Akerlof et al. (2019) about the desirable features of a carbon tax, the carbon tax rate in Canada is set to increase gradually and the tax revenue will be returned to households as lump-sum rebates to improve the fairness of the tax (Government of Canada 2017).

It should be noted at the outset that I do not intend to make any claims about how governments create policies in practice. That is, I do not claim that governments are in fact implementing carbon taxes with the intention of internalizing an externality, as economic theory suggests. Taxes are often used to induce consumers or producers to reduce their production or consumption of a good or activity independent of any claim about the presence of any externalities; that is, considerations other than the presence and size of an externality may both justify a tax or figure in determining a tax rate. Nonetheless, economists generally conceive of the carbon tax as a Pigovian tax, which is the type of tax that internalizes an externality. Indeed, economists recommend a carbon tax based on the view that it is a Pigovian tax, and therefore that it corrects an externality problem. Consequently, we ought to have a firm grasp of what an externality problem is to adequately understand carbon taxes and to assess how they should be designed.

The second reason that environmental problems are particularly interesting instances of externality problems is that an externality is, in some instances, conceived of as both that which is external to the market system or that which ought to be brought into the market system (Berta and Bertrand 2014). Insofar as the natural environment

can be interpreted as a large category of entities that are external to the market system, interpreting these entities as externalities thus implies that entities in the natural environment ought to be incorporated into the market system. In short, it contains both a descriptive and a normative set of assertions. But entities in the natural environment tend to have distinct properties from typical consumer goods that serve as the primary case for mainstream economic theory. Ecosystems, for example, are not typically introduced into consumption bundles or indifference maps (Spash 2008). Indeed, environmental ethicists have argued that environmental entities like ecosystems are valuable in and of themselves, as ends and not just means. Consumer goods, by contrast, acquire value because they are means to some further end (Jamieson 2001). The use of economic principles to explain and propose solutions to environmental problems thus offers a fruitful case study for unpacking assumptions of economic theory and market-based policies in a domain that is not obviously appropriate for economic theorizing.

Economic theory stipulates that the presence of an externality is a form of market failure. A market fails insofar as it fails to obtain an allocation of resources that is Pareto optimal, which is to say that it is inefficient. A market is efficient, or provides a Pareto optimal allocation of resources, when no Pareto improvements could be made; this means that no one could be made better off without making someone else worse off (Pareto [1909] 1971). According to the first law of welfare economics, in a perfectly competitive market, if the allocation of resources attains an equilibrium, then it is a Pareto optimal outcome. Any move from this equilibrium would make at least one person worse off (Blaug 1962). By the second law of welfare economics, any Pareto

optimal outcome can be attained in a perfectly competitive market given transfers of initial endowments. These laws require that several conditions hold, including the condition that there are no externalities, no asymmetries in information, and no barriers to entry or exit from any market (Mas-Colell et al. 1995). When one or more of these conditions are not met, a perfectly competitive market might fail to reach a Pareto optimal outcome. Hence the presence of an externality is considered a form of market failure because it causes an inefficient allocation of resources, which means that there exists space for Pareto improvements at the market equilibrium.

According to economic theory, efficiency is a desirable feature of markets because it indicates that scarce resources are used in a way that creates the most economic value. Economic value is a function of preference satisfaction; the most economic value is created when the satisfaction of preferences is maximized (Hausman and McPherson 2006; Colander 2015). Resources are squandered when they are not optimally allocated because, given the same limited resources, a reallocation of resources could result in the satisfaction of more preferences. For example, if a firm does not produce at its lowest cost, then it could have produced more goods with the same inputs at the same costs, thus generating more economic value with the same resources.

The quantity of goods produced in a perfectly competitive market is optimal when all the costs and benefits of production are included in the selling and buying decisions of economic agents. That is, resources are directed to their highest-value use when producers face the full costs and benefits of these resources. If the pollution created in the production of a good generates a cost that firms do not face in their

production decisions, for example, then these firms treat pollution as free when there is good reason to not regard pollution as free. This, in turn, causes these firms to misuse a valuable resource. If producers instead faced the full cost of pollution, then the production of pollution would be directed to its highest-value use. For example, the most economic value might by created by directing costly pollution to the production of solar cells, rather than to producing coal-powered electricity. More energy could be created in the long-run, and therefore more preferences for energy consumption could be satisfied, from the same quantity of pollution directed to the former use, as opposed to the latter use.

According to environmental economics, many environmental problems exist because the costs of consumption and production by-products, such as carbon dioxide emissions, are not shouldered by the consumers or producers that generate these by-products. That is, pollution is a negative externality both because it is unpriced and because it harms agents external to the action that generates the pollution (Hahn and Stavins 1992; Stavins 2011). Markets fail to reach an optimal allocation of resources when an externality is present. Therefore, environmental problems are market failures caused by externalities; in a market for a pollution-generating good, the allocation of resources attained at the competitive equilibrium is not Pareto optimal.

The policy solution to market failure caused by an externality is to internalize the externality through government intervention into the market. If pollution causes a market failure because the costs of pollution are not faced by producers or consumers, then a price can be set on pollution that accurately reflects its costs. The cost functions

of producers would then properly reflect the true costs of production and the prices of pollution-generating goods faced by consumers would also reflect the true costs of the goods. This can be achieved by setting a so-called Pigovian tax on pollution at the value of the externality at the socially optimal level of pollution (Weitzman 1974; Pearce 2002; Stavins 2011).

In theory, setting an optimal tax on an externality-generating activity leads to an efficient outcome because all the costs of the activity are taken into account in the buying and selling decisions of agents. Alternatively, the government can create a market for pollution permits. A cap and trade scheme, for example, sets a limit on the total quantity of pollution that can be emitted by a given industry, distributes shares of the total quantity of pollution among firms in the form of pollution permits, and then allows firms to buy and sell these permits. In theory, both a Pigovian tax and a cap and trade system can be used to create incentives for economic actors to generate pollution at the optimal level.

A policy that directly regulates how much each firm can emit, called a command-and-control policy in economics, can also be used to restrict pollution to the optimal level. A Pigovian tax or a cap and trade scheme are less costly than a command-and-control policy, however. This is because both a Pigovian tax and a cap and trade system direct pollution reductions to the firms who face the lowest costs for reducing pollution. For example, when pollution is taxed, firms decide how much they will pollute by considering their marginal costs of reducing their emissions. The firms with the highest marginal abatement costs will pollute the most, while a large quantity of pollution

reductions will come from firms who can inexpensively reduce their emissions to reduce their tax burden. Firms determine their optimal level of pollution by equating their marginal abatement costs with the marginal tax rate; therefore, firms with low marginal abatement costs will reduce pollution relatively more than firms with the high marginal abatement costs.

A command-and-control policy that regulates how much each firm can pollute is a costly way to reduce pollution because it does not distinguish between the differing cost structures of firms. A firm with low marginal abatement costs does not have an incentive to inexpensively and significantly reduce their emissions because the regulation requires it to reduce emissions to a predetermined level that is common across all firms in an industry. Therefore, this policy is more costly than market-based policies because it does not prioritize the least-cost emissions reductions. Furthermore, Pigovian taxes and cap and trade schemes generate government revenue that can be returned to polluters to offset the costs of the policy. Economists thus argue that Pigovian tax and cap and trade scheme are the most cost-effective methods for internalizing an externality.

The concept of an externality thus underlies dominant policies that address critical social and environmental problems like climate change. Even so, externalities are notoriously difficult to precisely define in economics. Typically, an externality is understood as an untraded spillover effect in economics; I call this *the broad view* of externalities. But as I will argue in chapter two, this view of externalities is inadequate because externalities are supposed to be policy-relevant, yet untraded spillover effects are everywhere in the social world. The broad view therefore does not provide justification

for any specific policy. This characterization of externalities is therefore too general for its intended purpose. Although a prominent concept in economics, the concept of an externality remains notoriously difficult to define (Mas-Colell et al 1995, 351; Papandreou 1994).

Alfred Marshall ([1890] 1961) developed the notion of external effects, a rudimentary version of the concept of an externality. An external effect for Marshall is any market activity by one firm that affects another without any voluntary interaction between the two firms. For example, economic development causes an increase in the cost of inputs because the demand for inputs increases as more firms enter the market. Hence an external effect changes a firm's costs. This is a spillover effect per se, but not the kind economists typically mean by *externality*. An externality is typically understood as an untraded spillover effect that is unmediated by market prices, but Marshall's external effects are mediated by prices. Marshall's notion of external effects is thus closer to what economists now call *pecuniary externalities*, which occur when the actions of some agents lead to changing prices that affect others. These are not the kind of externalities that economists are typically interested in because they are effects that are internal to markets (D. Hausman 1992a). Pecuniary externalities, for example, have zero net effects because an increase in the price of a good is a negative pecuniary externality for buyers but a positive pecuniary externality for sellers. This type of externality therefore does not cause Pareto inefficiencies.

Arthur Cecil Pigou ([1920] 2017) was instead concerned with external effects insofar as they reduce economic welfare. He understood national dividend to

approximate economic welfare, which approximates total welfare (Caldari and Masini 2011). Externalities cause a divergence between marginal private net product and marginal social net product, where the former includes costs and benefits that are faced by agents making market decisions and the latter includes all the costs and benefits realized in society at large. In theory, government intervention is justified when externalities are present. A tax or subsidy, if used properly, increases welfare by aligning marginal private and marginal social net product, therefore mitigating the effect of the externality and maximizing the national dividend. The contemporary understanding of a tax that internalizes an externality thus descends from Pigou; externalities cause a divergence in marginal social cost and marginal private cost and a Pigovian tax equalizes the two cost curves, thus resulting in the socially optimal level of production.

Ronald Coase ([1960] 2013) objected to Pigou's conception of an externality because it ignores the reciprocal nature of an externality (Medema 2009; Medema 2014). Pollution generated by A imposes an external cost on B, but if B were to prohibit A from polluting, then B would be imposing an external cost on A. Using a tax to internalize an externality, as Pigou proposed, thus favours one party over another without considering the effects of this tax on economic value. For example, a paper mill that is located upstream releases effluents into the river, which reduces the profits of a resort that is located downstream. A tax on the release of effluents would increase the profits of the resort at the expense of the paper mill. On the other hand, if property rights over the river were defined, the two firms would bargain over the release of effluents into the river. In doing so, economic value would be maximized because the

agent that could reap the highest economic value from using the river would pay the other for its use. When property rights are defined and there are no transaction costs, no externality will persist and economic value will be maximized because agents will bargain over the cost or benefit of the externality.

Hence, according to Coase, externalities persist because transaction costs are rarely negligible and property rights are often ill-defined. Coase's message, therefore, was not that bargaining will always eliminate externalities. Instead, he argued that a comparative institutional approach is necessary to determine the solution to an externality that maximizes economic value. That is, we must compare the effects on economic value of government intervention, market solutions, letting the externality persist, and having a single owner over the externality, in order to determine the best course of action in response to an externality.

Pigou and Coase agreed, however, that the government may be incapable of improving on market outcomes even in the presence of externalities (Aslanbeigui and Medema 1998). They agreed that we should be wary of giving government officials power to alter markets especially since they are subject to political pressure without competitive checks that exist in markets. Pigou and Coase also agreed that the administrative costs of taxes, subsidies, and regulations are often high and involve waste. Therefore, the costs and benefits of implementing the policy, including the distortions in markets caused by taxes, must be assessed before it can be claimed that a government should intervene in markets. Pigou and Coase differed, however, in their optimism about the abilities of governments to effectively implement policies that improve on market

outcomes. Pigou thought that the tendency of governments to appoint experts to assess policies is promising, and he did not think the distortions in markets caused by government intervention are particularly costly. Thus, he thought that government interventions in markets were often beneficial. Coase, on the other hand, thought that effective government interventions into markets were significantly impeded both by a lack of information held by government officials and high administrative costs. He had more confidence in markets than governments to maximize economic value (Aslanbeigui and Medema 1998).

James M. Buchanan (1962) similarly had a pessimistic view of the abilities of governments to improve on market outcomes. He argued that we can only claim that the state can effectively remove externalities when we inconsistently claim that all people, except policymakers, are self-interested; this is because only the altruistic policymaker faces the right incentives to develop effective market interventions (Marciano 2011, 246). For Buchanan, externalities exist because there are interdependencies between individuals. However, he argued that not all externalities ought to be internalized; instead, it is up to individuals to decide if an externality problem warrants attention (Marciano 2011, 244). If an externality is a large enough problem, then affected agents would exchange over the relevant externality and consequently adjust its production to the optimal level (Buchanan and Stubblebine 1962). If individuals are not willing to engage in exchange to eliminate an externality, then it must be the case that the externality is negligible, or irrelevant, because individuals do not expect to gain from

exchange over the externality (Marciano 2011). The persistence of irrelevant externalities is therefore consistent with an efficient allocation of goods, according to Buchanan.

This is an unsatisfactory understanding of an externality, however, because many externalities persist despite being relevant. Pollution, for example, is a relevant externality that cannot be eliminated merely by bargaining between individuals because the multi-agent nature of the externality precludes such a solution. Buchanan's view of an externality is therefore similar to Coase's view because both held that bargaining tends to remove an externality; however, Buchanan surpassed Coase in his optimism about the efficacy of bargaining solutions. Buchanan thought that the only relevant externalities are those that are eliminated by bargaining while Coase acknowledged that high transaction costs impede a bargaining solution, and thus result in persistent and relevant externalities.

There are many more ways an externality has been characterized in economics. Harold Demsetz (1996) follows Coase in defining externalities as the absence of properly defined property rights and high transaction costs. Kenneth Arrow (1969) describes externalities as a type of missing market. Similarly, Walter Heller and David Starrett (1976, 10) "define externalities to be a situation in which the private economy lacks sufficient incentives to create a potential market in some good and the nonexistence of this market results in losses in Pareto efficiency." On the other hand, James Meade (1952) understood externalities as unpaid factors of production. Don Fullerton and Robert Stavins (1998) instead interpret externalities as scenarios where some consequences of producing or consuming a good are external to the market, meaning that they are not considered by producers or consumers. The definition that has been

crystalized in textbooks holds that an externality occurs when an objective function depends on the unintended by-products of another's activities (Mas-Colell et al. 1995; Gaus 2008; Perman et al. 2013).

The differences in these definitions of externalities are nontrivial. For example, an externality understood as an unpaid factor of production limits its relevance to production byproducts and production decisions. An externality understood as interdependent utility functions instead extends its relevance into the social world. The social world is rife with effects between agents that are not mediated by prices, and therefore which could be modeled as interdependent utility functions. It is therefore unclear which unpriced spillover effects count as externalities.

Maurice Lagueux (2010) offers a detailed historical account of the changing understanding of externality. He argues that the best way to accommodate the various definitions of an externality in economics is to understand an externality as a residual concept in economics. It is something which is not the market. Nathalie Berta and Elodie Bertrand (2014) build on this account by showing that the conceptions of an externality offered by Coase and Arrow follow from the institutional framework they employ in their analysis. Hence, the way an externality is conceptualized depends on the definition of a market. The reason there are different types of externalities therefore stems from the variety of market frameworks, as understood in economic theory.

Berta and Bertrand (2014) thus argue that according to Coase, who employs a bargaining framework, an externality is an effect without property rights that can be internalized through negotiation. For Arrow, who employs competitive market

framework, an externality is an effect without parametric prices. They claim that these are distinct conceptions of externalities, which are unified only by the residual character of both conceptions. However, it is unclear why this is the case. The type of externality Arrow discusses is one that affects many people—a multilateral externality—whereas the type of externality that Coase discusses is one that occurs between two agents—a bilateral externality (Mas-Colell et al. 1995, 364). The perfectly competitive market model is appropriate to explain the former, while the bargaining model is appropriate to explain the latter. It does not seem that the conception of an externality itself is necessarily distinct in the two models; the difference is in terms of the number of agents affected by the externality.

More importantly, however, defining externality as a residual does little to clarify the concept of an externality. It does not specify which costs or benefits that are external to markets constitute an externality. If we want to internalize an externality, we need to know which costs we should measure, not just that they are costs that are external to markets. Even if an externality is indeed a residual concept, it is important to understand the components of this residual term.

Andreas Papandreou (1994, 281) attempts to clarify the "plural concepts that underlie the apparently homogenous notion of externality." He argues that there are two approaches economists have taken when analyzing externalities: the phenomenological approach and the general-equilibrium approach. The phenomenological approach holds that an externality is defined by an interdependence between the utility functions of agents. William Baumol and Wallace Oates (1975), for example, characterized an

externality in terms of an interdependence of agents, and this interdependence was characterized in terms of an interdependence between the utility functions of the agents as well as the degree of control held by an agent over the unpriced spillover effect. Papandreou suggests that this approach attempts to characterize an externality independently of its institutional context. He argues, however, that it is unsuccessful in this attempt. By attempting to define the degree of control held by an agent over the unpriced activity, the account devolves into a question of whether property rights are sufficiently defined. This means that this definition of an externality must address the institutional context of any given externality. On the contrary, he argues that the general equilibrium approach primarily aims to clarify the notion of market failure. The concept of an externality is employed, but not sufficiently clarified, in this analysis.

Papandreou (1994, 281) concludes, similarly to Lagueux (2010) and co-authors Berta and Bertrand (2014), that "there cannot be a unique good characterization of externality. Externality has come to denote many things, none of which separately, or in combination, seem to justify the appellation "externality", and more importantly, none of which do full justice to the important issues underlying this notion." It seems, then, that a single overarching characterization of the concept of an externality cannot capture its various historical usages. Accordingly, in this dissertation, I characterize the concept of an externality as it is treated in two prominent models of externalities in contemporary mainstream microeconomic theory that are most relevant to economic responses to environmental problems; I do not attempt to account for the many ways the concept of an externality has been employed in economics.

D. Hausman (1992a) argues that economists generally hold that externalities are unintended effects of A on B to which A and B have not both consented. He does not attempt to clarify the concept further; instead, he argues that welfare economics is mistaken in taking the existence of an externality to be an exceptional phenomenon. He argues that all aspects of social interaction involve unintended and untraded spillover effects. Therefore, economists should take the presence of an externality to be the rule rather than a rare occurrence that justifies a policy intervention. D. Hausman argues that economists ignore some of the most significant externalities by limiting their focus to nonpecuniary externalities. According to D. Hausman, pecuniary externalities are significant because they often cause more harm than non-pecuniary externalities. For example, competition creates pecuniary externalities by driving some firms out of business, which imposes significant harm on business owners and employees. This type of externality concerns justice rather than efficiency; he argues that economists should expand their understanding of an externality to include all significant harms relevant to markets. In doing so, economists could better analyze distributional effects of markets rather than merely inefficiency effects.

D. Hausman also argues that since externalities are the rule, not the exception, they are not something from which we can abstract when modelling markets. As such, it is a mistake to define the ideal market as one in which there are no externalities. He thus questions the relevance of the theory of market failure because it judges how well markets fare in comparison to perfectly competitive markets devoid of externalities, but this ideal is impossible in reality (D. Hausman 2008).

D. Hausman appears to take negative externalities to be synonymous with harms. Externalities matter because they are harms that go uncompensated. Thus, he argues that economists should not ignore the harms of injustice. It is not clear, however, that it is correct to interpret externalities as harms. One reason to think externalities are not merely harms is that being compensated for an externality does not necessarily eliminate the harm that is caused by the externality, although it does eliminate the externality. Eliminating the externality generated by pollution involves reducing the quantity pollution to its efficient level; but there is still likely a positive quantity of pollution at the efficient level that causes harm to some individuals. Internalizing an externality involves balancing the costs and benefits of the untraded activity, which will reduce but not entirely remove harms that are caused by the activity. Therefore, a negative externality is not merely a harm. If externalities are not harms, then it does not follow that the understanding of externalities should be expanded to include a broader class of harms. It might be the case that pecuniary externalities are important for social policy, but it does not follow that they should be included in the economic understanding of policy-relevant externalities.

Debra Satz (2012, 31-32) discusses externalities and market failure in her book *Why Some Things Should Not Be for Sale* to motivate her argument that there are moral limits to markets that cannot be explained by the theory of market failure. She claims that externalities are everywhere in the social world and thus economic theory implies that every interaction ought to be mediated by markets. She suggests that externalities are used selectively by economists—pollution counts as an externality while intolerance of

religious diversity does not—but there is nothing internal to the notion of externality that limits its reach into all aspects of society. While D. Hausman argues that the notion of an externality should expand to include harms caused by injustice, Satz implies that the notion of an externality expands too far into the social world. There are some things that should not be mediated by markets or conceived of as market goods. Furthermore, she claims that the interpretation of an externality in economics "feeds off moral theory done elsewhere"—in particular, the harm principle in liberal theory (Satz 2012, 32). This is an interesting claim that Claassen (2016) expands, explained below.

Satz's goal in this book is to develop criteria for the moral limits of markets, which she defines in terms of noxious markets. Noxious markets are assessed according to the values of weak agency, vulnerability, extreme harms to individuals, or extreme harms to society. The absence of a noxious market should not be considered a market failure because it is a case where markets should not exist in the first place. She claims that "the economists' generic view of externality is not fine-grained enough to allow us to distinguish the markets that score high on one or more of these parameters from other markets with third-party effects" (Satz 2012, 208). Satz thus attempts to define moral limits on markets. It is plausible that the theory of market failure should be accompanied by a moral argument about the moral limits of markets; nonetheless, it is not obvious that the concept of an externality implies that all interactions should be mediated by markets. Indeed, I will argue in this dissertation that not all unpriced spillover effects are externalities, and therefore that externalities do not necessarily proliferate in the social world. This means that economic theory does not imply that all

unpriced spillover effects should be addressed by economic interventions or mediated by markets.

Michael Sandel (2012) argues that markets are merely one way, and not always the best way, of organizing society. Sometimes, market organization is effective at allocating goods; other times, it erodes social norms that are more effective than markets at organizing society. For instance, trading carbon permits in a market implies that carbon emissions are a commodity. This might have the counterproductive effect of eroding the social norms that initially were successful in allocating this good (Bicchieri 2016). That is, establishing carbon emissions as a consumer good that can be bought and sold in a market might cause individuals to pollute more than if emitting carbon dioxide were merely frowned upon. Markets may erode the regulating nature of social norms and therefore lead to a worse social outcome (Nyborg 2000; Sandel 2012).

Rutger Claassen (2016) builds on Satz's claim that economists tend to use the concept of an externality opportunistically. He claims that nothing internal to the concept differentiates the policy-relevant externalities from the policy-irrelevant externalities. In line with D. Hausman (1992a), he argues that the problem with the concept is that it ignores harms that involve justice and freedom, yet it is these harms that typically call for state intervention. He argues that a harms principle, grounded in a theory of basic human interest, ought to be incorporated into the notion of an externality. It is not clear, however, that it is possible to incorporate Claassen's suggested interpretation of externality into economic theory. He argues that harms beyond the frustration of preference satisfaction should be incorporated into the notion of an

externality; however, there is no way to represent harm in economic theory other than in terms of preference satisfaction. Claassen's understanding of externality may be useful for analyzing government intervention in general; however, it does not help to elucidate the nature of externality in economics. Furthermore, I will argue in chapter three that the concept does, indeed, have an internal mechanism that distinguishes between policy-relevant and policy-irrelevant externalities.

Philosophical work on the application of economic theory to environmental problems has, for the most part, been more plentiful than philosophical work on the concept of an externality in economics. As I will show in chapter four, this work has important implications for assessing whether environmental problems can be interpreted as externality problems. Mark Sagoff (2004; 2007; 2010) argues that it is inappropriate to interpret environmental problems in an economic framework. He argues that social policy is a matter of ethics, not economics, and that environmental problems are a matter of social policy. Therefore, economics should not be used to explain and address environmental problems. This argument is grounded in his distinction between the consumer and the citizen. He argues that economic theory is only applicable to an individual in their role as a consumer who has self-interested desires. It distorts the values of an individual in their role as a citizen. That is, measuring value by one's willingness to pay is relevant to consumer values, not citizen values. Therefore, a measure of willingness to pay cannot capture the true value of the environment, which is a citizen value.

Hausman and McPherson (2006) similarly argue that environmental economics cannot capture the various ways in which the environment can be valued. In estimating the costs of pollution, an individual's willingness to pay for a reduction in pollution obscures other types of values for the environment she might have, such as values that are grounded in her moral principles. Willingness to pay is therefore an imprecise measure of well-being (Hausman and McPherson 2006, 286). Furthermore, welfare economics ignores distributional concerns, which are important to environmental policy. Hausman and McPherson do not argue, however, that these problems imply that economics is irrelevant to environmental policy. Instead, they maintain that using economic methods to estimate the costs and benefits of policy options is the best available method to determine how to balance the costs and benefits of pollution. They warn, however, that the virtues of quantification should not be exaggerated; the estimates of costs and benefits should be accompanied by moral deliberation in order to set policy that meets our social goals (Hausman and McPherson 2006, 286).

The use of cost-benefit analysis to adjudicate between competing environmental policies has received significant attention in the philosophical literature (Jamieson 2001). Cost-benefit analysis is a decision tool that is used to compare various resource-use options. If the total benefits of a given policy outweigh its total costs, then this policy passes a cost-benefit test. This means that undertaking this policy is at least worthwhile. Cost-benefit analysis, however, is primarily used to assess which of several competing policy options maximizes net benefit (total benefits net of total costs); therefore, it establishes which policy option is the most beneficial. John Broome (2012) argues that

cost-benefit analysis faces several limitations, including its inability to capture the full costs of climate change in monetary terms. For example, the total cost associated with the deaths caused by climate change cannot be captured monetarily. Nonetheless, he argues that cost-benefit analysis is a necessary decision tool when assessing climate change policy because it facilitates an assessment of the tradeoffs that are inherent to any climate policy option.

On the contrary, Stephen Gardiner (2011) argues that cost-benefit analysis is incapable at providing guidance on climate policy. One reason for this is that an assessment of the costs and benefits of climate change necessitates the impossible task of valuing costs and benefits that accrue to future generations. Estimating the present value of future costs and benefits requires the use of a social discount rate. A discount rate represents the idea that one dollar is more valuable to an individual now, for example, than it is a year from now. Economists typically understand the discount rate to be revealed by the interest rate because individuals will save insofar as they are sufficiently rewarded for saving, via interest. A social discount rate represents the value individuals place on the welfare of future generations. A social discount rate of zero, for example, places equal value on all generations.

Economists such as William Nordhaus (2014; 2017) argue that society's future consumption should be discounted according to the interest rate when estimating the social cost of carbon, which is the term for the externality generated by carbon dioxide emissions. Other economists such Nicholas Stern (2006; 2014a; 2014b), however, argue that discounting in this way is inappropriate when estimating the social cost of carbon.

Since climate change leads to lost lives, for example, and future lives are equally valuable to current lives, we ought not discount future consumption as heavily as the interest rate indicates (Broome 2012; Stern 2014a; 2014b). Stern (2006) and Nordhaus (2014; 2017) estimate significantly different values for the social cost of carbon, largely because of they employ different social discount rates (Anthoff and Tol 2013).

It is not obvious that that we should discount future welfare at all (Caney 2014; Gardiner 2011; Tarsney 2017). Accordingly, there has been significant debate among economists and philosophers over the social discount rate that should be used when estimating the social cost of carbon (Hansen 2011; Montuschi 2014; Parker 2014). Gardiner (2011) argues that setting the correct social discount rate is a near-impossible task; since one must choose a social discount rate to conduct a cost-benefit analysis for climate change, cost-benefit analysis cannot offer adequate guidance for climate policy. Indeed, Daniel Steel (2015) argues that it is impossible to estimate the expected costs and benefits of climate change to the degree of precision required by cost benefit analysis, and as a result this decision tool leads to unjustified delay in addressing the climate change problem.

The problem of settling on the correct social discount rate is indeed important because it is required to estimate the magnitude of the externality generated by greenhouse gas emissions. That is, the optimal carbon tax rate depends on the choice of the social discount rate (Broome 2012; Stern 2014b; Nordhaus 2017). This is an important issue in assessing the measurability of the externality generated by greenhouse

gas emissions; in this dissertation, however, I focus on the more basic issue of assessing how to interpret an externality even when this complex issue is set aside.

Accordingly, in the next chapter I assess the received view of externalities in economics. I argue that the view of externalities as untraded spillover effects is inadequate because externalities are supposed to be policy-relevant, yet untraded spillover effects are everywhere in the social world. This view is therefore too general for its intended purpose. By examining the prominent models of externalities in contemporary mainstream microeconomic theory, I develop an alternative characterization of externalities that I call the gains view of externalities.

I argue that untraded spillover effects generate two key outcomes in a two-agent constrained optimization framework. First, an externality occurs when an untraded activity generates untapped gains from exchange; that is, one agent or group of agents has a willingness to pay to reduce an untraded activity that diverges from the willingness to accept a payment for the reduction of the activity of one agent or group of agents. This outcome establishes the measurability of externalities in monetary terms. Second, an externality occurs when an untraded activity generates untapped welfare gains. This outcome establishes the policy relevance of externalities; if policy is directed toward enhancing welfare, then internalizing an externality is a legitimate policy aim. I therefore argue that an externality, and as posited in mainstream microeconomic models, is most often conceptualized as an untraded activity which generates untapped gains from exchange conjoined with untapped welfare gains.

In chapter three, I show that whether or not the gains view differs from the broad view of externalities depends on whether gains from exchange are invariably associated with welfare gains in the world, as they are in the model. To hold that the two types of gains are invariably linked in this way rests on a commitment to the view that welfare is the satisfaction of actual preferences. But D. Hausman's view that well-informed, unbiased, and self-interested preferences are evidence for welfare is more compelling than the view that preference satisfaction is identical to welfare.

This more compelling account of welfare, however, poses significant challenges for interpreting externalities in the world because it allows gains from exchange to diverge from welfare gains. If untapped gains from exchange are based on misinformed preferences, for example, then they are not informative of welfare gains. I argue, however, that a sufficient explanation of externalities in the world should characterize externalities as generating both untapped welfare gains and untapped gains from exchange.

There are two ways to interpret externalities that satisfy this criterion. First, an externality could be generated only when there are actual untapped gains from exchange that are informative of untapped welfare gains. Second, an externality could be generated when hypothetical untapped gains from exchange can be posited for a given untapped welfare gain. I argue that both interpretations are problematic. The first interpretation characterizes externalities in such a way that they may rarely arise in the world. The second interpretation characterizes externalities in such a way that they are unobservable and ubiquitous. Therefore, it is not clear how to interpret externalities in the world when

individuals hold false beliefs or are biased in the relevant context. Plausibly, this is a common state of affairs, at least in prominent policy problems like climate change. If this is right, then the primary concept that grounds a dominant policy approach to these problems is ambiguous; as it is characterized in microeconomic theory, it cannot provide useful guidance for policy. Therefore, if carbon taxes are justified, it is not on grounds that appeal to the concept of an externality.

In chapter four, I aim to show that the ambiguity inherent in the concept of an externality is not merely a theoretical concern. It has implications for how economists and philosophers assess the coherence of the contingent valuation method, which is a survey-based procedure for measuring externalities. I thus show the importance of being clear on the meaning of the concept of an externality by unpacking the criticisms developed by economists and philosophers of the contingent valuation method and by tracing these criticisms to a mistaken or inconsistent conception of an externality.

Finally, in chapter five I argue that climate change is not merely an externality problem, regardless of how externalities are interpreted. Even if private and social interests regarding climate change were perfectly aligned, meaning that all externalities were eliminated, we could still end up in a regretful state despite acting rationally. This is because climate change involves intertemporal choices that are uncertain and characterized by marginal costs that are uninformative of the total costs of polluting. The aim of maximizing social well-being under these conditions results in a counter-productive incentive to over-pollute.

The seemingly straightforward claim that climate change is an externality problem is therefore far from straightforward. It is unclear how to interpret the concept of an externality, which means that it is also unclear how one can accurately measure an externality. This is problematic for the use of a Pigovian tax to address climate change because this type of tax is determined by the magnitude of an externality. Furthermore, internalizing an externality will not necessarily solve the problem of climate change because climate change is not merely an externality problem. I thus conclude by suggesting that market-based policies are better conceived of as tools to achieve social goals rather than as instruments that solve market failures.

CHAPTER TWO

A Conceptual Clarification of Externalities

2.1 Introduction

The concept of an externality in economics plays a central role in justifying and guiding economic policies. Externalities are broadly understood as untraded spillover effects between agents. Carbon dioxide emissions, for example, are understood as externalities because they are untraded and negatively affect agents who are external to the market decisions that produce the pollution. The mainstream economic response to an externality is to recommend a market-based policy that "internalizes" the externality (Mas-Colell et al. 1995). One way to do this is to set a tax on the untraded activity at the value of the externality in equilibrium. A carbon tax is an example of such a policy. William Nordhaus (2014), for example, estimates the value of the externality generated by carbon dioxide emissions, and therefore the optimal value of the carbon tax, at $18.6 per ton of carbon dioxide emissions.

The concept of an externality is therefore foundational for dominant policies that address critical social and environmental problems like climate change. Yet, externalities

are notoriously difficult to define precisely in economics. The broad understanding of externalities as untraded spillover effects is inadequate because externalities are supposed to be policy-relevant, yet untraded spillover effects are everywhere in the social world. This characterization of externalities is therefore too general for its intended purpose. As I outlined in the previous chapter, there was a debate among economists between the 1960s and 1980s on how to solve this issue, which was largely unresolved. Despite this, there are generally accepted ways of modelling externalities in contemporary economics. My aim in this chapter is to characterize externalities as they are treated in these models.

In microeconomic models, most economists use a two-agent constrained optimization framework to develop key outcomes that are generated by untraded spillover effects. Through a close analysis of this model, I argue that untraded spillover effects generate two key outcomes. First, an untraded spillover effect generates untapped gains from exchange over the untraded activity. That is, one agent or group of agents has a willingness to pay to reduce an untraded activity that diverges from the willingness to accept a payment for the reduction of the activity of one agent or group of agents. This outcome establishes a key feature of externalities: they can be measured in terms of willingness to pay and willingness to accept. Second, an untraded spillover effect generates untapped welfare gains over the untraded activity. This outcome generates the policy-relevance of externalities: if policy is directed toward enhancing welfare, then internalizing an externality is a legitimate policy aim.

I therefore argue that an externality, as posited in mainstream microeconomic models, is most often conceptualized as an untraded activity which generates untapped

gains from exchange conjoined with untapped welfare gains. I show that externalities viewed this way are deductively implied by the broad view of externalities. That is, the assumption that there is an untraded spillover effect combined with the two-agent constrained optimization framework implies that there are untraded gains from exchange over the activity that are assumed to be identical to untraded welfare gains.

Nonetheless, the presence of an untraded spillover effect does not deductively imply the presence of these two types of gains in the world. I review prominent literature on the relationship between economic models and the world to emphasize that it is unwarranted to assume that untraded spillover effects in the world will have the same features that they have in the model. I suggest (but take up fully in chapter three) that my characterization of externalities is superior to the broader view because it makes explicit the key outcomes that must be generated by an untraded activity for it to be considered an externality. That is, if externalities are policy-relevant and measurable in the world, as economists tend to claim, then they must be untraded activities that generate untapped welfare gains and untapped gains from exchange.

2.2 Literature on Externalities

Externalities are broadly understood as untraded spillover effects between agents. However, this view of externalities, which I call *the broad view* of externalities, is problematic. First, if this characterization is correct, then externalities proliferate in the

social world. Many interactions, such as observing a stranger smile from a distance, are untraded spillover effects. But typically, we do not think this type of interaction is relevant to economics or policy. This characterization does not sufficiently target the type of externalities that seem to interest economists. Second, this notion of an externality implies that all interactions should be mediated by markets. Debra Satz (2012) argues that there is nothing internal to the notion of an untraded spillover effect that limits its reach into all aspects of society (32). She argues that this is inappropriate because not all aspects of society should be allocated by markets. For example, gift giving and parental love generate untraded spillover effects but ought not be mediated by markets.

Several prominent economists have taken steps to characterize this concept with greater precision. For example, James Meade (1952) characterized externalities as occurring when factors of production are untraded. Ronald Coase (1960) and Harold Demsetz (1996), respectively, defined externalities as the absence of property rights and the presence of high transactions costs. Kenneth Arrow (1969) argued that externalities are best understood as missing markets. Textbooks tend to draw on these definitions. In four textbooks I examined, the common approach defines externalities as occurring when an agent's utility function depends on the unintended by-products of the activities of other persons (Mas-Colell, Whinston, and Green 1995; Tietenberg and Lewis 2009; Perman et al. 2013).

Maurice Lagueux (2010), and co-authors Nathalie Berta and Elodie Bertrand (2014), argue, respectively, that, an externality is a residual entity or state of affairs. It is

anything which is external to markets. Hence, the way an externality is conceptualized depends on the definition of a market. The reason there are different types of externalities therefore stems from the variety of markets as understood in economic theory. The same problem with the broad view of externalities as untraded spillover effects, however, remains with this characterization. If an externality is merely that which is external to markets, whichever way markets are defined, then there is no way to delineate policy-relevant externalities and untraded activities that seem irrelevant to economic policy, such as the smile of a stranger.

In a more recent paper, Nathalie Berta (2017) restricts her analysis to Arrow's account of externalities. She argues that there are two important ambiguities in Arrow's characterization of externalities as missing markets. First, previous accounts specified that externalities were unintended and untraded spillover effects. On Arrow's characterization of externalities, however, it ambiguous whether the untraded spillover effect for which there is a missing market is unintended or deliberate. Therefore, Arrow seems to add an extra category of untraded spillover effects to the notion of an externality, namely untraded and intended spillover effects.

Second, Berta argues that it is ambiguous whether barter is a case of an externality on Arrow's view. Barter is a form of exchange in which agents exchange goods and services directly for other goods and services, rather than using a medium of exchange such as money to facilitate the exchange. Strictly speaking, goods exchanged through barter are *unpriced* insofar as they do not have a *monetary* price because the exchange takes place outside of a monetary system. Since monetary prices do not mediate barter

exchanges, barter can only be modeled as an interdependence between individuals in Arrow's framework. Therefore, a good exchanged through barter appears to be an externality when examined in Arrow's formal framework. As she shows, because barter is a form of market exchange, it should not count as an externality, per se. These two ambiguities show that, on Arrow's framework, the notion of an externality is reduced to interdependence between individuals. That is, an externality occurs any time there is an interdependence between individuals that is not mediated by a monetary price. The concept of an externality, on this view, is therefore too broad.

In this dissertation, I examine the concept of an externality in contemporary economic theory and argue that there are several problems encountered upon application of the concept to real-world conditions. Simply put, I question whether or not externalities so defined are manifest in the world. Moreover, of the set that might be actualized, it is not clear if they are the ones relevant to policy considerations. The first step in answering these questions is to understand how economists characterize externalities in their models.

2.3 The Treatment of Externalities in Contemporary Economics

2.3.1 Why externalities cause a market failure

There are two prominent models of externalities in contemporary economics. One treats externalities at the level of individual agents and the other at the level of

markets. The second one draws on the analysis secured by the first. As I will demonstrate, both fall short of capturing a concept of an externality that could be readily mapped onto the actual world.

The presence of an externality in a market is a type of market failure. A market is successful insofar as an allocation of resources is *efficient*, which, in welfare economics, is to say that it is *Pareto efficient* or *Pareto optimal*. A market is efficient when no Pareto improvements can be made, which means that no one can be made better off without making someone else worse off (Hausman and McPherson 2006; Tietenberg and Lewis 2009; Mas Colell et al. 1995). Markets are Pareto efficient when marginal costs are equal to marginal benefits. However, when there are costs or benefits that accrue to agents external to the market decisions, then we must distinguish between *private* marginal costs and benefits, which determine market outcomes, and *social* marginal costs and benefits, which are the costs that accrue to those internal and external to the market. This is illustrated in Figure 1 below. The equilibrium price and quantity in the market for cars, for example, occurs where private marginal costs equal private marginal benefits, at q* in the diagram. But the Pareto optimum occurs where social marginal costs equal social marginal benefits, at q^o in the diagram. The externality that is associated with the pollution generated by producing and using cars, for example, constitutes the difference between the two outcomes. So, when an externality is present in some market, the market fails to allocate resources optimally (Hahn and Stavins 1992; Stavins 2011). In this example, too many cars are produced.

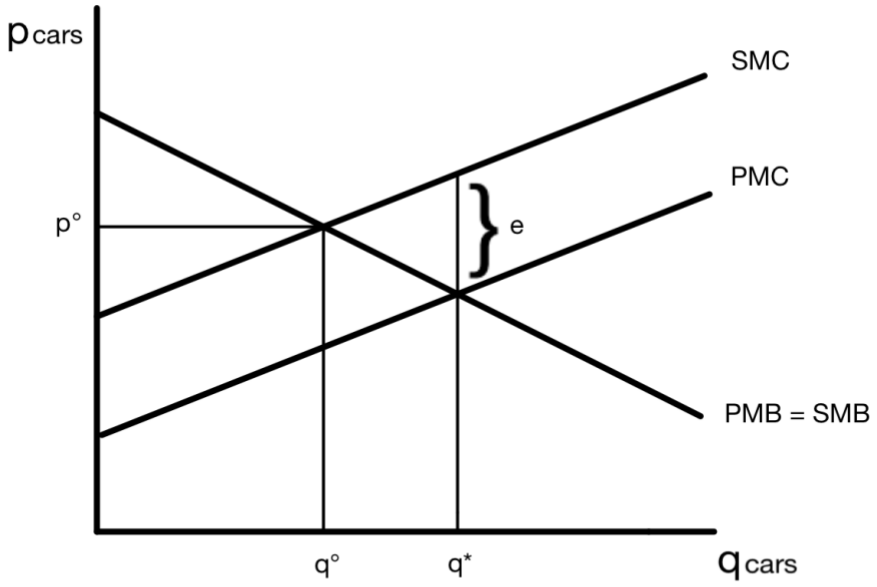

Figure 1: Market for Cars

"q" is the quantity; "p" is price; "q°" is the optimal quantity; "p°" is the optimal price; "q*" is the market price; "PMB" is private marginal benefit; "SMB" is social marginal benefit; "PMC" is private marginal cost; "SMC" is social marginal cost; "e" is the value of the externality.

This model demonstrates the nature of a market failure caused by an externality. However, it does so only by appealing to a vague conception of costs. That is, a negative externality is defined as the difference between social marginal costs and private marginal costs in equilibrium ("e" in Figure 1), but it does not make clear what social marginal costs are. Therefore, externalities are defined in terms of an imprecise conception of social marginal costs. Social marginal costs are generally described as *all costs* of a market action; and externalities are *all costs* that accrue to agents external to the market. But the nature of these costs is unspecified. Therefore, this model is underspecified because it

characterizes externalities in terms of social costs, but the notion of social costs is unspecified.

This model implies, however, that an externality is an untraded spillover effect which is associated with the production or consumption of a market good. For example, the greenhouse gas emissions that are generated through the production and consumption of cars are an externality. This means that the price of cars is lower than it should be, which means that it is lower than the socially optimal price, because the marginal costs and marginal benefits that determine the price of cars do not capture the full cost of the market exchange. This feature that ties externalities to existing markets seems plausible because it is reasonable to suppose that economic analysis is limited to market activities. This is inconsistent, however, with the way externalities are often discussed in economics. A classic example of a positive externality is a neighbour's beautiful garden, and a classic example of a negative externality is a noisy neighbour. These are clearly not market activities: the neighbour happens to enjoy gardening, and in doing so generates a positive externality for everyone who walks by and enjoys the sight of garden. Similarly, the neighbour happens to play the drums, and in doing so generates a negative externality for everyone she disturbs. Therefore, externalities need not be associated with any existing market. Instead, they can be conceived of as the absence of some market, such as a market in which neighbours exchange over noise disturbances (Arrow 1969). This could take the form of the drummer paying the neighbour for the right to make noise, or the neighbour paying the drummer to keep quiet, either by not playing the drums or by installing sound insulation (Coase [1960] 2013).

This model thus treats costs and benefits generally without specifying the nature of these costs and benefits. Instead, the nature of the costs and benefits that constitute the externality and the social marginal costs and benefits is made clear in a two-agent constrained optimization framework, to which I turn in the next section. I take this framework to be foundational to the economic understanding of the nature of externalities in the model.

2.3.2 Externalities in an optimization framework

A prominent graduate level microeconomics textbook (Mas-Colell, Whinston, and Green 1995) explicates externalities in a two-agent constrained optimization framework.[1] The broad characterization of an externality as an untraded spillover effect is assumed in the initial conditions of the optimization problem. That is, it is assumed that an untraded activity undertaken by one agent affects another agent. This is modeled by entering the untraded activity directly into the affected agent's utility function. More precisely, two consumers, consumer one and consumer two, each have preferences over a consumption bundle of n traded goods (x_{1i}, …, x_{ni}) and over some untraded action, h, that is taken by consumer one.[2] Since the action h is untraded, it enters directly into

[1] Versions of this framework for modeling externalities can be seen in Charles Plott (1966), Agnar Sandmo (1980), and Hal Varian (1994), for example.

[2] The subscript "i" denotes the i^{th} consumer. In this model, there are only two consumers, Consumer one and Consumer two. This means that i = 1, 2. Therefore, we can interpret the claim that the utility function of Consumer i is $u_i(x_{1i}, \ldots, x_{ni}, h)$ to mean that the utility function of Consumer 1 is $u_1(x_{11}, \ldots, x_{n1}, h)$ and the utility function of Consumer 2 is $u_2(x_{12}, \ldots, x_{n2}, h)$. This notation means that the maximization problems faced by the two consumers can be treated at once while their problems are the same. A specific value for the variable "i" is inserted to examine the problems faced by each consumer separately.

consumer i's utility function, which has the form $u_i(x_{1i}, \ldots, x_{ni}, h)$. It is also assumed that the partial derivative of consumer two's utility function with respect to h is non-zero, meaning that a change in h changes the utility of this agent. This is an essential feature of externalities: "because consumer [one's] choice of h affects consumer [two's] well-being, it generates an externality" (Mas-Colell et al. 1995, 352). This is to say that the untraded activity must have some effect on consumer two in order for it to generate an externality. If it had no effect, it would merely be some benign action taken by consumer one. The optimization problem therefore assumes that there is an untraded activity taken by one agent that decreases another agent's utility.

The point of the optimization framework therefore is not to describe the mechanism through which the untraded activity affects the agent, nor does it aim to describe why the activity remains untraded. Instead, it establishes two key outcomes that are generated by an untraded spillover effect. First, it shows the market outcome in the presence of an untraded spillover effect. At the competitive equilibrium, each agent maximizes her utility which is constrained only by her budget. The utility maximizing problem for consumer i is:

Max $u_i(x_i, h)$ subject to the budget constraint $p \bullet x_i \leq w_i$

The consumer's problem is to choose the bundle of goods (x_i) that maximizes her utility given the prices for the goods (p) and given that she has limited wealth (w_i). Importantly, consumer one chooses how much of the activity h both consumers will incur. That is, consumer two must "consume," to the detriment of her well-being, whatever level of h consumer one chooses to produce.

Consumer one chooses the level of h that maximizes her derived utility $ø_1(h)$.[3] That is, the first-order condition for an interior solution to the competitive equilibrium level of h, h*, is:

$$ø'_1(h^*) = 0$$

The function $ø'_1(h)$ is the partial derivative of derived utility with respect to h, which is the marginal benefit of h to consumer one. The first order condition tells us that derived utility reaches a maximum when the marginal benefit of consuming h is zero. More generally, $ø'_1(h^*) \leq 0$, with equality if $h^* > 0$. To see why this is the case, suppose consumer one produces h at a level less than h*. When h is less that h*, the change in utility given a change in h is positive. Hence there are welfare gains from producing more h. When h is produced at a level greater than h*, on the other hand, the marginal benefits from consuming h are negative. This means that consuming more of h reduces welfare. Therefore, utility is maximized after all benefits from consuming h are exhausted. Hence this framework shows the market outcome in the presence of an untraded spillover effect. Consumer one produces h up to the point where benefits from producing h are completely exhausted, which occurs when marginal benefits are equal to zero.

Secondly, the optimization framework shows that the competitive equilibrium in the presence of an untraded spillover effect is suboptimal or inefficient. Therefore, it demonstrates why externalities are undesirable. The Pareto optimal level of h, h^o, in

[3] Derived utility is $v_i(p, w_i, h)=\max u_i(x_i, h)$ subject to $p \bullet x_i \leq w_i \equiv ø_i(p, h) + w_i$. Since prices are unaffected by the choice of h, the price vector is suppressed. Therefore, derived utility is expressed as $ø_i(h)$. See Mas Colell et. al (1995, 352-354) for more detail.

contrast to the competitive equilibrium, maximizes the joint surplus of the two consumers.[4] That is, h^o is determined by solving:

Max $ø_1(h) + ø_2(h)$

The first-order condition for an interior solution to h^o is:

$ø'_1(h^o) = – ø'_2(h^o)$

This says that the Pareto optimal outcome (h^o) occurs where the marginal benefit of h to consumer one ($ø'_1(h)$) is equal to the marginal cost of h to consumer two ($– ø'_2(h)$), as illustrated in Figure 2 (Mas-Colell et al. 1995, 354). To see why this is the case, suppose consumer one produces h at h*. At this level of h, consumer one's marginal benefit of h is zero ($p_1^* = ø'_1(h^*) = 0$), whereas consumer two's marginal cost of h is greater than zero ($p_2^* = – ø'_2(h^*) > 0$). Consumer one is willing to accept any payment that is equal to or greater than p_1^* to reduce h by one unit at h*, and consumer two is willingness to pay consumer one any price that is equal to or less than p_2^* to reduce h by one unit at h*. Hence there are a range of prices, between p_1^* and p_2^*, for which exchange could occur over the quantity of h. This is to say that there are unexhausted gains from exchange at h*. The total value of these unexhausted gains from exchange is the shaded region in Figure 2. Mutually beneficial trades are possible any time the willingness to pay for a reduction in h is greater than the willingness to accept a payment to reduce h. The socially optimal level of h (h^o) occurs where willingness to pay is equal to willingness to accept, and thus all gains from exchange are exhausted.

[4] An outcome is Pareto optimal if there is no possible reallocation of goods that will make someone better off without making someone else worse off.

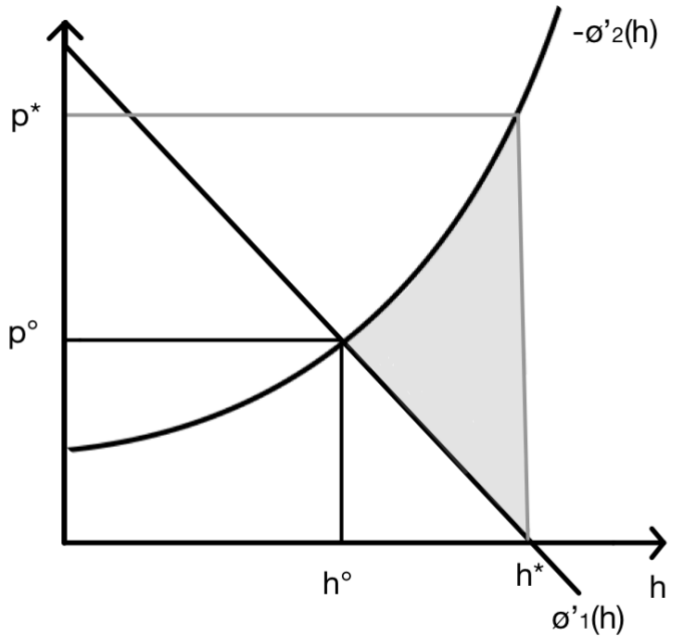

Figure 2: Externality

As long as a negative externality is present, which requires that $ø_2(h)$ is decreasing in h for at least some levels of h, the competitive equilibrium level of h is greater than the Pareto optimal level of h. That is, h is over-produced and underpriced. The policy solution to an externality is thus to create incentives for the agent to produce h at the Pareto optimal level, h^o. One way to do this is to internalize the externality by setting a marginal tax on h at p^o, which is the price of h at the social optimum. When the marginal tax rate t is equal to $–ø'_2(h^o)$, shown in Figure 3, consumer one maximizes utility not by setting $ø'_1(h) = 0$ but instead by setting $ø'_1(h) = – ø'_2(h^o)$. This means that consumer one produces h just up to the point where her marginal benefit from h is equal to her

marginal cost of h, which in this case is the marginal tax rate. Therefore, the competitive outcome when the marginal tax is set at the marginal value of the externality, called a Pigovian tax, is the Pareto optimal outcome (h^o).

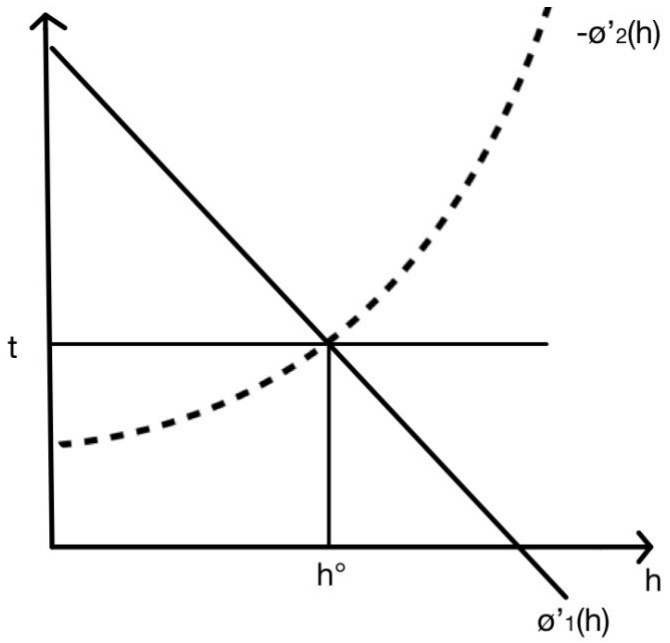

Figure 3: The Optimal Tax Rate

A carbon tax, in theory, is an example of a Pigovian tax. The optimal carbon tax rate is thus set at the price that equalizes the marginal costs and marginal benefits of carbon dioxide emissions (Weitzman 1974; Pearce 2002). It should be noted, however, that it is not clear that the same conception of an externality is being employed by economists who estimate the social cost of carbon in integrated assessment models. Economists such as William Nordhaus (2014) and David Anthoff and Richard Tol (2010) use macroeconomic models to estimate the externality generated by carbon

dioxide emissions. The model I examine is supposed to be the microeconomic foundation for such macroeconomic models. That is, the model I examine offers a precise account of how a tax can internalize an externality and is supposed to explain, at the individual level, the same phenomenon that macroeconomic models explain at the level of the entire economy. However, it is not clear that the macroeconomic concept of an externality captures the aggregation of the phenomena explained by the microeconomic concept.[5]

2.4 Characterizing Externalities in the Model

In this section, I develop a characterization of externalities as they are treated in a constrained optimization framework, called the *gains view* of externalities. This framework reveals assumptions of the modeler and establishes two key outcomes caused by untraded spillover effects, each with an associated feature of externalities. The first outcome is untapped gains from exchange, which gives externalities the feature of measurability. The second outcome is untapped welfare gains, which gives externalities the feature of policy relevance. I will argue that although untraded spillover effects invariably generate these outcomes and features in the model because of the deductive relationship between willingness to pay, preference satisfaction, and welfare, it is unwarranted to assume that this is the case in the world. I will argue that the gains view

[5] As I will outline in chapter six, this is a topic of future research.

of externalities is superior to the broad view of externalities as unpriced spillover effects because it highlights the criteria that an untraded spillover effect must meet to qualify as an externality in the world.

2.4.1 Assumptions of the modeler

The choices made in setting the initial conditions of the optimization framework are revealing of how externalities are conceptualized in economics. First, it is assumed that there is an untraded activity that is produced by one agent because it increases their utility. Utility is understood as ordered preferences that are frustrated or satisfied. It does not explain why this activity is untraded or why it remains untraded. Coase ([1960] 2013) instead examines why an activity might remain untraded and therefore unpriced, while the model examined here establishes the outcome that follows from the presence of an untraded activity. Coase argues that an activity will remain untraded when property rights are not properly defined or when the transaction costs are sufficiently high to prevent exchange over the activity. That is, Coase argues that no externality would persist if property rights were well-defined and transaction costs were zero. If an agent had a willingness to pay greater than another agent's willingness to accept for an activity, then these agents would engage in trade over this activity and thus the externality would be eliminated.

Of course, and as Coase acknowledged, transaction costs are rarely zero and sometimes a certain type of activity cannot be effectively circumscribed by property rights. Pollution faces both of these barriers to exchange. Transaction costs are high

because, for example, the relevant agents are sometimes on opposite sides of the globe. Furthermore, property rights cannot be clearly defined over pollution because it is an indivisible good. As I explained in chapter one, cap and trade systems are an attempt to create the conditions necessary for trade to occur over carbon dioxide emissions; however, the choice of the quantity of pollution that constitutes one unit of pollution that can be owned is part of the policy problem. Defining the optimal quantity is the flip side of the problem of establishing the optimal price for a carbon tax. Both require estimating the value of the externality. The lack of property rights or high transaction costs thus explains why something might remain untraded, but it does not explain why this matters. The two-agent constrained optimization framework, on the other hand, takes as given that the activity is untraded, for whatever reason. Instead, it shows why this is a problem and how it can be solved. As I will argue, it matters because there are untapped gains from exchange that are associated with untapped welfare gains.

This model also assumes that the untraded activity generated by one agent reduces another agent's utility. This agent cannot alter the quantity of the activity that is produced because the activity is assumed to be untraded. Note that characterizing this activity as *untraded* is more precise than *unpriced* because, as the model shows, the activity has a price even though this price is not realized through exchange. Indeed, this activity has several prices that are faced by different individuals. The price faced by the producer is zero, the price faced by the receiver of the activity is positive, and the Pareto optimal price is the price at which the willingness to pay equals the willingness to accept for the activity. What is assumed, therefore, is that there is an untraded activity that has an effect

on an agent external to the production decision. Recall that this assumption constitutes the broad characterization of an externality as an untraded spillover effect.

2.4.2 Outcomes that establish features of externalities

2.4.2.1 Untapped gains from exchange

The optimization framework therefore contributes to the characterization of externalities insofar as it establishes the nature of the outcome that is caused by an untraded spillover effect. There are two key outcomes that establish necessary features of an externality. First, an untraded activity generates an outcome in which there are untapped gains from exchange. There is a potential for exchange when willingness to pay for some good or activity is not equal to willingness to accept at the competitive equilibrium for the untraded activity. There are gains to be made from exchange because rational and perfectly informed agents in the model are only willing to pay for a good up to the point where they are at least indifferent between purchasing the good and not purchasing the good. That is, they are willing to pay for a good only insofar as the exchange does not reduce their utility level. This means that the *gains* are gains to preference satisfaction. There are untapped gains from exchange at the competitive equilibrium h* in Figure 2 because the willingness to pay to reduce h by one unit is p*, which is greater than the willingness to accept a payment to reduce h by one unit, which is any positive price. Hence agents would exchange if they could, and the exchange would better satisfy their preferences. This outcome of an untraded spillover effect establishes an important feature of externalities: they have a magnitude. The optimal

price of the untraded activity is the price at which gains from exchange are exhausted. That is, the optimal price is the price at which willingness to pay is equal to willingness to accept. Hence the magnitude of an externality, which determines the value of the marginal tax that would internalize the externality and restore efficiency, is derived from the measures of willingness to pay and willingness to accept, for a given activity.

2.4.2.2 Untapped welfare gains

Second, an untraded activity generates untapped welfare gains. That is, there are potential Pareto improvements at the competitive equilibrium in the presence of an externality. In economic theory, preference satisfaction, given certain assumptions regarding initial endowments, is assumed to constitute welfare. The gains to be made through exchange are gains to preference satisfaction, and preference satisfaction is assumed to constitute welfare. Therefore, in economic theory, gains from exchange are essentially tied to welfare gains. Willingness to pay is the partial derivative with respect to h of the agent's utility subject to the budget constraint, and the satisfaction of preferences (utility) is assumed to be what generates welfare. Put another way, welfare is utility (preference satisfaction), and the change in utility given a change the untraded activity subject to the budget constraint is the willingness to pay for the untraded activity. Willingness to pay is therefore a measure of welfare gains. Gains from exchange are therefore invariably associated with welfare gains in the model. Since the untraded spillover effect is modeled as directly entering the utility function, there must be an

associated willingness to pay for the activity, and there must be potential welfare gains associated with changing the level of the activity.

Therefore, the optimization framework establishes that externalities are undesirable because untapped welfare gains are undesirable, and the externality is eliminated when the correct price is set on the activity. This model establishes the features of these outcomes that are crucial to the policy implications of an externality; they have a magnitude and they are policy relevant. It therefore shows how to internalize an externality by setting a tax on the untraded activity at the value of the externality in equilibrium, which is the price at which willingness to pay equals willingness to accept. Thus, the tax exhausts potential Pareto improvements.

Note that the potential exchanges in the case of externalities are often qualitatively different to potential exchanges in standard markets where buyers and sellers interact. This is because it is assumed at the outset that direct exchange over the untraded activity is inhibited for some unspecified reasons (although Coase specifies these reasons to be high transaction costs and poorly-defined property rights). This means that the untapped gains from exchange over an untraded activity cannot be truly *tapped* because no exchange can take place. Furthermore, the economic solutions to an externality do not mimic exchanges in a traditional market where parties who incur costs from bringing a good to market receive payment from those who benefit from the good. That is, polluting firms, who benefit from pollution because it lowers their costs of production, do not pay harmed individuals for the pollution. Nor do individuals who would benefit from reduced pollution pay firms to pollute less.

Instead, a policy is used to create the outcome that would have occurred if exchange over the untraded activity were possible. That is, the untapped gains from exchange indicate the optimal quantity of pollution. A policy tool is then used to induce polluters to emit pollution at this optimal level. A marginal tax on pollution emissions can achieve this by setting the tax rate at the optimal price of pollution; this type of tax is called a Pigovian tax. A cap and trade scheme sets the total allowable pollution at the optimal quantity of pollution, allocates permits to polluters that gives them the right to emit a portion of the total allowable pollution, and allows permit-holders to trade these permits. Both policies involve no exchange between polluters and those harmed by pollution, but instead use incentive mechanisms to induce the production of pollution at the optimal level, and the optimal level is defined by the untapped gains from exchange.

Therefore, the policy is designed to eliminate untapped gains from exchange. It does not exhaust untapped gains from exchange because no exchange occurs. The policy does not, however, eliminate or exhaust potential welfare gains. Arguably, a redistribution of income could maximize welfare, thus exhausting potential welfare gains. But this is not a Pareto improvement because a redistribution of income would make some agents better off while making other agents worse off. Pigovian taxes, therefore, are not concerned with maximizing welfare; instead, they are designed to eliminate potential Pareto improvements. They do this by creating incentives for producers of the untraded activity to produce the activity at the level that would be achieved if there were a market for that activity. If there were a market for that activity, the potential Pareto improvements would be exhausted at the market equilibrium, which is the level at which

there would be no additional mutually beneficial exchanges. If there were a market for pollution, for example, the payment received by polluters to reduce their emissions would compensate them for their abatement costs.

No exchange occurs when a Pigovian tax is implemented, however. Because of this, Pigovian taxes put producers in a state that they disprefer; the implementation of the policy makes producers worse off because they must pay to pollute or install abatement technology without compensation that would have been received through exchange if there were a market for pollution. In response to this issue, economists typically accompany the recommendation for a Pigovian tax with the recommendation that those who pay the cost of internalizing the externality be compensated by transfer payments from others. This might take the form of using the government revenue from the Pigovian tax to compensate producers in the form of lump sum payments. This does not alter the incentives created by a Pigovian tax because producers maximize profits by reducing pollution to the socially optimal level and receiving the lump sum payment to off-set their costs; keeping the level of pollution unchanged and receiving lump sum payments is more expensive to the producer because the tax costs are plausibly significantly greater than the cost of installing abatement technology in the long run. Therefore, in principle, Pigovian taxes require no one to make sacrifices in terms of their welfare (Broome 2018).

In sum, I have argued that the two-agent constrained optimization framework shows that an externality occurs when an untraded activity generates both untapped gains from exchange and untapped welfare gains. This is the gains view of externality.

2.5 Implications of the Gains View of Externalities

In a utility maximizing framework, untraded spillover effects invariably generate untapped gains from exchange and welfare gains. The implication of this, in theory, is that all untraded activities that have an effect on others ought to be mediated by markets or by market-based policy. That is, the broad view implies the gains view of externalities. This result follows from the nature of economic theory, where agents can only be affected by something via an influence on their utility function. Therefore, an effect will always have an associated potential gain from exchange and welfare gain. An untapped gain from exchange is a divergence in a willingness to pay for an activity and a willingness to accept a payment for the reduction in an activity. Willingness to pay for an activity is the marginal cost of the activity to the agent and willingness to accept is the marginal benefit of the activity. Marginal cost or benefit is the change in constrained utility given a change in the untraded activity. Hence gains from exchange are utility gains (which are understood as welfare gains) because willingness to pay is a partial derivative of utility. In this framework, then, *all* untraded spillover effects between agents generate both untapped gains from exchange and untapped welfare gains.

This is the feature of externalities that is criticized by philosophers such as Amartya Sen (1977) and Debra Satz (2012). Satz (2012, 31-32) discusses externalities and market failure to motivate her argument that there are moral limits to markets that cannot be explained with the theory of market failure. Moreover, she claims that untraded spillover effects are everywhere in the social world and thus economists are

unwarranted when they assert that every interaction ought to be mediated by markets. She suggests that the concept of an externality is used selectively by economists. For example, pollution is treated as an externality while intolerance of religious diversity is not, but this distinction is an implicit normative choice on the part of the theorist rather than a distinction internal to the notion of an externality (Satz 2012, 32). Sen (1977) offers a critique of the notion of an externality in his analysis of the problems with the economic conception of rationality. He argues that sympathy, a normal human emotion, generates an externality and therefore an inefficiency, according to economic theory, because sympathy is an unpriced spillover effect. That is, another person's well-being has an unpriced effect on my well-being when I sympathize with them.

It seems plausible to assume that untraded spillover effects proliferate in the social world. Activities as common as social interactions generate untraded spillover effects. But economic theory does not necessarily imply that all untraded spillover effects in the world are policy-relevant externalities. The nature of the policy-relevance of untraded spillover effects is generated in the model; that is, untraded spillover effects are policy-relevant insofar as they generate untapped welfare gains that are measured by untapped gains from exchange. It is not obvious that untraded activities behave in the same way in the world and the model. Therefore, the inference from the model to the world must be examined in order to establish the conditions under which untraded spillover effects in the world have the same features of untraded spillover effects in the model, which generate their specific policy implication. Therefore, Satz's claim that economic theory implies that all untraded spillover effects ought to be mediated by

markets is plausible insofar as it is a claim about untraded activities only in theory. It is not plausible, however, that this claim is true of the world; economic theory does not imply that all untraded spillover effects in the world should be mediated by markets. This is because untraded activities in the world do not necessarily generate the same outcomes that are specified in the model and which generate the specific policy implications that concern Satz.

In the next chapter, I will argue that not all unpriced spillover effects in the world are policy-relevant externalities, as I have characterized them. That is, the broad view is not identical to the gains view of externalities in the world. This is a case where the model does not map directly onto the world. Prior to this, however, I will motivate the importance of distinguishing claims about untraded spillover effects in the model and the world by analyzing prominent perspectives on the nature of modelling in economics. If the epistemic reach of models were properly understood, this might produce a more plausible characterization of externalities. Insofar as this helps to establish which untraded activities are policy-relevant, policy makers would be better served.

2.6 The Relationship between Models and the World

In this account of externalities in the model and the world, I do not intend to make novel claims about the nature of models in economics. However, I make the minimal claim that if the policy implications of externalities that are established in the

model are used to determine and justify policies in the world, then the features of an externality that generate these policy implications in the model should also be attributes of externalities in the world. The two-agent constrained optimization problem establishes that an untraded spillover effect generates untapped welfare gains that can be measured by untapped gains from exchange. On this account, the untraded activity is overproduced and underpriced. A tax set at the value of the activity where willingness to pay is equal to willingness to accept eliminates the potential gains from exchange and potential welfare gains and thus establishes an efficient output of the activity. The same reasoning is used in the world. Economists use various techniques, such as surveys, to determine the willingness of human subjects to pay for unpriced activities which serve, in turn, as the basis for constructing a demand schedule (Mitchell and Carson 1989). This demand schedule can be used to estimate the optimal policy response, such as a tax set at the efficient price (Stavins 2011). Economists thus treat externalities in the world in a way that conforms to the findings of the model. Attention is needed, however, to the relationship between the model and the world to justify this symmetrical treatment of untraded spillover effects as represented both in the model and the world.

There are several insightful philosophical analyses on the use and construction of models in economics. I will single out three, by Nancy Cartwright (2007), Mary Morgan (2002), and Robert Sugden (2002). It is important to note at the outset that each of these authors is well-versed in econometrics and have written a substantial body of scholarship that critiques the methodological efforts of economists to measure and assign correlations, if not causal relations, to various parameters. Nancy Cartwright is also an

acclaimed philosopher of physics and has made seminal contributions to the literature on the problems of causal ascriptions in quantum mechanics. She holds that scientists assign fundamental causal relationships, what she calls *tendencies* or *capacities*, that are obscured if one acknowledges the full complexity of the world (Cartwright 1983). The aim of scientists is to discover these tendencies, and in the case of modern science, they often construct models. For Cartwright, unlike other philosophers such as Larry Sklar (2000) and Robert Batterman (2002), a model is not justified by its empirical success. In fact, like Milton Friedman (1953), she argues that the truth per se does not explain well. Models aim to capture relations that hold universally, and are therefore not easily verified.

Cartwright argues that there is reason to suppose that a causal tendency derived in a model is unobservable in the world because it is likely to be obscured by confounding causes (Cartwright 2007, 222). Therefore, there is no reason to suppose the model should be predictive or descriptive of our observations. Instead, a good model is one that successfully isolates a tendency. The models that can do this are analogous to what she calls *Galilean experiments*, which are lab experiments that eliminate all possible confounding causes in order to establish the effect of the causal mechanism in question (Cartwright 2007, 223). *Galilean idealizations* are mathematical models that successfully eliminate all confounding causes in order to determine the effect of the cause in question. The model thus deduces an outcome that could also be established by conducting a Galilean experiment.

Galilean idealizations are built from general principles that hold in the target situation. These idealizations construct a system from which a result can be derived that is isolated from confounding causes; the use of general principles in a sense constructs the controlled environment we would construct in a laboratory. The derived result is a true causal relationship, or a tendency, because it is derived in isolation from all confounding causes. As such, the result can be carried outside the experiment because it is a true relationship in the target situation.

Cartwright doubts that models that take the form of Galilean idealizations have been formulated in mainstream economics. As she observes, economists aim for deductive rigour. As such, they construct models similar to models in physics in which the results deductively follow from general principles. Unlike physics, however, economics lacks general and uncontroversial principles. Whereas the structure of the mathematical system in physics models is constructed with general principles, the structure of the mathematical system in economic models must instead be constructed with assumptions. Some structure is required to construct a mathematical system from which results can be derived, and economic theorists thus create structure in this structure with assumptions rather than general principles. Cartwright (2007, 227) tells us that it is not uncommon for economic models to include a list of over a dozen assumptions, such as the assumption that workers are wage maximizers, or that consumers and firms have perfect information. The two-agent constrained optimization model of externalities, for example, assumes that agents are utility-maximizers that are rational and have perfect information.

These assumptions are not true of the world but instead help construct a simple model economy that is tractable. These model-specific assumptions allow for deductively validated results where universal principles are scarce. However, the cost of deductive rigour in economics is that the results that are deduced in models are tied to special circumstances (Cartwright 2007, 229). For both models in physics and in economics, results that are deduced within models depend on the structure of the model. This is unproblematic if the structure is true of the world, which Cartwright argues is the case in physics. This dependence, however, is problematic in economics because the structure of the model is not true of the target scenario in the world. This is because the model is given structure through strictly false assumptions rather than general principles. Individuals never possess perfect information, and yet this assumption is often included in mainstream economic models. Therefore, a result deduced in an economic model is not a tendency, but rather a tendency plus the confounding effect of the false assumptions that are built into the model.

According to Cartwright, then, models can be used to find out about tendencies when they are Galilean idealizations; they isolate a single cause and derive the true effect of this cause alone. Economics lacks principles that are true of the world, and therefore must use assumptions to deduce a causal result. But then economic models cannot be used to find out about tendencies because any causal relationship they derive depend on the assumptions of the models. Therefore, economic models only imprecisely capture universal behaviours we observe in the world, such as an increase in demand leading to an increase in price. Economic models do not unearth laws that are universally true of

economic systems; instead, the behaviours derived in economic models are only true of the specific, false assumptions of the models.

Cartwright's view implies that current economic models are largely uninformative of the world. But in practice, models are important to how economists understand the world; as we see in the case of an externality, a concept developed in a model is used to establish economic policy recommendations that are influential to the design of actual policies in the world, such as the carbon tax. This indicates that the models that economists construct are different from the models that physicists construct. Physics has well-confirmed background theory to build and de-idealize models; furthermore, it has well-behaved systems that are conducive to the discovery of true principles. Economists attempt to model human behaviour and human-constructed systems. Like ecology and meteorology, economics has comprehensive foundations but lacks well-confirmed background theory and well-behaved systems (Morgan 2012, 14). As such, it is plausible that the understanding of models as a way to perform Galilean experiments is appropriate to physics but not to economics.

Mary Morgan (2002) holds a constructivist view of economic models. Economic models are constructed to represent hypothetical economies that incorporate our general theoretical claims or hypotheses about the economic world (Morgan 2002, 193). They do not isolate a causal relationship that exists in the world. Instead, economic models draw on extant economic theory and construct their own integrated world. This "world in the model" serves an important pedagogical function. There is much to learn from a model-based form of inquiry; by examining and perturbing the model world, we can learn about

relationships that were not transparent at the time the model was formed. This form of inquiry, however, is distinct from an inquiry that uses the model to understand the world. The view of economic models as constructed worlds is therefore faced with the problem of explaining how the insights gained from a model-based inquiry relate to the world representationally. Morgan responds to this problem with an account of narratives that connect the model world to the real world, which I will explain below.

Morgan argues that we learn about economic theory by using and manipulating economic models. Economists must pose questions in order to understand and make use of the internal dynamics that the mathematics of the model provides (Morgan 2002, 183). Consider a supply and demand diagram that represents the market for coffee. The horizontal axis represents the quantity available in the market, measured in units of coffee, and the vertical axis represents price per unit. The demand curve for coffee is normally downward sloping, justified by the idea that more coffee will be demanded as the price decreases. The supply curve is upward sloping, justified by the idea that more suppliers will be willing to produce more coffee as the price increases. The intersection of the two curves is the equilibrium point where supply equals demand. This is the equilibrium because the price will settle at the level where the quantity supplied just equals the quantity demanded; any price above or below this level will gravitate toward the equilibrium because, for example, excess coffee supplied creates downward pressure on the price. The analysis of this model is therefore intended to illustrate the mechanisms that obtain in an actual market even though they reside, strictly speaking, in the world of the model.

To understand the dynamic nature of the model, however, Morgan argues that we must pose questions and subsequently offer an interpretation of what the model predicts. For example, Morgan asks, "what would happen if the income of consumers increased?" Income level was originally contained in the *ceteris paribus* clause; now, we are perturbing the system to determine a different effect. An increase in income will shift the demand curve to the right because at every price level consumers now demand more of the good. We see that at the previous price, demand exceeds supply which leads to upward pressure on the price. Price will increase just until the new equilibrium point is reached at a higher price and higher quantity. In using and manipulating the model, Morgan argues, we come to understand the dynamics of the model and learn something about our hypothetical economy.

Indeed, an examination of the construction of a demand curve helps to illustrate the extent to which simplifying assumptions are employed in supply and demand models to construct a hypothetical economy. The demand curve for the market for coffee is constructed by aggregating individual demand curves, which in turn are derived from the utility functions of individuals. An individual demand curve for coffee is determined by the marginal utility an individual receives from consuming incremental units of coffee, subject to their budget constraint. There are no utility functions in the world, however; utility functions are an index of preferences in which a number is assigned to represent an individual's ordered preferences (D. Hausman 1992b).

Utility functions can be used to represent preferences only when preferences are complete, transitive, and continuous (Debreu 1959; D. Hausman 1992b). Transitivity

requires that an individual does not have circular preferences; if preferences are transitive, then when X is preferred to Y and Y is preferred to Z, X must be preferred to Z. Completeness requires that an individual can rank all options. Clearly it is implausible that most individuals have a settled ranking of all options. Continuity requires that, "for every option x both the superior and inferior sets are closed" (D. Hausman 1992b, 17). This requires that, for example, an individual ranks the same good according to a single criterion. Continuity is not a requirement of rationality, unlike the conditions of transitivity and completeness; arguably, it is instead required for "the mathematical idealization involved in the use of real numbers" to represent preferences (D. Hausman 1992b, 17-18). Agents are also assumed to have perfect information and foresight. Goods are assumed to be infinitely divisible, even though no actual good has this property. Furthermore, economists do not know the particular shape of the utility functions of individuals (D. Hausman 1992b, 39). This means that additional assumptions about the shape of an individual's utility function are required for economists to explain or predict how demand responds to price or income shocks. The extensive use of idealizing and simplifying assumptions in the construction of a demand function thus helps to illustrate that the mathematical representation of demand strain credulity.

It is important, then, to understand how the insights gained from a model-based inquiry relate to the world representationally. Morgan argues that a hypothetical economy constructed in a model is related to the real world through a narrative explanation. The interpretation of the market for coffee above, and the typical type of explanation

economists give of their models, takes the form of a narrative explanation because, like stories, they feature events that occur consecutively with ambiguous causal claims (Morgan 2002, 186). For example, the sequence of events is well defined (there is an increase in income, which is followed by an increase in demand) and it is unclear whether an increase in income causes a price increase or whether it is correlated with a price increase. Thus, Morgan argues that it is through narrative explanations of economic models that economists relate their models to the world. They try to re-introduce the complexity of the world to the results of the model through questioning the model. We do not expect our explanations to be exactly true, but instead to be correlated with economic systems in the world (Morgan 2002, 197). That is, our interpretations, rather than background theory or the soundness of the results, relates economic models to the world.

Insofar as economists use narrative explanations to relate the model world to the real world, they explain and enlighten. It is not satisfactory, however, that the narrative structure alone explains how the two worlds are related. The content of the narrative structure is plausibly a determinant of a successful relation of the models to the world. Robert Sugden (2002) argues that the explanation involves inferences from the model world to the real world. Like Morgan, Sugden holds that economic models are constructions because *ceteris paribus* clauses do not exhaust the extent of the idealization. Furthermore, he also draws an analogy between models and fiction: like fiction, models present to us an alternate yet similar world to the actual world. He argues, however, that inductive inferences rather than narrative explanations connect the model and the world.

That is, through modeling, economists construct credible worlds that are connected to the real world by way of inductive inference.

Sugden takes George Akerlof's (1970) model of "the market for lemons" as an exemplar of economic models. Sugden analyzes this model in order to determine how economic models explain. This model describes a simple market for used cars. In constructing this market, "Akerlof sets up an imaginary world which makes no pretense to be realistic" (Sugden 2002, 109). In this world, there are two types of traders, type one and type two. There are *n* cars that differ only in quality, which is measured in monetary units. Each group of traders aims to maximize an aggregate utility function, which specifies both that all traders are risk-neutral and that the value a buyer ascribes to a car is greater than the value a seller ascribes to that car. He assumes that all cars are initially owned by type one sellers, and the quality of the car is known only by the seller. All traders know the average quality of all traded cars, however. These simplifying assumptions cannot be understood as general principles or bracketed with a *ceteris paribus* clause. Therefore, Akerlof constructs a model that relies on the particular structure that he writes into it. He is not isolating a system by stripping away irrelevant causal factors; instead, he is using unrealistic assumptions to construct a model world that "allows him to focus on those features of the real market that he wishes to analyse" (Sudgen 2002, 110).

Akerlof then investigates the level of trade that occurs in this model world. He argues that if any trade takes place, then there must be a single price for cars. This is because buyers are completely unaware of the quality of any given car, and therefore are

not willing to pay different prices for different cars. But sellers are aware of the quality of their car; sellers of good quality cars require a higher price for their cars than sellers of low quality cars. Therefore, the market price is determined by the price of the high-quality cars. This means, then, that the value of the average quality of cars is always lower than the price for cars. Since buyers are not be willing to pay a price that is higher than the average quality of cars, no trade will take place at any single market price.

If information is symmetric, instead, which means that seller and buyers have the same information about the distribution of the quality of cars in a market but no information about the quality of any given car, then trade takes place as it would in any standard market. The result of the model is therefore that the higher the degree of asymmetry of information in the market, whereby the seller has more information about the quality of the car than the buyer, the lower is the volume of trade that takes place. When the seller has much more information than the buyer, the buyer does not purchase the car even when it is a good quality car and it is priced at the value of a good quality car.

As Sugden explains, Akerlof's result is too strong to explain the price difference between new and used cars in real markets. He found that no trade occurs when there is asymmetric information, which is certainly not the case in real markets. "Presumably, then, Akerlof sees his model as describing in extreme form the workings of some *tendency* which exists in the real used-car market, by virtue of the asymmetry of information which (he claims) is a property of that model" (Sugden 2002, 111). This tendency is that low quality cars drive good quality cars out of the used-car market; in the real world, this

tendency has the effect of reducing the average quality of a car in this market, and this low quality explains the low price of used cars compared to new cars.

Akerlof thus explicitly constructs his model in such a way that what is deduced from the model depends on the assumptions of the model. Sugden interprets Akerlof to be using the model world to make a claim that a regularity is the effect of a set of causal factors. In this model, Akerlof claims that a regularity occurs (here, that good products are driven out by bad products) and that this regularity can be explained by some set of causal factors (here, that sellers are better informed than buyers). That is, Akerlof constructs an imagined model world and shows that, under specific conditions and holding all else constant, an increase in the degree of asymmetry of information (the causal factor) decreases the volume of trade (the regularity). Given that this causal relationship is derived from a model world, this finding tells us nothing about the relevant variables in the real world. Despite this, Akerlof uses the findings of this model to draw more general conclusions that are at least credible, if not true, of all markets (*ceteris paribus*). He does this without explaining why he thinks he is justified in using the results of the model world to explain a regularity in the real world.

Sugden argues that Akerlof uses inductive reasoning to fill the gap between the specific result of the model and the general claim. Let R stand for the regularity, which is that bad products drive out good products. Let F stand for the set of causal factors, which is that sellers are better informed than buyers. Sugden argues that we can interpret Akerlof to be reasoning in any of the following three ways:

1. Explanation: In the model world, R is caused by F; F operates in the real world; R occurs in the real world; Therefore, there is reason to believe that in the real world, R is caused by F.
2. Prediction: In the model, R is caused by F; F operates in the real world; Therefore, there is reason to believe that R occurs in the real world.
3. Abduction: In the model world, R is caused by F; R occurs in the real world; Therefore, there is reason to believe that F operates in the real world.

In any of these forms of reasoning, there is an inductive leap from premises to conclusion. "It seems, then, that Akerlof's … method is not purely deductive: it depends on induction as well as on deduction" (Sugden 2002, 126). The question of how a model is used to explain or predict phenomena in the marketplace, then, rests on the justification of this inference, a matter that appears to have no solution.

Given that Akerlof's model is paradigmatic of current practice in economics, Sugden argues that this shows that economists are in a vulnerable place regarding their use of results derived from models. They do not, strictly speaking, deduce their core propositions from the assumptions. If they did, then Akerlof would not have to qualify his hypothesis as corresponding to the specific assumptions in the model. Sugden is therefore in implicit agreement with Cartwright that the results derived in an economic model are dependent on their initial assumptions and therefore the direct application of these results to our world is unjustified. However, Sugden does not conclude from this that economic models are unsuccessful. Instead, he argues that "to understand what Akerlof and Schelling are doing, we have to realize that the results that they derive

deductively within their models are not the same as the hypotheses they want us to entertain (Sugden 2002, 123). To assess whether Akerlof is justified in hypothesizing that the model results hold in the world, we must assess the strength of the induction from the model to the world.

Sugden rejects the claim that robustness checks justify the inference. A robustness check is a system of reasoning whereby theorists alter the model to determine how much the result relies on the various assumptions built into the model. If a model is robust, then the result does not depend (or weakly depends) on the assumptions in the model, and there is good reason to believe that the result will hold across a wide class of models. Sugden argues, however, that robustness checks keep us within the realm of models. They can justify the inference from one model to a wide array of other models but not to the real world. That is, robustness checks establish internal, but not external, validity. Instead, he argues, we ought to envision models as representing counterfactual worlds that are connected to the real world by a relation of similarity.[6] The more similar the model world is to the real world, the more warrant we have in believing that the causal relationship determined in the model holds in the real world. This is clearly problematic, however, because the most similar possible world to the real world is plausibly complex, while models construct simple worlds. Sugden thus warns that this difference in

[6] Sugden's appeal to possible worlds that are connected to the real world by a relation of similarity mirrors David Lewis's (1973) counterfactual theory of causation. This theory famously employs possible world semantics for counterfactuals. For example, the counterfactual "if A were the case, then B would be the case" is true just in case a possible world where both the antecedent and the consequent are true is closer to the actual world than any world where the antecedent is true and the consequent is false. A relation of similarity between the actual world and possible worlds therefore determines the truth condition of a counterfactual.

complexity shows that "we ought to be very cautious about making inferences from the latter to the former" (Sugden 2002, 129). Sugden argues that we appeal to a different type of similarity when judging the strength of the inference from a model world to the real world. We judge the strength of the inference by assessing whether the model constructs a credible world, meaning we could imagine our world could be like the world of the model. That is, Sugden thinks the credibility of the inductive leap increases the more we think that the model world describes the way the real world could be.

For example, the empirical claim Akerlof infers from the results of his exploration of the model world are justified insofar as we think the model describes a possible used-car market. "We recognize the significance of the similarity between … model markets and real markets by accepting that the model world *could* be real—that it describes a state of affairs that is *credible*, given what we know (or think we know) about the general laws governing events in the real world" (Sugden 2002, 131). Sugden's claim that Akerlof's model describes how the world could be is not compelling, however. The long list of unrealistic assumptions strains the credulity of this model world. Furthermore, Sugden claims that we judge the credibility of a model based on what we already know about the general laws governing events in the world. If this is right, then it is plausible that we think Akerlof's model of the market for lemons describes a credible world only because it confirms our prior beliefs about the used-car market. The model of the market for used cars draws attention to unrecognized implications of what we take to be plausible characteristics of markets. We learn from exploring the model, but a further set of problems need to be addressed to test the outcomes against real-world cases. Sugden's

account of models in economics helps to expose the problem of justifying inferences from the model to the world; however, it does not help us to understand why some inductions from the model world to the real world are deemed sound or explanatory and others are not.

To conclude, economic models are not Galilean idealizations. Economics lacks the background theory required to construct a system from which tendencies can be derived. Cartwright's conclusion that there is nothing more positive to say about economic models is not compelling, however. If economic models are useful at all, then it must be for reasons other than those that make models in physics useful. It is thus compelling that economists, as opposed to physicists, construct model worlds from which they deduce results specific to that construction. As illustrated by Morgan and Sugden, however, it is unclear what economic models can explain about the real world when they are interpreted as constructions. If Sugden is right, then they can only explain insofar as we are justified in making inferences from the model world to the real world. However, the conditions under which this type of inference is justified remain unspecified.

These accounts establish that we are not justified in assuming that untraded activities will have particular features in the world just because they have those features in the model. Attention must be paid to the conditions under which the inference from the model to the world holds when making claims about untraded spillover effects and externalities. These are issues to which economists typically pay little attention. This is apparent in leading microeconomic textbooks that characterize externalities purely in

model terms, where an externality is defined as occurring when an agent's utility function depends on unintended by-products of the activities of another (Mas-Colell, Whinston, and Green 1995; Tietenberg and Lewis 2009).[7] This is also apparent in the work of leading contemporary environmental economists who employ the concept of an externality without analysis of the concept itself or its applicability in different contexts (Stavins 2011; Nordhaus 2014; Holland et al. 2016). Characterizing externalities in terms of their attributes that should be present in both the model and the world, rather than in terms of their characteristics in the model alone, thus offers a more accurate depiction of externalities; furthermore, it offers guidance for interpreting externalities in both the model and in the world.

Additional insights can be drawn about externalities from these views on modeling in economics. Sugden's account helps to reveal some interesting characteristics of the two-agent constrained optimization framework that distinguishes it from the type of model constructed by Akerlof. Similar to Akerlof's model, we can interpret the two-agent constrained optimization framework as constructing an imagined world in which, holding all else constant, the presence of an untraded activity that benefits some individuals and harms others leads to an overproduction of that activity. This is best interpreted in Sugden's predictive form of reasoning about model results. The reasoning would go something like this: In the model world, R (the regularity, which is that untraded activities are overproduced) is caused by F (the set of causal factors, which is

[7] As I explained above, there are no utility functions in the world, per se; utility functions are an index of preferences. The definition of an externality as occurring when an untraded activity enters into a utility function is therefore a definition of an externality as it is modeled in economics.

that there is an activity that benefits some and harms others and yet this activity is untraded); we observe in the world that there are untraded activities that benefit some and harm others, such as the act of polluting. From this, we infer that the regularity is also present in the world (that these untraded activities are overproduced). This is to say that, in the model, it is deduced that the presence of an untraded activity that affects agents external to the production or consumption decision causes an overproduction of the activity. Economists observe that there are untraded activities that have spillover effects in the world. They infer, however, that the regularity (that the untraded activity is overproduced) is present. Interestingly, then, the model for externalities takes a predictive form of reasoning, whereby it is inferred that untraded activities are, in fact, overproduced.

We cannot observe that untraded activities like pollution are overproduced. We can observe that a portion of the population thinks we emit too much pollution, and we can observe that pollution harms the health of people and the environment. But these observations do not necessarily imply that pollution is overproduced, in the economic sense of the term. *Overproduction* means that the quantity of pollution that is produced is not Pareto optimal. It could be the case that many individuals think that we emit too much pollution at the Pareto optimal level. This could occur if individuals do not have a sufficiently high income to pay polluters to reduce pollution, which could be the case if the costs of reducing pollution were high. This would mean that these individuals have insufficiently high willingness to pay, compared to the willingness to accept a payment for the reduction of pollution, to facilitate exchange. Therefore, given willingness to pay

and willingness to accept, the current level of pollution is Pareto optimal. This does not mean, however, that there is not still a desire for a reduction in pollution on the part of individuals that would translate into potential exchange, and thus potential Pareto improvements, if they had a higher income.

The model of externalities therefore defines the sense in which pollution is undesirable; it is undesirable insofar as there are potential Pareto improvements. *Potential* Pareto improvements are not observable; instead, they are inferred to exist because it is assumed that individuals are willing to pay to reduce a harmful activity according to the degree to which it frustrates their preference satisfaction. Therefore, in addition to inferring the causal relationship between the causal factors and the regularity, this model infers that the regularity is present, which is that the activity is overproduced insofar as there exists potential Pareto improvements, and thus a willingness to pay for an untraded activity that is sufficiently high compared to the willingness to accept on the part of producers.

Therefore, Sugden's framework for understanding models highlights an interesting feature of the model of externalities. The two-agent constrained optimization problem posits the nature of the problem it examines; untraded spillover effects are problematic because they generate a Pareto inefficiency. The causal factors in the world are observed, but the particular character of the regularity is posited in the model. Therefore, the solution to an externality problem, which is established in the model, is coherent to the extent that untraded spillover effects, in fact, generate a Pareto inefficiency; a Pigovian tax solves an externality problem by eliminating the Pareto

inefficiency. Therefore, the two-agent constrained optimization model of externalities is best interpreted in a predictive form of reasoning. It is not just the causal relationship between R and F that rests on an inductive inference from the model to the world; R itself is posited in the world using induction.

Morgan's framework for understanding models is also enlightening of the way economists model externalities. Economists insert an unpriced spillover effect into a two-agent constrained optimization problem and, in doing so, they learn about the effect of an unpriced effect in the hypothetical economy. This model is therefore used to question what happens in the presence of an untraded spillover effect, and to establish a solution to the negative outcome it uncovers. That is, the researcher uses the internal dynamics of a constrained optimization framework to determine why unpriced spillover effects are undesirable and to posit solutions. We have seen that they are undesirable because there are untapped gains from exchange that are associated with untapped welfare gains. The solution is thus to exhaust the potential gains, which can be done with a Pigovian tax. Characterizing externalities in terms of their untapped gains from exchange and untapped welfare gains thus works to incorporate into the meaning of externalities the necessary features that a narrative should contain to connect the model world to the real world.

2.7 The Gains View of Externalities

I have argued that the gains view of externalities is implied by the broad view of externalities in the model. I also argued, however, that there is no reason to suppose that this implication holds in the world. In the model, untraded activities that affect other agents invariably generate gains from exchange and welfare gains because of the particular structure of the model. It is not necessarily the case that this will hold in the world, however. There are many types of untraded spillover effects for which individuals have no associated willingness to pay. For example, an individual might be positively affected by observing a beautifully restored nineteenth-century house. As a result, there is an untraded spillover effect between the owner's restoration efforts and the observer's welfare. However, as much as the observer might value well-maintained historic houses, she is not willing to pay for the restoration of privately owned buildings. The unpriced spillover effect of restoring a home therefore counts as an externality on the broad view, but it does not count as an externality for this observer on the gains view. If, however, there are some individuals that are willing to contribute to some of these restoration efforts, then the restoration efforts generate an externality for these individuals on both the gains view and the broad view. Intuitively, then, the two accounts are not equivalent in the world. I will argue in the next chapter that this intuition is well-motivated. What counts as an externality on the gains view is contingent on whether individuals have a willingness to pay for a given unpriced spillover effect.

Unpriced spillover effects in the model generate a particular kind of outcome: they generate untapped gains from exchange that are associated with untapped welfare gains. These features lend themselves to a particular kind of solution because the welfare effects are measurable in terms of gains from exchange. Therefore, a policy tool that exhausts these gains from exchange using measures of willingness to pay can exhaust Pareto improvements. I suggest that we call this kind of untraded spillover effect an *externality*. These are the category of untraded spillover effects for which economic policy solutions are relevant. That is, Pigovian taxes are relevant to the subset of untraded spillover effects that are measurable in terms of willingness to pay that is indicative of welfare gains.

2.8 Conclusion

In this chapter, I developed the gains view of externality. I argued first that a close examination of the model of externalities shows that the economists assume that there is an untraded effect without explaining why it is untraded. I then argued that the process of modelling an untraded spillover effect in a two-agent constrained optimization framework establishes two necessary outcomes of externalities. Untraded spillover effects generate untapped gains from exchange that are associated with untapped welfare gains. These outcomes are crucial to the way economists understand externalities because

they establish the measurability and policy-relevance of externalities, which are necessary features of externalities. That is, these features of externalities generate the policy implications that are inherent to economists' theoretical and practical understanding externalities. Externalities can be "internalized" through a market-based policy such as a Pigovian tax, and doing so establishes an efficient outcome. The value of the Pigovian tax is determined by a measure of the potential gains from exchange over the untraded activity. Therefore, the gains view of externalities holds that an externality should be characterized as untraded spillover effects that generate untapped gains from exchange that are associated with untapped welfare gains.

The broad view of externalities implies the gains view of externalities in the model. That is, all untraded activities in the model generate untapped gains from exchange that are associated with untapped welfare gains. Nonetheless, this does not imply that all untraded spillover effects in the world also have these attributes. In the next chapter, I will argue that not every unpriced spillover effect in the world is an externality, as I have characterized the term. If externalities have the policy implications that economists typically claim they have, then externalities cannot be any untraded spillover effect. Instead, externalities must be untraded activities that generate untapped gains from exchange—this generates the claim that externalities have a dollar value—and welfare gains—this generates the policy relevance of externalities. That is, I will argue that although Satz is right to claim that untraded spillover effects proliferate in the social world, the untraded spillover effects that generate externalities—those for which a market or policy should be established—may rarely be manifest in the world. The notion

of an externality is therefore more limited in the social world than it might appear on the broad view of externalities.

CHAPTER THREE

The Instantiation of Externalities

3.1 Introduction

As I argued in the previous chapter, an externality in the model is an untraded activity that generates gains from exchange that are associated with welfare gains. Whether or not this characterization differs from the broad view of externalities depends on whether gains from exchange are invariably associated with welfare gains in the world, as they are in the model. To hold that the two types of gains are invariably linked in this way rests on a commitment to the view that welfare is the satisfaction of actual preferences. But, as I will show by building on the work of Daniel Hausman (2012), welfare is not merely the satisfaction of preferences. As a result, there is no reason to suppose that gains from exchange are invariably associated with welfare gains. The gains view of externalities therefore counts as externalities only a subset of untraded spillover effects. On this view, externalities are a subset of untraded spillover effects for which there are gains from exchange that are properly related to welfare gains.

There are several problems with the theory that welfare is identical to the satisfaction of preferences. D. Hausman (2012) offers the most compelling arguments. He argues that well-informed, unbiased, and self-interested preferences are evidence for, but not identical to, welfare. This account of welfare, however, poses significant problems for interpreting externalities in the world because it allows for the divergence of gains from exchange and welfare gains. If untapped gains from exchange are based on misinformed preferences, for example, then they are not informative of welfare gains. When this is the case, it is unclear whether there is an externality generated by the untapped gains from exchange, or whether there is an externality that is generated by the untapped welfare gains. As I will argue, an externality requires both types of gains that are properly related.

There are two ways to interpret externalities that satisfy this criterion. First, an externality could be generated only when there are actual untapped gains from exchange that are informative of untapped welfare gains. Second, an externality could be generated when hypothetical untapped gains from exchange can be posited for a given untapped welfare gain. I argue that both interpretations are problematic. The first interpretation characterizes externalities in such a way that they may rarely arise in the world. The second interpretation characterizes externalities in such a way that they are unobservable and ubiquitous.

This argument supports two conclusions. First, it is important to characterize externalities in terms of their relation to gains from exchange and welfare gains because the two types of gains need not be linked in the way posited by economic theory.

Second, it is not clear how to interpret externalities in the world when individuals hold false beliefs or are biased in the relevant context. I speculate that this is a common state of affairs, at least in prominent policy problems like climate change and environmental problems in general. If this is right, then the primary concept that grounds a dominant policy approach to these problems is confused. I thus conclude this chapter by problematizing the concept of an externality without offering a solution. As this chapter will demonstrate, there is no satisfactory definition of the concept of an externality. This in turn poses substantial problems for economic policies such as the carbon tax.

3.2 Problems with the Preference Satisfaction Theory of Welfare

Externalities in the model are cases of untraded activities that generate untapped gains from exchange and untapped welfare gains. Gains from exchange are measured by a divergence between willingness to pay and willingness to accept for a certain activity. Willingness to pay is the dollar value that will make an individual indifferent between paying for the activity and acquiring it, or not paying for the activity and thus not acquiring it. It is the partial derivative of an agent's utility function, subject to their budget constraint. Their utility function is based on a schedule of ordered preferences with a specific ascribed functional form. Reaching a higher utility curve means that more preferences have been satisfied. Satisfying preferences is what constitutes the welfare of agents in the model. Hence the link between gains from exchange and welfare gains in

the model derives from the deductive relationship between willingness to pay and preference, and the assumption that preference satisfaction is identical to welfare.

I have also argued that the constrained-optimization framework establishes attributes that are crucial to the concept of an externality, in the sense that these attributes should be present in both externalities in the model and in the world. Gains from exchange are a necessary feature of externalities that give externalities their magnitude. Welfare gains are a necessary feature of externalities that give externalities their policy relevance. Therefore, understanding the link between gains from exchange and welfare gains is central to understanding the nature of externalities in the world. Gains from exchange and welfare gains are linked via the relationship between willingness to pay, preference satisfaction, and welfare. Therefore, understanding this relationship is of critical importance.

Mainstream economists, for the most part, claim that willingness to pay, preference, and welfare are linked in the world in the same way that they are in the model. First, they assume that willingness to pay meaningfully expresses preference satisfaction. Second, they equate preference satisfaction and welfare (Hausman and McPherson 2006; D. Hausman 2012; Sumner 1996). These two assumptions imply that untapped gains from exchange are invariably linked to untapped welfare gains in the world. That is, any time there is a divergence in an agent's willingness to pay for an activity and another agent's willingness to accept for the same activity, there are potential mutually beneficial exchanges. These exchanges increase the preference satisfaction of both agents. Therefore, these exchanges also increase welfare. Hence, an exchange that

occurs at a price less than one agent's willingness to pay and greater than the other agent's willingness to accept a payment for the activity will increase the welfare of both the agents.

This view of welfare allows for a relatively simple move from claims about the model to claims about the world. An untraded spillover effect, in virtue of affecting an agent and in virtue of the assumption that individuals care only about their welfare, is an effect on the welfare of an agent. An untraded spillover effect that affects the welfare of others does so by increasing or decreasing their preference satisfaction. This is because welfare is understood to be identical to preference satisfaction. Furthermore, preference satisfaction is expressed in terms of a willingness to pay for the activity. That is, an agent is willing to pay just up to the point where they are indifferent between enduring a certain quantity of the welfare-reducing activity and paying to reduce that quantity of the activity. Hence, an untraded spillover effect always has a marginal welfare effect, and this welfare effect is a reduction in preference satisfaction; this reduction in the satisfaction of preferences is expressed in terms of a willingness to pay to mitigate the reduction in preference satisfaction. Therefore, a welfare effect always has an associated potential gain from exchange.

The preference satisfaction theory of welfare therefore drives the claim that externalities in both the model and the world are invariably generated by untraded spillover effects. This is because, on this view of welfare, these effects invariably produce untapped gains from exchange that are identical to untapped welfare gains. The broad view of externalities as untraded spillover effects, then, is consistent with the gains view

of externalities as generating both untapped gains from exchange and welfare gains, provided one holds the preference satisfaction theory of welfare.

In this chapter, I set aside any issues with the assumption that willingness to pay can always express preference satisfaction; I will return to these issues in chapter four. That is, I accept that a positive willingness to pay for an activity indicates that acquiring a specific activity at a lower price will increase preference satisfaction. What matters in this chapter is determining whether gains from exchange can diverge from welfare gains. If preference satisfaction is not identical to welfare, then regardless of our view of the relationship between willingness to pay and preference satisfaction, it must be the case that gains from exchange can diverge from welfare gains. I therefore focus on the assumption that preference satisfaction is identical to welfare. I argue that preference satisfaction is not identical to welfare and, on the contrary, satisfying a preference can be counterproductive to increasing welfare. I then show how this complicates the characterization of externalities in the world because it shows that gains from exchange can be disassociated from welfare gains.

The preference satisfaction theory of welfare holds that the satisfaction of actual preferences constitutes welfare. It is notoriously problematic because individuals are often under-informed and hence unwittingly prefer what is worse for them rather than what is better for them (D. Hausman 2012, 79). Consider a scenario in which individuals live downwind from a newly established factory that emits both particulate matter and steam, both of which are untraded activities. The individuals begin to suffer from frequent headaches shortly after the factory is established. They believe that the steam is

causing their headaches. Consequently, they have a preference and associated willingness to pay to reduce steam emissions; the factory also has a willingness to accept a payment for the reduction in steam emissions because they will accept a payment for the reduction that is equal to or greater than the marginal cost of reducing the steam emissions by installing abatement technology. Therefore, there are untapped gains from exchange over steam emissions.

The willingness to pay of the individuals for a reduction in steam, however, is grounded in a false belief that steam is harmful to them. As it turns out, particulate matter instead is the emission that is giving them headaches and, unbeknown to them, also increases their risk of developing heart disease. Therefore, their manifest preference for steam reductions is not indicative of what will increase their welfare. Presumably, if the agents were better informed about the effect of the emissions on their health, they would prefer the reduction of particulate matter rather than steam. Satisfying these informed preferences would increase their welfare. This means that the untapped gains from exchange over steam emissions are not associated with welfare gains, and the welfare gains from reducing particulate matter emissions are not associated with any actual untapped gains from exchange. Willingness to pay is informative of actual preferences, but actual preferences in this scenario are not informative of welfare gains. Hence the preference satisfaction view of welfare is inadequate when preferences are mistaken.

The argument that untapped gains from exchange invariably imply untapped welfare gains thus rests on a problematic claim that the satisfaction of actual preferences

is identical to welfare. Individuals often have preferences that rest on false beliefs, and when this is the case, preference satisfaction does not coincide with welfare. D. Hausman (2012) suggests three reasons to explain why economists are inclined to take preference satisfaction to be identical to welfare gains. First, he suggests that they are accustomed to idealizing assumptions that individuals are self-interested and perfectly well-informed. Therefore, it is a natural step to take welfare to be the satisfaction of preferences because when these assumptions hold, this is a reasonable view.

Second, economists defend their identification of preference satisfaction and welfare on "the grounds of epistemic and philosophical modesty" (D. Hausman 2012, 80). They do not attempt to say what is good for individuals. Instead, they hold that individuals know what is best for them and express this through their preferences. As D. Hausman (2012, 80) retorts, however, "there is nothing philosophically modest about the claim that preference satisfaction determines well-being."[8] The claim that the satisfaction of actual preferences constitutes welfare tells us exactly what welfare is. Furthermore, this view must be defended against its counterintuitive implications, such as the implication that satisfying misinformed preferences, such as the preference for a decrease in steam emissions, increases welfare.

Third, economists have typically condemned paternalistic policies that coerce people for their own good. If what is good for an individual is what they prefer, however, then there is no possibility to coerce an individual for their own good using welfare-improving policies. Giving individuals what they prefer is not coercive, and on

[8] I follow D. Hausman (2010) in using the terms *welfare* and *well-being* interchangeably.

the preference satisfaction theory of welfare, giving individuals what they prefer constitutes welfare. Therefore, welfare-improving policies, on this view, are not paternalistic. Finally, as D. Hausman observes, the conflation of multiple uses of the terms *utility* and *satisfaction* can account for the adherence of economists to a preference satisfaction theory of welfare. Utility can refer to pleasure or a preference ordering. Satisfaction can refer to a feeling and to whether the world matches an agent's preferences (D. Hausman 2012, 80). Equivocation on both utility and satisfaction can make it tempting to hold that preference satisfaction is identical to well-being because satisfying the preference produces a pleasurable mental state such as happiness or the feeling of satisfaction. But this is an incorrect interpretation of utility in economics. Utility in economics is ordered preferences, and utility is satisfied when the world matches this preference. Therefore, an adequate defence of this view cannot rely on a mental-state interpretation of utility and well-being.

Furthermore, preferences are not always self-interested, and when this is the case, they are not indicative of welfare. Amartya Sen (1977) famously argues that agents form preferences according to non-egoist considerations such as sympathy and commitment. Preferences that are motivated by sympathy acknowledge that one's own welfare depends on the welfare of others. This is consistent with egoism in that satisfying these preferences increases welfare, although they are not structured based on egoistic considerations. Preferences that are motivated by commitment, however, are

inconsistent with the view that satisfying preferences increases welfare.[9] These preferences are formed according to moral concerns without attention to one's own welfare. For example, people make sacrifices in their welfare to protect the environment or boycott certain companies, and these actions are grounded in a commitment to certain moral principles rather than egoistic concerns.

Derek Parfit (1984) also argues that satisfying a preference is sometimes irrelevant to welfare. He gives the following example to motivate this claim. Suppose a person has a preference that a stranger recover from a fatal disease. Unbeknownst to him, the stranger recovers. His preference is therefore satisfied because the state of affairs matches his preference, and yet this has no bearing on the welfare of this person because the preference has nothing to do with him. That is, the satisfaction of a preference that is not self-directed does not bear on welfare.

Preferences also sometimes conflict. When this is the case, the satisfaction of actual preferences cannot indicate what is better for a person. For example, an agent might prefer to purchase pre-made meals that produce a significant amount of plastic waste, but also prefer to reduce the amount of plastic waste they create. The preference satisfaction view of welfare cannot privilege one preference over another, and therefore cannot account for welfare when preferences conflict. On Sen's (1977) view, we could explain this type of conflict to stem from the agent having distinct preference rankings. The former preference stems from egoistic considerations while the latter stems from

[9] In chapter four, I argue that this is not necessarily the case. For the purposes of this chapter, however, I concede that preferences based on a commitment to moral principles are not properly related to welfare.

moral considerations. But the preference satisfaction theory cannot account for welfare when this happens because satisfying one preferences results in the frustration of the other preferences. Therefore, satisfying either preference seems to both increase and decrease the welfare of the agent.

A related issue to the problem of conflicting preferences is the philosophical problem of akrasia. Historically, the problem of akrasia was also called the problem of weakness of will. However, Richard Holton (1999) argues that weakness of will is a failure to act on one's (past) intentions. This is different from akrasia, however, since an akratic act is defined as an act that is contrary to one's (current) better judgment. The problem of akrasia thus concerns whether it is, strictly speaking, possible for agents to act against their best judgment. For example, an agent buys pre-made meals rather than cooking a meal from scratch even though she believes that cooking a meal from scratch was the best thing to do, all things considered. The problem is whether this analysis of such an action is veridical. Plato maintained that, "no one who either knows or believes that there is another possible course of action, better than the one he is following, will ever continue on his present course" (Protagoras 358b-c, cited in Stroud 2014). R. M. Hare (1952) similarly denies that a person can genuinely act contrary to what they judge to be the best course of action, all things considered. If they appear to succumb to akrasia, then we ought to believe that they are not genuinely acting in accordance with their best judgment. They might not be fully capable of acting otherwise, like Ulysses and the Sirens (Elster 1979). Hare argues that evaluative judgment leads to action, and it

follows from an individual having done an action that she judged that action to be the best available option for her in the moment (Stroud 2014).

This view reaches a similar conclusion to revealed preference theory in economics, which states that whatever agents choose must be what they prefer. The weak axiom of revealed preference stipulates that the manifest preference is identified with choice. If economists hold that preference is a judgment by an agent of what is best for her, all things considered, as D. Hausman (2012) argues, then revealed preference theory denies the possibility of akrasia. D. Hausman criticizes economists for not even considering the role of false beliefs about the available options when they identify preference with choice. Economists hold that if a person chooses x over y, then they must prefer x to y regardless of whether they falsely believed y was a feasible option (D. Hausman 2012, 27-28). That is, economists do not in general allow for counter-preferential choices.

Donald Davidson (1980) argues that akratic action is possible but irrational. He suggests that the inclination to deny its possibility is the result of ignoring the importance of the *all things considered* clause in the definition of akratic action. He argues that practical reasoning starts from what he calls prima facie judgments. These judgments focus on particular ways in which one option might be better than another. They do not commit an agent to an overall judgment of betterness. He argues that, as our practical reasoning progresses, however, we eventually have to compare options in light of several considerations, rather than one. This leads to an all-things-considered judgment. But this judgment is still relational in the sense that it is derived from the particular ways in which

the options compare. It is possible to make an all-things-considered judgment (conditional judgment) without making a corresponding overall evaluative judgment (unconditional judgment) of the options (Stroud 2014).

His position rests on the view that acting contrary to unconditional evaluations is impossible. That is, one cannot choose A over B when she judges B to be better than A according to an overall evaluative judgment. But they can act contrary to conditional evaluations. That is, they might judge A to be better than B, all-things-considered, but think that they might not have all the evidence necessary to act on this judgment. Thus, if an agent conditionally judges that A is better than B, but unconditionally judges that B is better than A, then they choose B. Akrasia thus occurs when an agent does not act in accordance with their conditional judgment, which is an all-things-considered judgment. He argues that this is practically irrational, and agents should act in accordance with all the relevant and available evidence; that is, it is rational to act in accordance with conditional judgments.

Following Davidson (1980), the general view in philosophy is that akratic action is possible (Tappolet 2003). The puzzle of akratic action has thus evolved from a puzzle of its existence to a puzzle of how it fits within theories of practical reasoning and agency. Indeed, action contrary to one's best judgment seems like a common phenomenon. John Searle (2001, 220) questions "why anyone would doubt or even be puzzled by the possibility [of akrasia] since in real life it is so common." Sarah Stroud (2014) states that the all-too-common occurrence of remaining on the couch instead of grading papers is an instance of akratic action. If akratic action is indeed possible and furthermore if it is

common, which seems plausible, then actual preferences that are revealed through the choices of agents often do not indicate what is better for an agent, even according to the judgment of agents about what is better for them.

Preferences also change over time, and the preference satisfaction theory of well-being cannot explain the intuition that satisfying past preferences does not increase welfare. That is, it cannot explain why current preferences take priority over old preferences (D. Hausman et al. 2017). D. Hausman (2012) offers two explanations for this intuition, both of which indicate that preference satisfaction cannot constitute welfare. First, old preferences are irrelevant to welfare because as we learn more, we update our preferences. Our preferences evolve as we improve at judging the sorts of things that make us better off. Therefore, satisfying current preferences should take precedence over past preferences. This answer assumes, however, that preferences are determined by a judgment of what will make one's life go better. Therefore, it cannot be the case that merely satisfying preferences makes our lives go better; something is not better for us just because we prefer it. Instead, our judgments of what make us better off determine our preferences. Welfare, while correlated with preference satisfaction, is better understood as distinct from it.

The second explanation for the intuition that satisfying old preferences does not increase welfare is that satisfying an agent's past preferences means that you are giving them what they no longer want. If you give someone what they no longer want, then you are giving them something that will no longer give them satisfaction (D. Hausman et al. 2017, 134). We want different things at different points in our lives, and past wants are

no longer relevant to the sorts of things that we currently want. This implies that we care about getting what we want because it gives us a feeling of satisfaction. But this answer assumes that the *feeling* of satisfaction we have from getting what we want, not preference satisfaction, makes us better off. Such a claim marks a shift from a preference satisfaction theory of welfare to a mental-state theory of welfare. That is, it implies that the mental state that is caused by having a preference satisfied is what is intrinsically good for an individual. For example, suppose an agent wanted a new SUV last year. They never got that new car, and this year decided that they now want a new electric car. On the mental state view, getting the electric car, as opposed to getting the SUV, is good for this person because the feeling of satisfaction of getting the thing they want is what makes their life go better. Since they no longer want the SUV, getting it will not give them the pleasurable feeling of getting what they want. On the preference satisfaction view of welfare, however, satisfying a preference—meaning that the state of affairs matches the preference—is what constitutes welfare rather than any feeling. It is not clear, on the preference satisfaction theory of welfare, why it is the case that the world matching a new preference, but not an old preference, increases welfare.

Preference change is therefore a theoretical problem for the preference satisfaction theory of welfare because it cannot explain why satisfying current preferences should take precedence over satisfying past preferences. Explanations for this intuition seem to devolve into an account of welfare that is distinct from merely satisfying preferences. As Hausman, McPherson, and Satz (2017, 134) observe, preference change is more than a theoretical problem for the preference satisfaction view

of welfare. It is also a practical problem for setting policies according to preferences. Policies and institutions change the preferences of the people at which these government interventions are directed. This could occur through advertising or propaganda, for example. But policies and institutions also change the context in which people form their preferences. The policies that are supposed to be determined by preferences thus change those same preferences. It seems, then, that policies might not truly address current preferences, which are the preferences that are relevant to the individuals. If welfare is equivalent to the satisfaction of actual preferences, then it is unclear how policies can be directed at increasing welfare when actual preferences, and thus the sorts of things that might increase welfare, change in the context of new policies and institutions. Hausman, McPherson, and Satz (2017, 135) thus question, "how concerned should one be about satisfying current preferences if one judges that they are likely to change? Should one aim to modify preferences so that they will be easier to satisfy? How should one choose between either satisfying existing preferences or modifying preferences first and then satisfying the modified preferences?"

There has been a move among some economists and psychologists, including Daniel Kahneman, Robert Sugden, and Richard Thaler, to reject the preference satisfaction theory of welfare in favour of what D. Hausman (2010) calls a "new hedonist economics." New hedonists suggest that economic theory should shift from a preference satisfaction theory of welfare to the view that increasing happiness improves welfare. These economists are typically influenced by findings in behavioural economics that show systematic errors in judgements made by individuals about their welfare. That is,

people do not merely make mistakes about what will make them better off; they systematically misjudge what will make them better off in some contexts. For example, work in behavioural economics has suggested that people often have a status quo bias whereby individuals adhere to the status quo more often than is predicted by rational choice theory (Samuelson and Zeckhauser 1988; Thaler and Benartzi 2004). This was demonstrated through a study of the effect of the default option on the participation rate in a retirement savings plan (Samuelson and Zeckhauser 1988). As it turned out, what was set as the default option has a significant effect on the choice made by individuals about retirement savings. If the default was to not contribute to a retirement savings plan, individuals typically saved less toward retirement. If the default option was to contribute substantially to retirement savings plans, then individuals did not opt out and thus saved significantly more towards their retirement (Thaler and Sunstein 2008). Hence, preferences in this context are a poor indicator of what is the welfare-improving option for an individual and, instead, their behaviour reflects no settled preferences on the matter.

It is thus unsurprising that economists who are influenced by the findings in behavioural economics take issue with the view that preference satisfaction is constitutive of welfare. Economists such as Sugden, or psychologists such as Kahneman, maintain that there are only two options for interpreting utility. Either utility is a representation of preferences, as is the case in contemporary economic theory, or utility is a measure of net pleasure, as was the case among the early neoclassical economists such as William Stanley Jevons and Francis Ysidro Edgeworth. But accepting the

"dichotomy between preference and pleasure leaves welfare economists with an unappetizing choice. Either they maintain (absurdly) that people never make mistakes—whatever people choose or prefer is best for them—or they can climb out on a shaky philosophical limb and espouse a hedonist view of human welfare" (D. Hausman 2010, 326).

D. Hausman (2010) argues that welfare economics should not shift to a mental-state theory of welfare that holds that the goal of economic policy is to make people happy. This just raises more problems than not, since it requires a definition of happiness. The two notions of happiness that economists have considered are attitudinal happiness, where happiness is a judgment or an attitude about how one's life is unfolding, and experienced happiness, where happiness is a feeling or a matter of one's mood (D. Hausman 2010, 333). This difference has empirical consequences, and there are cases when agents report inconsistent accounts of their happiness depending on which definition they adopt. In order to avoid this outcome, the new hedonists have argued that, to get an accurate picture of well-being, they must measure *moment utility*, which is the quality of a momentary experience, and aggregate the measures utility in a single moment to calculate total utility.

D. Hausman (2010) argues, however, that there are significant problems with this view of welfare. One problem he identifies is that the policy interventions implied by the findings of the new hedonists are likely not welfare-improving. They recommend that people reduce the time they spend doing things they find unpleasant in the moment; their studies find that such activities include both going to the dentist and caring for

one's children. But this would result in a myopic policy of maximizing current net pleasure, and that is not likely to maximize happiness over a lifetime. Especially with non-consumption goods, individuals invest in future satisfactions. Individuals endure activities they do not find pleasurable in the moment in order to increase their long-term well-being. Reducing the time doing these activities is consistent with the new hedonist proposal—for example, reducing the time parents spend with their young children—and yet this policy would have a long-run effect of reducing welfare. Hausman (2012, 336-7) argues that people's preferences are determined by their own calculations of tradeoffs between the present and the future. These calculations, he argues, are not all uninformed or biased, as the new hedonists suppose. Individuals do things they expect to be unpleasant now because they expect that it will make them better off in the long run. The notion of welfare as preference satisfaction therefore captures the tradeoff we make between momentary pleasures and long-run well-being. Happiness, understood as momentary pleasures or feelings of satisfaction, cannot capture such trade-offs over time or our general assessments of how to make investments now to improve our lives in the long run. Happiness thus defined fails to capture our general sense of well-being (Angner 2013).

D. Hausman argues that a substantial oversight of the new-hedonists is that they ignore the most common view of welfare among philosophers. This is the view that welfare consists in preferences that are sufficiently "cleaned up." Self-directed preferences that are informed and rational, that is, preferences that are sufficiently "laundered" or "purified", constitute welfare (Gauthier 1986; Goodin 1986). Although

welfare consists in the satisfaction of laundered preferences, what is taken to be sufficiently laundered varies in the scholarly literature. Typically, these preferences must be self-directed and well-informed. Larry Sumner (1996, 122) claims that the identification of laundered preferences with welfare has "achieved the status of an unquestioned axiom," and enlists the views of several prominent moral philosophers, such as John Rawls and R. M. Hare, to support his claim.

Taking the satisfaction of laundered or purified preferences to be identical to welfare solves the most serious problems with the preference satisfaction view of welfare. For example, laundering preferences eliminates the problem false beliefs. The satisfaction of preferences that are formed on the basis of a false belief or insufficient information do not constitute welfare. While actual preferences evolve as agents learn about what makes them better off, purified preferences do not change because they are fully informed. Furthermore, suitably purified preferences never conflict because they are fully informed and rationally formed. D. Hausman (2012, 84) suggests that we may want to take anti-social preferences, such as racist or sadistic preferences, to be mistaken preferences that should be purified away. This is more controversial than the type of laundering that targets obviously mistaken beliefs or incomplete information. This is because it purifies away something that might be closer to an agent's values rather than mere mistakes in forming preferences. However, arguably, these anti-social preferences cannot withstand rational scrutiny and therefore would be eliminated when preferences are purified.

The purified preference view, as opposed to the view that takes the satisfaction of actual preferences, assumes that researchers know something about the welfare of agents that is unapparent to the agents themselves. It is not what someone actually prefers that is equivalent to welfare. Instead, it is what someone should prefer that is equivalent to welfare. For example, Peter Railton (1986, 16) argues that "an individual's good consists in what he would want himself to want, or to pursue, were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free from cognitive error or lapses of instrumental rationality." Researchers thus launder or purify actual preferences in line with the correct information and the sorts of things they think agents would pursue if they were perfectly rational in their pursuit of increasing their welfare.

"To take well-being to be the satisfaction of purified rather than actual preferences shifts the emphasis of what people do prefer to what they should prefer" (D. Hausman 2012, 86). We can only purify preferences to the degree that we think we know something about the agent's well-being that is unapparent to them—we know what they should prefer in order to make themselves better off. The purification of preferences thus introduces an objective list view of welfare into preference satisfaction views of welfare (D. Hausman 2012, 86). Objective list theories, broadly speaking, provide lists of goods that improve the welfare of everyone regardless of whether anyone actually desires those good. For example, individuals who have strong friendships are better off than individuals who do not, regardless of whether these people want strong friendships (Fletcher 2016). On this view, then, we can claim that the agents in the steam and

particulate matter example should prefer a reduction in particulate matter because we know, unlike the agents, that this will improve their health, and improving health increases everyone's well-being. So much for their own perceptions of their well-being. They are simply wrong about what will make them better off.

The purified preference view avoids the problems of the actual preference view of welfare at the cost of making well-being less measurable. If economists accept this view, they would have to assess the extent to which actual preferences diverge from purified preferences. This is problematic because purified preferences are unobservable, and therefore would have to be constructed by the researcher. It may be reasonable, however, to take actual preferences as a proxy for purified preferences (D. Hausman 2012, 84). This seems reasonable, however, only when actual preferences are sufficiently well-informed, unbiased, and self-interested. Therefore, researchers must have an idea of what counts as a sufficiently informed, unbiased, and self-interested preference on the purified preference theory of welfare. As I will show in chapter four, they often do not have such a well-formed idea about what counts as a sufficiently purified preference.

Furthermore, the proxy account is problematic in cases where preferences are significantly mistaken. The preference for reduced steam emission cannot be a proxy for a purified preference insofar as the object of the preference is entirely mistaken. That is, the actual preference for reduced steam cannot be a proxy for the purified preference for reduced steam, because purifying preferences eliminates any preference over the quantity of steam released. This actual preference might instead be seen as evidence for a preference for reducing whatever is causing the poor health of the individuals. Therefore,

we might re-interpret the actual preference for steam reductions as an actual preference for a reduction in harmful emissions, and thus take this as a proxy for the purified preference for a reduction in harmful emissions. This is not simply taking the actual preference as a proxy, but instead it is using an actual preference as a guide for how to construct the missing preferences for a decrease in particulate matter emissions

There may not always be actual preferences to use as a guide or proxy when estimating purified preferences. For example, the steam and particulate matter example can be altered so that the only harmful effect of the particulate matter emissions is an increase in the chances of developing lung cancer, of which the affected individuals are unaware. Therefore, they have no actual preferences for reduced steam or particulate matter emissions. However, they would prefer a reduction in particulate matter if they knew about these harmful effects. There is no actual preference in this case to act as a proxy or guide to constructing laundered preferences for a reduction in particulate matter emissions. In this setting, it is unclear how economists could construct preferences and the associated willingness to pay for particulate matter emissions in order to measure welfare.

D. Hausman (2012) thinks that the real problem with the proxy view, however, is that all preference satisfaction theories of welfare hold a mistaken view of welfare. "The fact that Jack prefers x to y does not make it the case that x is better for him than y, no matter what conditions one imposes on his preferences" (D. Hausman 2012, 84). Therefore, according to D. Hausman, it is never appropriate to take the satisfaction of actual preferences to be a proxy for welfare gains because the satisfaction of preferences,

no matter how purified, is not identical to welfare gains. The first argument D. Hausman gives for the distinction between purified preference satisfaction and welfare takes aim at Parfit's (1984) version of the purified preference view of welfare. Parfit specifies that *self-directed* informed preferences constitute welfare. These are the preferences that are concerned with our own lives (Parfit 1984). It is plausible that some preferences are clearly about the lives of other people, and in virtue of this, do not bear on the welfare of the individual who holds the preference. To use the same example, if I have a preference for a stranger to recover from a terminal illness, and, unbeknownst to me this stranger recovers, it does not make my life go any better.

Parfit's view that self-directed preferences constitute welfare is not plausible, however, when we consider preferences that are directed at individuals that are involved in one's life. For example, if, unbeknownst to Jill, her daughter becomes a thief, then her welfare decreases because in this state of affairs, her preference to be a good parent is frustrated. That is, her welfare decreases because a preference she holds that is about her own life—being a good parent— is frustrated. On the other hand, Parfit argues that if unbeknownst to Jack, his child dies, then his welfare does not decrease. This is because the preference that is frustrated—that his child live a long life— is not about Jack's own life. This is a counterintuitive result of the view that self-directed preferences constitute welfare. It is hard to see why Jill's life is worse because she has a criminal daughter but Jack's life is not worse when his child dies. Parfit thinks this is the case because Jill's preference is self-directed and Jack's preference is other-directed. But this distinction does not seem to track our intuitions about what makes a life go well. Does this imply

that Jack's preference is really self-directed rather than directed toward the welfare of his child, or does it mean that other-directed preferences are sometimes relevant to one's welfare? It is difficult to determine which preferences are about one's life and which are not.

D. Hausman (2012) argues that we need not worry about clarifying this distinction because it is the wrong distinction to draw in the first place. Whether preferences are self- or other-directed is not relevant to the determination of which states of affairs increase or decrease welfare. This is demonstrated by a preference for self-harm, which is self-directed but also welfare decreasing. Similarly, the trivial preference that one be descended from Charlemagne is self-directed but irrelevant to welfare (D. Hausman 2012, 85-6). There is little reason to think that the satisfaction of any self-directed preference increases the welfare of the agent. D. Hausman (2012, 84) argues that "the distinction that is needed lies between those of Jack's preferences that are directed toward promoting his own well-being and those that are not." If Jack's preferences are formed on the basis of making his own life go better, then these preferences are good evidence for what will increase his welfare. But this presupposes a conception of welfare that is independent of preference satisfaction. Jack can form his preference on the basis of increasing his welfare only if welfare is something distinct from preference satisfaction. As D. Hausman concludes, preference satisfaction must be distinct from welfare.

This view gives a more intuitive ruling on the welfare of Jack and Jill. Although Jack's preference that his child remain alive is a preference that is not self-directed, it can

be interpreted on D. Hausman's view as being directed at his own welfare. This is because Jack's welfare is affected by the welfare of individuals for whom he cares. Sympathetic preferences bear on one's welfare (Sen 1977). It would certainly make his life go worse if his child died; therefore, his preference that his child stay alive is, perhaps in part, self-interested but not self-directed. I agree with D. Hausman that the distinction between self- and other-directed preferences does not track the sorts of preferences that enhance welfare. I do not agree, however, that replacing this with the distinction between self-interested and not self-interested preferences solves this problem. It is plausible that Jack's preference is not directed at increasing his own welfare even though satisfying this preference does in fact increase his welfare. Since this issue does not bear on the aim of this chapter, however, which is to show that the satisfaction of actual preferences is not identical to welfare, I set it aside until chapter four.

D. Hausman's (2012, 86) second argument against the identification of purified preference satisfaction with welfare is that welfare, unlike preferences satisfaction alone, has "moral pull." If individuals want something just for the sake of getting it, then their getting that thing is of no moral importance to others; that is, they have no reason to help this person get what they want (Nagel 1986; D. Hausman 2012, 86). If someone wants something, then we want to know the reasons they have for wanting it. If we deem these to be good reasons for wanting something, then we have reason to think that it is good that they satisfy this preference. D. Hausman thinks that a good reason for wanting something is that getting it makes one's life go better. We have reason to help people increase their welfare, but not to help them satisfy any want. Hence, not all preferences

are indicative of welfare gains. This argument thus establishes that there must be a notion of welfare that is distinct from the satisfaction of preferences. Furthermore, this argument also shows that welfare, not merely the satisfaction of preferences, is of moral importance. Indeed, it is convincing that once we bar the conflation of the good feeling of having one's preferences satisfied with the state of the affairs satisfying one's preferences, *simpliciter,* it is hard to see why satisfying any preference, however purified, is good except insofar as satisfying these preferences contributes to well-being.

It is plausible that individuals typically have reasons for their preferences. Someone might want a new computer just because it is new. That is, they derive pleasure from replacing their old computer with the latest model not because it offers more computing power but simply because it is the latest model. Or they might derive satisfaction from the sheer fact that they can get what they want. But all of these reasons seem as though they could be directed at enhancing one's life. If we are to distinguish between the preferences that are directed at increasing welfare from those that are not, by assessing the reasons one has for holding the preferences, then we must have as a background condition some idea about what constitutes welfare. D. Hausman thinks, however, that we do not need a full-fledged theory of well-being to determine what makes the lives of individuals go well. We have a good general idea of what sort of things make our lives go well—friends, health, or successful accomplishments of worthwhile projects—despite not securing a well-confirmed theory of welfare (D. Hausman 2012, 92). If we think that individuals derive pleasure from getting what they want and that pleasure contributes to welfare, however, then it seems that most preferences that are

well-informed are also evidence for welfare gains. This might explain why economists have conflated preference satisfaction and welfare; it is often the case that satisfying preferences are evidence for welfare gains because individuals feel pleased when they get what they want, and pleasure contributes to welfare (D. Hausman 2012, 87). Nonetheless, they are distinct, as we have seen.

The satisfaction of purified preferences and welfare gains are correlated because they are often derived from common sources. When individuals are self-interested and good judges of what enhances their well-being, both goodness and happiness help to determine their preferences as well as to enhance their welfare. People structure their lives and projects so as to aim to make their lives go better, and thus form their preferences accordingly. Therefore, preference satisfaction is not equivalent to welfare and does not necessarily imply welfare gains. Nonetheless, preference satisfaction is often evidence for welfare gains because the pursuit of welfare often guides the formation of preferences. "When people are self-interested, their preferences will match what they believe will benefit them. If with respect to the matters at hand individuals are good judges of what will benefit them, then economists can use people's preferences as evidence concerning what promotes their welfare" (D. Hausman 2012, 89).

D. Hausman thus demonstrates the merits of an evidential view of the relationship between preference satisfaction and welfare gains. On this view, the only substantial claim we need to make about welfare is that it is not the satisfaction of preferences, however purified theses preferences might be. Therefore, we can be neutral, to an extent, on what actually constitutes welfare. The right sort of preference is evidence

for welfare, regardless of what welfare is. When individuals are self-interested, they prefer x to y because they believe that x will make their lives go better. When individuals are good judges of what will benefit them, then preferences are good evidence of what will increase their welfare (D. Hausman 2012, 88-89). This is to say that people prefer things because it is good for them, but it is not good for them because they prefer it. Therefore, preferences are evidence for what will promote welfare when preferences are self-interested, and individuals are well-informed and free of significant biases concerning the good in question. The evidential account thus shows that deferral to preferences is not always appropriate when assessing the welfare of agents. If the mistakes of individuals are systematic, so that they do not cancel out over a large population, or where preferences are systematically distorted by biases, or where preferences are not self-interested, then the preferences of individuals cannot be used as evidence for what is good for them. Therefore, when an individual forms a preference directed at increasing their welfare and yet they are wrong about what makes their life go well, then this preference is not relevant to their welfare.

What matters the most for the characterization of externalities in the world is that the satisfaction of actual preferences is not identical to welfare. This is important for externalities because gains from exchange are determined by willingness to pay, and willingness to pay is derived from the actual preferences of individuals. Since the satisfaction of actual preferences, however, is not identical to welfare, gains from exchange are not invariably associated with welfare gains. When gains from exchange are derived from misinformed or biased preferences, they are uninformative of welfare gains.

This holds regardless of whether we take the satisfaction of purified preferences to be identical to welfare or to be evidence for welfare.

On the purified preference view, actual preferences are equivalent to welfare if agents are sufficiently informed, unbiased, and rational. When this is the case, gains from exchange will track welfare gains. When this is not the case, gains from exchange will not track welfare gains. On the evidential view, actual preferences are evidence for welfare when agents are sufficiently informed, unbiased, and self-interested. When this is the case, gains from exchange are evidence for welfare gains. On both views, gains from exchange are derived from willingness to pay and willingness to accept for an untraded activity. This means that they are not identical to, but are derived from, actual preferences. Therefore, on both views, gains from exchange are not equivalent to welfare gains. However, there is a stronger connection between the two types of gains on the purified preference view. Since the satisfaction of purified preferences constitutes welfare, willingness to pay is therefore a measure of preference satisfaction and welfare. Therefore, gains from exchange are a measure of welfare gains. But on the evidential view, gains from exchange is a measure of preference satisfaction but is only evidence for welfare gains.

In sum, I have argued in support of two claims. First, externalities in the model occur when an untraded activity generates untapped gains from exchange and untapped welfare gains. In the model, these two types of gains never diverge because economic agents in the model are perfectly informed, rational, and self-interested. Second, untapped gains from exchange can diverge from welfare gains in the world because

welfare is not the satisfaction of actual preferences. When individuals are mistaken or biased, their preferences are uninformative of their welfare gains. Therefore, their willingness to pay, which is derived from their preferences, is also unconnected to welfare. In this case, untapped gains from exchange are uninformative of welfare gains. In the next section, I illustrate the problems this presents for interpreting externalities in the world.

3.3 Interpreting Externalities in the World

On the gains view, an externality is an untraded activity that generates both untapped gains from exchange and untapped welfare gains. The presence of untapped gains from exchange establishes the measurability of externalities, and the untapped welfare gains generates the policy relevance of externalities. In the world, however, untapped gains from exchange can diverge from welfare gains because actual preferences are sometimes uninformative of welfare, as I argued in the previous section. That is, the satisfaction of actual preferences is not identical to welfare gains, and they diverge when preferences are uninformed or irrational.

Recall that in the steam and particulate matter example, the agents express a specific preference to reduce steam emissions, and consequently there are gains from exchange over steam emissions. However, no welfare gains actually transpire from reducing steam because their preferences are grounded in the mistaken belief that steam

is bad for their health. Reducing particulate matter emissions would instead improve their welfare, but there is no willingness to pay among the agents for a reduction in particulate matter emissions. Hence, there are untapped gains from exchange over steam emissions with no associated welfare gains, and there are untapped welfare gains over particulate matter emissions with no associated gains from exchange. The question thus remains whether an externality is generated at all when actual preferences are uninformative of welfare, and if so, then whether it is the gains from exchange or the welfare gains that is associated with the emergence of an externality.

The first option is that the externality is generated by the steam emissions. On this interpretation, externalities necessarily involve untapped gains from exchange regardless of whether they are informative of welfare gains. Like markets, externalities are generated by actual preferences. That is, externalities are a sort of latent market in which the psychological conditions necessary for the formation of a market are present—a willingness to exchange on the part of the supplier and consumer—and yet no market forms. This interpretation aligns with Ronald Coase's ([1960] 2013) framework for understanding externalities. He argued that externalities exist only when exchange is obstructed by high transaction costs or ill-defined property rights. But for these factors that prevent exchange from occurring, a market would form because agents have a willingness to pay and a willingness to accept a payment for a reduction in steam emissions that is conducive to exchange. Therefore, the externality would be eliminated through exchange if transaction costs were lower or if there were well-defined property rights. The value of the externality is the equilibrium price that would emerge in this

market. Hence, this outcome gives externalities their measurability in terms of willingness to pay. This is an important feature of externalities because the policy response to externalities is to internalize the externality by setting the optimal price on the untraded good (Hahn and Stavins 1992; Stavins 2011). Such a market-based policy thus mimics the outcome of a market that would exist if exchange were unhindered.

The sole purpose of policy in economics, however, is to increase welfare. This often takes the form of seeking out Pareto improvements (D. Hausman 2010). There is no reason to create a policy that mimics the outcome of a market, therefore, if there are only gains from exchange but no welfare gains. Steam emissions therefore cannot generate an externality, because externalities, as I have characterized them, are necessarily policy relevant.

Steam emissions cause untraded effects that generate untapped gains from exchange but no welfare gains. This type of untraded effect could be called a *policy-irrelevant externality*; however, a policy-irrelevant externality is a trivial concept. It does little more than say that more (mistaken or biased) preferences could be satisfied if exchange occurred over an untraded good for which there is a willingness by each of the two parties to engage in exchange. This does little conceptual work beyond what the concepts of supply and demand already offer.

There might be interesting things to learn about markets and latent markets that are irrelevant to the welfare of individuals, but this is not relevant to the present inquiry. The type of externalities that are given attention in economics are the ones that should be internalized through a market-based policy response. As such, I use the term *externality*

to refer to policy-relevant externalities. Steam emissions are untraded spillover effects, but they are not externalities, in this sense, because they are irrelevant to policy. *Untraded spillover effect* thus refers to a broad category that includes both policy-irrelevant and policy-relevant untraded spillover effects.

Untapped welfare gains are therefore indispensable to externalities in economics. Consider the interpretation of an externality as an untraded activity that generates untapped welfare gains, regardless of whether it generates untapped gains from exchange. That is, the particulate matter emissions generate an externality in the previous example. Interestingly, this interpretation is synonymous with the broad view of externalities as untraded spillover effects. Therefore, it has the same problems as the broad conception of externalities. If an untraded activity generates an externality whenever it affects another agent's welfare, then externalities would be an incredibly broad concept. Minute social interactions, such as observing a stranger smile, would count as an externality. Therefore, the concept would offer no way to distinguish between externalities that seem policy-relevant, like pollution, and those that seem irrelevant to policy interventions, like a smiling agent. Indeed, a policy intervention into many supposed externalities generated by social interactions would be entirely inappropriate.

Importantly, however, externalities cannot be generated without any regard to gains from exchange. This is because externalities are measurable, which is a feature of the presence of untapped gains from exchange. If there are welfare gains over smiling strangers but there are no plausible gains from exchange over this unpriced activity, then

there is no externality generated by this activity that has an associated market-based policy. Understanding externalities as they are manifest in the world and conjoining this with a more plausible theory of welfare which is distinct from actual preferences therefore constrains the scope of externalities. Externalities are a subset of untraded spillover effects for which there are both welfare gains and gains from exchange. If a person is not willing to pay for the smile they observe, despite deriving some welfare benefit from this smile, then it cannot be considered an externality.

If we interpret externalities to be generated by gains from exchange without any necessary link to welfare gains, then they capture the trivial idea that there are always gains from exchange to be made insofar as individuals wish to exchange. If externalities are generated by welfare gains without any necessary link to gains from exchange, then we are left with the trivial, if not tautological, claim that welfare is increased by reducing an activity that reduces welfare. It seems better, for the purposes of both theory and policy, to forge a concept of an externality that captures the idea that some welfare-reducing activities could be addressed by exhausting the associated potential gains from exchange.

3.3.1 Externalities arise when actual preferences are sufficiently informed

As I argued in the previous section, welfare gains are essential to policy-relevant externalities; however, externalities also have a magnitude. Therefore, a sufficient explanation of externalities in the world should characterize them as generating both

untapped welfare gains and untapped gains from exchange. There are two ways to interpret externalities in the world that meet these criteria.

The first way to interpret externalities is consistent with the way I discussed externalities in the world in the previous section: externalities arise only when an untraded activity generates actual gains from exchange that are informative of untapped welfare gains. That is, a smiling stranger does not generate an externality because there is no willingness to pay or accept a payment for this type of activity. On this view, then, externalities arise only when actual preferences are reliable guides to welfare and the agents have a willingness to pay and accept a payment for the activity that are derived from the actual preferences of the agents. On this view, there is no externality present in the example of the steam and particulate matter emissions. This means that economists do not have the resources to recommend a precise policy response to increase welfare by reducing pollution in this context. The best they can do is to recommend some positive tax on particulate matter emissions because, given the right conditions, a tax on a welfare-reducing activity is a cost-effective way of inducing producers to reduce the quantity of the activity. But they cannot specify how much the activity should be reduced, or how to weigh the trade-off between the benefits derived by producers and the costs incurred by external agents caused by the production. If the pollution generates an externality, on the other hand, then economists have the resources to recommend a tax that gives a precise solution for weighing these costs and benefits: set the tax where marginal cost equals marginal benefit.

Suppose that the agents in the example instead held the correct belief that the particulate matter causes harm to their health. Accordingly, they have a willingness to pay to reduce particulate matter emissions that is derived from their actual preferences. Are actual preferences sufficiently good evidence of welfare when agents do not have blatantly false causal beliefs? They might have the correct causal belief about what is causing their headaches but still not know that long-run exposure to particulate matter causes other health concerns like an increased risk of chronic lung disease. Or, they could know that they are at increased risk of these conditions, but have a present bias such that they underestimate how much developing these health issues later in life will reduce their overall welfare. Or, they could also have a poor grasp of probability, like many of us, so that their preferences do not respond to this type of probabilistic evidence.

Even if an individual understood the probability of developing these disorders given their exposure to particulate matter emissions, the probabilities that are established in health studies are statistical averages. Is an average taken from a large population informative of an individual's risk? Did it attend to the specific exposure, genetic record, or other related environmental factors, that might also inform preferences? If to be informed means that one has true beliefs and accurate prior probabilities about future effects, then actual preferences would never be sufficiently informed to constitute evidence of welfare. If preferences were never sufficiently informed, then externalities would never be manifest in the world because gains from exchange would never constitute sufficiently strong evidence for welfare gains. The more stringent the

requirements are on informed actual preferences, the less likely it is that externalities will arise in the world.

Regardless of what counts as sufficiently informed, if agents have blatantly false beliefs regarding the effect of the activity in question on their welfare, then it is clear that they have insufficiently informed actual preferences for the manifestation of an externality. This view of externalities therefore constrains externalities in the world to untraded activities that generate untapped gains from exchange that are based on actual preferences that are informative of welfare. If agents have a willingness to exchange and yet this willingness to pay is based on a false belief, then an externality is also not manifest. Furthermore, if one agent is affected by an untraded activity and has informed relevant preferences, and yet has no willingness to pay for the activity, for whatever reason, then an externality is not present. The actual preference view of externalities tracks the views of individuals about whether it is appropriate to purchase certain goods or activities. On this view, then, we would have to accept that if an informed and unbiased individual is willing to pay a stranger for a smile, and if the stranger is also informed, unbiased, and willing to accept such payment, then an externality is generated, however strange this seems.

On this interpretation of externalities, Satz is incorrect to claim that all untraded spillover effects are externalities. The untraded spillover effects for which there are untapped gains from exchange that are informative of welfare are externalities, which implies that they should be mediated by markets. This means that there could be externalities for goods that we think ought not be for sale. If one individual is willing to

pay for sex, and another individual is willing to accept a payment for sex, and yet no exchange transpires, then this untraded activity generates an externality if both parties are informed, unbiased and rational; contrary to common moral beliefs, this implies that this activity ought to be mediated by markets or market-based policies. Therefore, this account of externalities draws a boundary around untraded spillover effects that tracks the views of individuals about what sorts of things for which they are willing to pay and accept a payment. But it clearly cannot accommodate views about the sorts of activities individuals should be prohibited from exchanging, as Satz highlights. For example, Satz argues that one type of "noxious" market is one that takes advantage of the vulnerabilities of individuals. People take on dangerous jobs without adequate compensation because they have no alternative.

Despite the benefits of this view, however, it is inconsistent with the way economists treat externalities in the world. Some of the most prominent (supposed) externalities, such as the externality generated by carbon dioxide emissions, are notoriously associated with false beliefs. A large portion of the US population holds the false belief that carbon dioxide emissions do not cause climate change and therefore are not harmful (Egan and Mullin 2017). This means that the minimal requirement that preferences are based on true beliefs relevant to the untraded activity in question rules out many examples of externalities, even if they are untraded and welfare-relevant activities. A fallacious belief about the causal factors relevant to climate change renders immeasurable the entire scope of the welfare gains from carbon mitigation. This is because potential gains from exchange over carbon emissions are uninformative of

welfare for those individuals who hold false beliefs about climate change. Therefore, carbon dioxide emissions do not generate a policy-relevant externality, on this view, for those individuals who hold false beliefs. Surely it is not reasonable to suppose that welfare-increasing policies should only respond to those who hold true beliefs. The welfare effects ought to accrue to everyone, yet the gains from exchange are informative only of the welfare of some individuals.

But economists claim that greenhouse gas emissions are indeed an externality which is a cause of the problem of climate change (Akerlof et al. 2019). Nordhaus (2014) states that he is estimating the magnitude of the externality generated by greenhouse gas emissions when estimating the social cost of carbon, which is the price at which the optimal carbon tax should be set. Interpreting externalities to be generated by willingness to pay that is derived from actual preferences is therefore inconsistent with the treatment of externalities in economics. Externalities are generated by untraded activities even when actual preferences are insufficiently informed.

The view that externalities are constituted by reports of willingness to pay derived from actual informed and unbiased preferences implies that externalities are observable phenomena. That is, preferences for non-market goods are observable, and therefore potential welfare gains are observable, when individuals report their preferences or reveal their preferences through their choices. This view, however, is inconsistent with the way economists treat some externalities in the world. This is because economists claim that untraded spillover effects that are associated with significant false beliefs and systematic biases, like greenhouse gas emissions, are nonetheless externalities that can be measured

for the sake of setting a welfare-enhancing policy. But the view of externalities expounded in this section holds that these sorts of untraded spillover effects are not externalities because the actual gains from exchange are unassociated with welfare gains.

3.3.2 Externalities arise from welfare gains and hypothetical preferences

The second option for interpreting externalities is that an externality arises when an untraded activity generates untapped welfare gains for which there are associated hypothetical gains from exchange. I define hypothetical gains from exchange as the gains from exchange that would arise if agents had informed and unbiased preferences. That is, hypothetical gains from exchange are determined by what individuals would be willing to pay or accept for an untraded activity if they had sufficiently informed preferences.

The benefit of retaining some version of untapped gains from exchange in the presence of uninformed preferences is that it is a monetary phenomenon, and therefore can be measured in dollars. A hypothetical willingness to pay could be posited, for example, from information about the actual willingness to pay of individuals, as well as their false beliefs. But it is not obvious whether this would provide a meaningful measure of welfare gains. Gains from exchange are meaningful insofar as they indicate the degree to which a good satisfies preferences after taking into account one's wealth and market prices. However, it is not clear that hypothetical willingness to pay can provide this sort of information. Even if preferences could be laundered such that the resulting preferences are what individuals would prefer if they had true beliefs, there would still be

an issue of determining how individuals would allocate their wealth across these new preferences to construct measures of willingness to pay. As Amartya Sen (2002) notes, it is difficult to determine the relevance of the budget constraint for willingness to pay estimates. Constructing hypothetical measures of willingness to pay from laundered preferences, however, would require knowing exactly how each individual makes allocation decisions given their constraints.

Economists are accustomed to constructing counterfactuals when estimating causal effects in econometrics. It is thus not absurd to suppose that economists might develop a way to estimate what an individual would be willing to pay in a counterfactual scenario. I do not address this observation in this dissertation; however, it is the basis of future work on the use of integrated assessment models to estimate the value of the social cost of carbon. Economists such as Nordhaus (2014) estimate the externality using information about the difference between GDP projection with and without climate change. He seems to assume that this approximates what individuals would be willing to pay if they were rational and informed. It is not obvious, however, that any individual would be willing to pay this amount even if they were rational and informed. That is, it is not clear that Nordhaus is estimating an externality at all, at least as it is characterized in economic theory. The question is therefore what Nordhaus is measuring and whether this is how externalities should be characterized in the world. Nonetheless, more attention would have to be paid to how individuals construct preferences and associated willingness to pay to develop an adequate method for constructing these counterfactuals.

As it stands, more research is needed to determine whether hypothetical willingness to pay can be estimated non-arbitrarily.

This interpretation of externalities retains the feature of untapped welfare gains, which generates the policy implications of externalities; however, the gains from exchange are posited as hypothetical entities. In this sense, an externality is not a market waiting to emerge, as Arrow (1969) suggests; instead, it is a market that would emerge in a possible world where individuals have informed, unbiased, and self-interested preferences. In the steam and particulate matter example, this means that the externality is generated by the particulate matter emissions because these emissions generate untapped welfare gains for which hypothetical willingness to pay for the reduction of particulate matter emissions could be constructed. The description of preferences in the hypothetical account of externalities is therefore idealized, and as such, does not readily transfer to the actual world. On Sugden's (2002) analysis of models in economics, then, there has not been an inference from the model world to the real world, but instead from the model world to another possible world in which all agents have informed preferences.

This interpretation of externalities counts as externalities the untraded activities that have welfare effects; this is compelling insofar as externalities are considered to be policy relevant in economics and insofar as policies are directed at welfare effects. Moreover, because hypothetical willingness to pay is a monetary construct, there is a possibility of arriving at a measurement of an externality, given a suitable theory of the relationship between welfare, purified preference, and hypothetical willingness to pay.

The implication of this view is that externalities could arise everywhere depending on how preferences are purified or laundered and how hypothetical willingness to pay is constructed. That is, a researcher could assume that a hypothetical willingness to pay could be constructed for all welfare effects. If this is the case, then externalities would be generated by any minute welfare effect. Therefore, like the broad view of externalities, this interpretation of externalities cannot distinguish between policy-relevant externalities and minute social interactions. It is not clear, however, what is meant by hypothetical willingness to pay, whether it could be meaningfully estimated, or whether it could meaningfully convey information about potential welfare gains. Hypothetical willingness to pay is far removed from the agent and their actual preferences, and therefore it is also far removed from the common assumption in economics that agents tend to know best what is good for them. The construction of hypothetical willingness to pay by researchers who, by engaging in this task, would be claiming to know better than agents what is good for the agents, is inconsistent with the liberal principles that ground revealed preference theory.

Positing hypothetical entities, however, is not uncommon in economics. A demand schedule for a normal good is hypothetical in the sense that it represents what individuals would be willing to pay for a good if they had the requisite knowledge to know what amounts they might purchase were the price to vary. That is, demand for a good is only observable at one quantity-price coordinate, which is a single point on the demand schedule. In practice, however, estimating a demand schedule is important to firms that are introducing a new product and must make production decisions, for

example. Consequently, a common concern in marketing literature is whether an individual's stated willingness to pay in a survey or experimental context is indicative of what she would be actually willing to pay when she is faced with the real option of purchasing the good. There is evidence that stated and actual willingness to pay often do not align. The explanation that is typically given for this asymmetry is that individuals do not properly consider their budget constraint in a hypothetical choice scenario.[10] The estimated demand schedule, which is constructed using stated willingness to pay, is therefore uninformative of the actual demand at various prices.

This means that gains from exchange estimated with stated willingness to pay, even if based on informed preferences, is uninformative of actual gains from exchange determined by actual willingness to pay. Individuals tend to reveal preferences in a market that are distinct from the preferences they reveal via a hypothetical market they must imagine. This is an intractable problem for externalities because a willingness to pay for an externality is always, in some sense, hypothetical. By definition, there is no market in which to reveal one's willingness to pay for an externality because it is generated by an untraded activity.

Positing a demand schedule for an externality is therefore even more problematic than positing a demand schedule for a good. Both interpretations of externalities face the problem of the divergence that arises between the hypothetical and actual willingness to pay. There is no price-quantity point on the demand schedule that is observable for any

[10] Bhatia and Fox-Rushby (2003) and Wang et al. (2007) make this point in the context of estimating a demand function for new-to-market consumer goods, whereas Arrow et al. (1993) and Christie (2006), consider this point when assessing the contingent valuation method.

untraded activity, unlike traded activities. It is therefore unclear whether researchers could accurately test the degree to which stated willingness to pay for an untraded activity is informative of their actual willingness to pay. Furthermore, externalities have the additional requirement that actual preferences correspond to welfare in the right way. When individuals hold false beliefs on the hypothetical interpretation of externalities, it is not only the context of choice but also the belief structure that inhibits the revelation of the true willingness to pay that is informative of welfare gains. When the belief structure of an individual causes a deviation in stated and true willingness to pay, it is unclear how to test whether any constructed willingness to pay captures what individuals would actually pay if they had the right beliefs and if there were a market in which they could reveal their actual willingness to pay.

Hypothetical willingness to pay in the world is therefore more idealized than the concept of a demand schedule in the world. Like a demand schedule for a market good, externalities are not actualized in the world. Unlike a demand schedule, however, externalities have no observable point of actual, as opposed to stated, willingness to pay. Furthermore, an estimated demand schedule for an externality based on stated willingness to pay is inaccurate when agents have insufficiently informed beliefs, and therefore stated willingness to pay is altered to align with what researchers think individuals would be willing to pay if they were sufficiently informed. Externalities on this view are therefore hypothetical entities that are used to posit a magnitude for a welfare effect in terms of a hypothetical construct of demand for a good.

Some of the problems for externalities in the presence of uninformed preferences might be remedied by aggregating individual willingness to pay over a large population to determine the overall value of an externality. For example, if the effect of false beliefs on preferences in a population were randomly distributed, then the distortionary effects of the false beliefs on the value of the externality might cancel out, on average. In this case, the total value of the externality could be a reliable estimate of the potential welfare gains of the untraded activity despite the presence of some uninformed preferences. Positing hypothetical willingness to pay could therefore be avoided in this case even though some actual preferences are uninformative of welfare. The problem of false beliefs on the value of an externality therefore arises in small populations, where false beliefs in a population are not random, or where false beliefs are widespread. For example, there might be no problem with reliably estimating the value of the externality generated by greenhouse gas emissions from stated willingness to pay if there were evidence that individuals in the population are as likely to under- or overestimate the effect of climate change on their welfare. However, behavioural economists have shown that there are several underlying psychological mechanisms that cause people to make systematic mistakes in some decision scenarios. For example, individuals tend to be present-biased, meaning that they typically undervalue the welfare effect of long-run decision problems such as saving for retirement (Thaler and Sunstein 2008). If most people tend to underestimate the long-run effect of climate change on their welfare, then the aggregate estimate of the value of the externality would be biased downwards.

I have therefore argued that the actual preference view of externalities is the least problematic because it is observable and relies on the judgment of individuals about what sort of preferences can be expressed in monetary terms. It limits externalities to scenarios in which there is a latent market and in which the gains from exchange are informative of welfare gains. It therefore solves the problem of distinguishing externalities from all social interactions and is consistent with the predilection among economists to take the preferences of individuals as given. But this view is inconsistent with the way they treat externalities in the world. For example, studies demonstrate that there are systematic biases and false beliefs concerning the welfare effects of greenhouse gas emissions, yet this is the most prominent example of an externality in economics (Akerlof et al. 2019).

The hypothetical preference view of externalities is the most problematic, but also the most consistent with the way economists treat externalities in the world. This view implies that any unpriced effect that has a welfare effect counts as an externality, and thus economists cannot adequately distinguish externalities from any other type of untraded spillover effect. It also requires positing hypothetical willingness to pay, and it is unclear whether this can be done in a meaningful way. It seems that the interpretation, by economists, of actual instantiations of an externality faces significant difficulties. On one interpretation, the scope of an externality is significantly narrowed to the point where externalities might rarely arise in the world. On the other interpretation, the scope of an externality depends on the researcher's view of which welfare effects should have

an associated willingness to pay and what the magnitude of the willingness to pay should be.

3.4 Conclusion

Treating externalities as simply unpriced spillover effects assumes that in the world, like in the model, gains from exchange are invariably associated with welfare gains. This rests on the view that welfare is the satisfaction of actual preferences. I argued that this account of welfare is mistaken. Therefore, there is no unqualified reason to suppose that gains from exchange are invariably associated with welfare gains in the world.

When gains from exchange and welfare gains diverge, it is unclear how to interpret externalities. If the gains from exchange must be derived from actual preferences, then externalities only arise when agents are sufficiently informed. However, it is unclear what counts as *sufficiently informed*. On this view, externalities may never arise in the world depending on what counts as sufficiently informed, and would not arise for individuals who do not rationally respond to evidence. Therefore, externalities do not proliferate in the social world on this view, but only arise when there are gains from exchange that are informative of welfare.

On the other hand, one could hold that hypothetical gains from exchange associated with welfare gains are sufficient for the actualization of externalities. However, it is unclear whether hypothetical willingness to pay is a meaningful construct. Externalities could proliferate in the social world, depending on what welfare gains the researcher assumes can be appropriately represented as hypothetical willingness to pay. Each possible interpretation of externalities is therefore problematic. Either the scope of an externality is significantly narrowed, or it depends on hypothetical constructs that are unobservable and whose measurement is problematic. If externalities in the world are understood to have these microeconomic foundations, then more work is needed to understand the formation of preferences and willingness to pay, to understand what counts as sufficiently informed preferences, and to ascertain if hypothetical willingness to pay is a meaningful concept and construct.

One possibility I did not consider in this chapter is that another way to construct hypothetical willingness to pay that is informative of welfare is for researchers to manipulate the context in which individuals form their preferences. In this way, actual preferences could become informative of welfare. In the next chapter, I show that this is done in contingent valuation studies. In so doing, I argue that the problems with interpreting externalities in the world impede the attempts of researchers to estimate the value of externalities using the contingent valuation method.

CHAPTER FOUR

The Misplaced Controversy over the Contingent Valuation Method

4.1 Introduction

In the previous chapter, I argued that there are several plausible ways to interpret externalities in the world given the way they are characterized in the canonical model. In this chapter, I aim to show that the ambiguity over the concept of an externality is not merely a theoretical concern. It has implications for how economists and philosophers assess the coherence of the contingent valuation method, which is a procedure for measuring externalities. I thus show the importance of being clear on the meaning of the concept of an externality by unpacking the criticisms developed by both economists and philosophers against the contingent valuation method and by tracing these criticisms to their mistaken or conflicting conceptions of an externality.

Contingent valuation studies use surveys to garner information about the willingness to pay of individuals for non-market goods. This method is used extensively

in government or applied economic research to value local environmental externalities such as irrigating from river ecosystems and, perhaps most famously, the damages caused by the Exxon Valdez oil spill in 1989 (Maas and Svorenčík 2017). This method, however, is dismissed by many mainstream economists (Banzhaf 2017). Jerry Hausman (2012), a prominent critic of contingent valuation, argues that the method is "hopeless" for several reasons, all which centre on the inability of surveys to elicit information about true preferences. In a revealing statement about the mainstream opinion of contingent valuation, Timothy Haab et al. (2013, 593) state that their motivation in publishing an extended response to J. Hausman's (2012) criticisms is to "urge the community of economists to recognize that the intellectual debate over contingent valuation is still ongoing." The mainstream opinion of economists therefore seems to be that the concept of an externality is meaningful, but the contingent valuation method is an unreliable method of measuring externalities.

In this chapter, I unpack the criticisms developed by economists and philosophers of the contingent valuation method and trace these criticisms to their mistaken or conflicting conceptions of an externality. In doing so, I aim to show that significant controversy over the contingent valuation method is driven by ambiguity in the concept of an externality; that is, problems that are associated with the contingent valuation method are in fact problems with the phenomenon it is designed to measure. This conclusion calls for a redirection of critiques from the contingent valuation method towards the concept of an externality.

I first explain the contingent valuation method and the prominent criticisms of the method that are made by some economists. I illustrate that the debate centres both on what is taken to be evidence that stated preferences are sufficiently informed and unbiased and on what counts as sufficiently informed and unbiased preferences. The former concerns how to test the accuracy of the measuring instrument, whereas the latter concerns how to interpret an externality in the world. The way an externality is interpreted alters the perceived success of the contingent valuation method; since contingent valuation is the prominent method to incorporate non-market benefits of environmental resources into cost-benefit analyses, its perceived success in measuring an externality is of significant practical importance.

I then explain the criticisms of the contingent valuation method made by philosophers Daniel Hausman (2012) and Mark Sagoff (2007). D. Hausman argues that contingent valuation specifically targets preferences that are unrelated to welfare. Therefore, the reports of willingness to pay elicited through this method are meaningless. I argue that this criticism is mistaken and that the intuition that some preferences cannot be accurately expressed in monetary terms is better grounded in the problematic relationship between willingness to pay and preferences, not the relationship between preferences and welfare. Sagoff adopts this focus and argues that environmental preferences are "citizen preferences," which means they cannot be expressed monetarily. I argue that this is plausible if it implies that environmental goods have intrinsic value that denotes moral standing.

Even if this is right, however, it is a mistake to claim that the contingent valuation method is useless because environmental goods have intrinsic value. The value of an externality is not a measure of the entire cost or benefit of a good, broadly interpreted. The value of an externality is an estimate of potential welfare gains, which are expressed through a willingness to pay that is properly related to welfare. Willingness to pay is properly related to welfare when it is generated from informed and unbiased preferences. It should be expected, then, that the measure of the value of an externality excludes the intrinsic value of a good, which is a value independent of the effects on human welfare. I conclude that controversy over the contingent valuation method arises from ambiguity in the concept of an externality.

4.2 The Contingent Valuation Method

The "value", or willingness to pay, an individual has for a good can be determined either by observing her market behaviour or by asking her what she is willing to pay for the good (Schelling 1968). Market behaviour, however, only reveals the *use value* an individual has for a good. That is, an individual is willing to pay for a good in a market insofar as she derives value from acquiring and using the good. Following Krutilla (1967), however, economists typically hold that the total economic value of goods is composed of use and non-use value (Fullerton and Stavins 1998). Krutilla specifies two types of non-use value. The first type is existence value, whereby an individual values the

knowledge that an environmental good exists, despite having no intention of using the environmental good. The second type is an option value whereby an individual values the possibility of keeping open the option of using the good in the future. Non-use value is not displayed in market behaviour because it is not tied to a desire to acquire or use the good.

The use value of a non-market good can be measured by observing market behaviour. The travel cost method, for example, measures how much individuals spend to use an environmental good, such as a national park. Their travel and accommodation costs reveal how much they are willing to pay to visit the park. The hedonic pricing method instead measures price differentials between goods that are equivalent except for some non-market feature. For example, the price difference between two houses that are equivalent but for their air quality reveals the willingness to pay of those individuals to live in an area with cleaner air. Hence, willingness to pay associated with use values of non-market goods can be revealed through market behaviour.

These measures, however, exclude the non-use value individuals might have for these goods. For example, some individuals might be willing to pay a certain amount to use a national park, but others value that the national park exists despite having no desire to visit the park. Similarly, some individuals might value that there are some areas of a city with cleaner air, despite not directly benefitting from the cleaner air. The non-use value portion of total economic value therefore cannot be reliably observed through market behaviour. Some individuals might donate money to environmental causes and, in doing so, reveal a non-use value they hold for an environmental good. This payment,

however, is likely to be uninformative of the willingness to pay of all individuals for a given environmental good because, among many reasons, some individuals might free-ride on the conservation efforts of others. Therefore, the total economic value of a good that has use and non-use value can only be determined by asking individuals how much they are willing to pay for the non-market good.

Contingent valuation is the only method for estimating the total economic value of non-market goods that have both use value and non-use value (Landry and List 2007). This method uses surveys to elicit individual reports of willingness to pay for non-market goods. It is often, although not exclusively, used to estimate the total economic value of environmental goods. S. V. Ciracy-Wantrub (1947) is credited as the founder of the contingent valuation method because he first proposed that welfare economics take advantage of the developments in survey research methods in social psychology and ask individuals their willingness to pay for goods when that value cannot be revealed in markets (Banzhaf 2017). Robert K. Davis (1963) is credited with conducting the first contingent valuation study as an alternative to the travel cost method (Banzhaf 2017).

In contingent valuation studies, participants are first asked warm-up questions to gauge their understanding and views of the environmental problem in question. They are then given information about the environmental problem in question. The market mechanism for eliciting reports of willingness to pay is then constructed. This includes both the payment vehicle, which is a hypothetical channel through which a payment would be made, and the elicitation question, which contains the rules for behaving in the market (Banzhaf 2017). For example, the payment vehicle could be a tax and the

elicitation question could be a simple question of the participant's willingness to pay. Finally, the participant is asked follow-up questions about their perception of the survey, whether the hypothetical scenario seemed plausible to them, their income, and their demographics (Mitchell and Carson 1989). If they did not understand the scenario or did not find the hypothetical scenario believable, then their responses are typically dropped or adjusted. The reports of willingness to pay are then aggregated by researchers to form a demand schedule, or total benefit function, for the non-market good. This function can then be used to estimate the total economic value of the non-market good. This estimate is typically used in a cost-benefit analysis of some land-use proposal, but sometimes it is also used to determine the optimal price and quantity of the environmental good itself.

Consider an example of a typical contingent valuation study. Carson, Wilks, and Imber (1994) conduct a contingent valuation survey with the goal of estimating the Australian public's value—or their total economic benefit—of adding an undeveloped parcel of land to an existing national park. They use the contingent valuation method to estimate the average willingness to pay for the preservation of this land by conducting a survey over a random sample of the Australian population. They aggregate the average stated willingness to pay for the preservation of this land to estimate the total value the Australian public ascribes to the preservation of this land. They estimate the total benefit of preserving the land to be $435 million, which they contrast to the estimated $102 million that would be generated by mining the land. The contingent valuation method thus allows for a comparison of the environmental benefits of a natural resource to the

more traditional forms of economic benefits, such as profits that are generated by using the natural resource to produce consumer goods.

Contingent valuation has been used extensively to incorporate the benefits of environmental preservation into cost-benefit analyses directed at resource-use decisions. As Carson (2012) reports, "almost 60 percent of the estimates in the very large Environmental Values Reference Inventory (EVRI) database maintained by Environment Canada in conjunction with the U.S. Environmental Protection Agency and the environmental agencies in several other countries come from contingent valuation" (28). Examples include the use of the contingent valuation method to estimate the willingness of residents to pay higher water tariffs to decrease water pollution in Fuzhou, China (Jiang, Jin and Lin 2011), the willingness of the US public to pay for climate change efforts (Aldy, Kotchen, and Leiserowitz 2012), and the willingness of Oregon residents to pay for measures that reduce fire-risk to old-growth forest in the Pacific Northwest (Loomis, Gonzalez-caban, and Gregory 1994).

The contingent valuation method is not limited to the environmental context. It has also been used to estimate the willingness to pay to develop vaccine programs in Africa (Jeuland, Lucas, Clemens, and Wittington 2009), and the willingness to pay of residents to pay higher taxes to fund a sports stadium (Johnson and Whitehead 2007). Given the prominence of the contingent valuation method in policy contexts, it is therefore important to determine whether this method can accurately measure the total value of an externality, as it is designed to do.

4.3 Long-standing Concerns with Surveys in Economics

4.3.1 Incentives to misreport willingness to pay

Mainstream economists are typically suspicious of surveys. One source of this suspicion is Paul Samuelson's (1954) theory of public goods. This theory shows that individuals have an incentive to underreport their value of a public good because public goods are nonrivalrous and nonexcludable. That is, one person's use of a public good, like clean air, does not reduce the amount of the good available for others to use, and no individuals can be excluded from using the good, respectively. Samuelson (1954, 389) argues that selfish agents will give deceptive signals about their value of a public good in order to "snatch some selfish benefit." They will free-ride on the contributions of others because they can use the good without paying the full cost of their use given that they cannot be excluded from using the good. Therefore, surveys that ask individuals how much they are willing to pay for a public good are unreliable because individuals have an incentive to misrepresent their actual willingness to pay for the good.

Contingent valuation studies, at least those that are structured according to the current best-practice guidelines of the method, attempt to control for Samuelson's concern by presenting the individuals with closed-format rather than open-format questions (Mitchell and Carson 1989). That is, it is no longer acceptable to ask an open-format survey question in which an individual is merely asked how much they are willing to pay for a good (Hanemann 1991). Asking open-ended questions is understood to

invite reports of willingness to pay that, for strategic reasons, deviate from an agent's true willingness to pay (Arrow et al. 1993). Instead, individuals are asked closed-format questions where they respond to questions that ask them to accept or reject the purchase of an environmental good at a given price. They respond with a "yes" or "no" rather than a report of their willingness to pay for the good. This is understood to eliminate the incentive to underreport willingness to pay for public goods because the good is not provided at all if they answer "no". They would be worse off if they rejected an offer at their true willingness to pay for an environmental good because they would not receive the good that they prefer receiving at that price.

An additional benefit of closed-format questions is that they are less cognitively demanding than open format questions. It is a difficult task to report on the price that would make one indifferent between purchasing and not purchasing a good. It is also an unfamiliar task, since the individuals that are typically surveyed in contingent valuation studies belong to societies where bargaining is uncommon (Haab et al. 2012). Instead, they are accustomed to taking market prices as given, and thus face dichotomous choices in markets analogous to closed-format questions in surveys. They decide whether or not to purchase a good at its market price; they do not decide exactly how much they *would* pay for a good, if they had such a choice. Therefore, surveys that aim to estimate an agent's willingness to pay should have the familiar characteristic of market choices where individuals are price-takers. Asking individuals dichotomous choice questions regarding their willingness to pay for a good is therefore considered to be an important feature of well-conducted contingent valuation studies.

The assumption that agents are price-takers, meaning that any single agent cannot influence the price of a good, is a core assumption of perfection competition in economic theory (D. Hausman 1992b, 35). Allowing for bargaining complicates models of exchange because any outcome of this type of exchange depends on bargaining power and skill. Perfect competition also requires that sellers are price-takers. Hence the dynamics of supply and demand, whereby competition leads to an elimination of excess supply or demand and hence an equilibration of supply and demand at the market-clearing price, are not explained through individuals negotiating over the price of a good. Instead, the equilibrating dynamics of supply and demand stem from individuals making dichotomous choices over whether to enter the market by accepting the going price and purchasing the good or exiting the market by rejecting the going price. The dynamics of supply and demand thus rest on the assumption that there is a critical mass of buyers and sellers such that, even if there is significant exiting from a market when prices are high, no single buyer or seller has price-setting power. It rules out the possibility, for example, that only one consumer constitutes the demand function for a good at a sufficiently high price, because in this case this individual would have some market power over the price of the good, thus violating the assumption of perfect competition.

The assumption of price-taking agents is thus a core element of perfectly competitive markets. This assumption is an idealization of the fact that consumers cannot typically single-handedly influence the price of goods in large markets. The price-taking assumption in the contingent valuation method functions to eliminate the distortions that are created when individuals hold any power during exchange rather than

to approximate the role of consumers in a real market. The contingent valuation method creates a hypothetical scenario where the respondent could plausibly be regarded as a price-maker because surveys usually ask individuals about their willingness to pay to acquire an entire environmental good. Since this good is indivisible, they are treated as if they are the single purchaser of the good. Therefore, if an individual is asked simply what they would be willing to pay for the environmental good, they are treated as if they are a price-maker. Bargaining power and skill are thus a relevant consideration in this context and individuals have an incentive to underreport their actual willingness to pay in order to secure more benefit.

The treatment of individuals as price-takers in the contingent valuation is thus not a response to how a market exchange would occur if there were an actual market for the non-market good. Instead, the assumption that the survey should create a hypothetical market in which individuals are price-takers rests on the view that true economic preferences are masked when individuals have bargaining power or when there is an opportunity to free-ride on the efforts of others.

4.3.2 Scientific realism and instrumentalism

A second source of suspicion toward surveys in economics is Milton Friedman's (1953) influential argument in favour of scientific instrumentalism. The traditional form of instrumentalism holds that theories are merely instruments, or tools, for predicting observable phenomena (Chakravarrty 2017). Therefore, instrumentalists typically deny that scientific theories give us knowledge of unobservable phenomena. Furthermore,

terms for unobservables have no literal meaning, and statements involving these terms are not candidates for truth or falsity. The most common form of scientific realism, on the contrary, is defined by a positive epistemic attitude toward the content of the best scientific theories and models (Chakravarrty, 2017).

This positive epistemic attitude involves both metaphysical, semantic, and epistemological commitments. First, scientific realism is committed to the mind-independence of the world that is studied by the sciences. Second, it is committed to a literal interpretation of scientific claims about scientific entities. This means that descriptions of both observable and unobservable entities in scientific entities have literal meaning. Instrumentalists, instead, hold that descriptions of unobservables are merely instruments for the prediction of observable phenomena. Third, scientific realism holds that theoretical claims constitute knowledge of the world. This contrasts to Bas van Fraassen's version of instrumentalism, for example, which recommends belief in theoretical claims about observables derived from our best theories, but argues that we should adopt an agnostic attitude concerning unobservables.

D. Hausman (1992b, 287) argues that the dominant instrumentalist position in economics, expounded by Milton Friedman (1953), concerns the goals of science rather than the epistemic achievements of scientific theories. Friedman argues that the goal of the science of economics is to develop tools that enable reliable predictions. This contrasts to realism about the aims of science that holds that science aims to produce true descriptions of the world (Chakravarrty 2017). That is, Friedman foregrounds the goal of accurate prediction for science, and denies the importance of discovering new

truths about the world and explaining phenomena (D. Hausman 1992b). On this view, then, there is no reason to assess whether the assumptions of a theory are "unrealistic". That is, it does not matter if the assumptions are not true of the phenomena to which the theory is applied. What matters for the assessment of a theory is its predictive success for "the class of phenomena the hypothesis is designed to explain" (Friedman 1953, 214). Therefore, the success of a theory depends solely on its ability to predict the specific phenomena it was designed to "explain" (D. Hausman 1992b, 165).

This view helps to clarify Friedman's dismissal of survey results. As D. Hausman (1992b) recounts, Richard Lester (1946) attempted to test standard assumptions of economic theory, including the assumption that firms attempt to maximize expected returns, by conducting surveys of businesspeople. He found, unsurprisingly, that firms did not behave as described by the theory of the firm. But for Friedman, "the realism of a theory's assumptions or the truth of its uninteresting or irrelevant implications is unimportant except insofar as either restricts the theory's scope. Since economists are not interested in what businesspeople say, it makes no difference what Lester's surveys show" (D. Hausman 1992b, 165). The theory of the firm is designed to explain firm behaviour, not the understanding businesspeople have of their own behaviour, or the reasons they might have for acting in a certain way.

Friedman thus argues that there is no reason to suppose that an individual's understanding of their own behaviour must correspond to the theory that explains their behaviour. He argues that a model of the distribution of leaves on a tree that treats the tree as if it can instantaneously move its leaves around the branches is a good model for

tree growth if it predicts accurately the distribution of leaf growth. But of course, the tree is not making a maximizing choice in this way, nor can it move its leaves instantaneously. Similarly, individuals need not actually understand their own behaviour or act according to the same principles that are found in economic models. A billiard ball player acts *as if* she is computing complex mathematical formulas, and thus a good model of her behaviour would treat her as such (Friedman 1953, 223). But she is not actually acting in this way and indeed her actions might be governed by random chance. A survey of her understanding of her own behaviour therefore should not be expected to correspond with the way the model treats her behaviour. Individuals behave *as if* they are maximizing their utility, which has an associated willingness to pay that is properly connected to their welfare. Firms act *as if* they equate marginal costs to marginal revenues.

On Friedman's view, then, the model of externalities is successful if it accurately predicts the outcome of an untraded spillover effect, which is that the untraded effect is overproduced. There is little reason to ask people what they are willing to pay for a good because individuals should not be expected to report on their preferences in a way that corresponds to the economic model of their behaviour. Furthermore, an individual's reflection on their economic behaviour or preferences is not the phenomena economic models were designed to explain. Therefore, predictive failure of the model when applied to understanding this phenomenon—for example, when applied to understanding the willingness to pay reported by individuals—is irrelevant to judgments of the success of the model.

Economists who estimate willingness to pay in order to estimate the total value of an environmental good think that individuals can report on their economic preferences and that this practice is meaningful, if not to economic theorizing, then at least to setting economically-informed policy. Economists who conduct contingent valuation surveys cannot understand economic models to be merely tools that enable prediction of a narrow range of phenomena. They spend considerable time debating whether the assumptions of the model hold in the world, and how better to align survey findings with the structure of the model. That is, they attempt to create a hypothetical context in which the stated preferences of individuals cohere with the assumptions in the economic model of externalities. Furthermore, coherence with the assumptions of the model is taken as evidence that the contingent valuation survey has been conducted well. Friedman's hugely influential instrumentalist position thus bolsters suspicion about the coherence of the contingent valuation method and the relevance survey results to economic theorizing.

4.4 Criticisms of the Contingent Valuation Method within Economics

The prominent criticisms of the contingent valuation method by economists were developed partly in response to the prominence of this method in the environmental policy context in the late twentieth century in the United States. In 1980, the Department of the Interior (DOI) decided to use consumer surplus as the framework for assessing

damages when suing parties for hazardous releases. Consumer surplus is the difference between a consumer's willingness to pay and the price they have to pay to obtain a good. In the case of a non-market good or activity, consumer surplus associated with that good or activity is the total value of the externality as I characterized it in the previous chapter. In 1981, Ronald Reagan expanded the role of cost-benefit analysis in public policy decisions, especially in decisions made by the Environmental Protection Agency (Banzhaf 2017). This meant that more environmental policy was made on the basis of cost-benefit analysis, which required an estimate of costs, or damages, measured by consumer surplus. In 1989, the DOI incorporated non-use values into their definition of damages (Banzhaf 2017). That is, the willingness to pay that enters into the estimation of consumer surplus was required to include a measure of willingness to pay for the existence or option value of the environmental good. Since contingent valuation is the only method to set a monetary value on non-use values, it is unsurprising that it rose to prominence in the environmental policy context.

Significant controversy over the validity of the contingent valuation method, however, emerged after it was used to estimate the damages caused by the Exxon Valdez oil spill in 1989. This oil spill occurred in Prince William Sound, which is a remote area of Alaska; fittingly, it was determined that losses to non-use value constituted a significant portion of the total damages caused by the spilled crude oil. Contingent valuation was therefore used to estimate a monetary value for these non-use value damages (Carson et al. 1992). The U.S. National Oceanic and Atmospheric Administration (NOAA) assembled a "blue ribbon panel" composed of prominent

economists such as Kenneth Arrow to assess contingent valuation method (Arrow et al. 1993). The panel concluded that contingent valuation was reliable in valuing damages that include non-use values when certain conditions are met, including that respondents are giving extensive information about the proposed policy, that they are reminded of their budget constraint and the available substitutes, and that in-person interviews are conducted (Arrow et al. 1993).

Exxon's response to the use of contingent valuation to assess the damages of the oil spill was unusual in that it did not contest the dollar value determined through contingent valuation. Instead, Exxon sponsored research in opposition to the contingent valuation method itself. This research was conducted by authoritative economists who were nonetheless inexperienced with contingent valuation studies (Maas and Svorenčík 2017). These Exxon-funded economists published a compendium of critiques of the contingent valuation method (J. Hausman 1993). The typical structure of these critiques was to conduct a contingent valuation study and to argue that the findings of these studies establish that contingent valuation cannot produce reliable estimates of value (Hanemann 1994). Banzhaf (2017, 228) argues that there was "a reasonable cross-section of views" about the validity of CV that "mirror[ed] differences in the larger profession" prior to Exxon's attack on the contingent valuation method. After this debate, contingent valuation largely lost its credibility in mainstream economics, although not within the policy domain (Banzhaf 2017). Harro Maas and Andrej Svorenčík (2017, 340) similarly argue that "by sponsoring countervailing studies, Exxon thus clearly changed the playing field of CV research."

This episode in the history of the contingent valuation method is important because it spawned several significant and long-standing criticisms of the method. Jerry Hausman, who was a prominent Exxon-sponsored economist that presented evidence for the invalidity of contingent valuation in the early 1990s, argues that contingent valuation has not progressed in the years following this debate (J. Hausman 1993; 2012). He thinks the method has not resolved fundamental problems that were identified during the debate on the estimates of the Exxon Valdez oil spill damages. Indeed, he thinks that these problems are irresolvable. In the following three sub-sections, I explain these purported problems and their respective responses. For each, I argue that the criticisms and responses stem from a difference in views, both about what evidence is used to claim that contingent valuation is an accurate measuring instrument and about what counts as an accurate measurement of an externality.

4.4.1 Hypothetical response bias

The first of the three fundamental problems with the contingent valuation method, according to Jerry Hausman (2012), is that individuals have a hypothetical response bias. In a contingent valuation study, individuals are not offered the actual opportunity to purchase a good. Therefore, they are reporting on what they think they would be willing to pay for a good if they had the option of purchasing the good. That is, they are reporting on their intention to pay rather than demonstrating their willingness to pay through an action. J. Hausman (2012, 44) takes this to be an intractable problem with

contingent valuation because "what people say is different from what they do," and he assumes that true preferences are revealed only by what people do.

There is evidence that individuals tend to overstate their willingness to pay in a hypothetical setting (Carson and Groves 2007). Marketing researchers, who attempt to estimate demand for new-to-market goods, account for this by dampening the estimated demand from the hypothetical study. Haab et al. (2013) argue that contingent valuation studies can employ the same sort of tactics. Furthermore, they argue that there is a new line of research that investigates the role that incentives play in reducing the divergence between hypothetical and actual willingness to pay. Therefore, they think that a divergence is not evidence that contingent valuation is inherently flawed, but instead that there is still much to learn about how to create the right incentives that will align hypothetical and actual willingness to pay.

Both sides of the debate therefore agree that actual willingness to pay, which is revealed through a purchase, is the accurate measure of an externality; therefore, they agree that hypothetical bias must be controlled. They disagree, however, on whether a contingent valuation study can ever be conducted in such a way that individuals report a willingness to pay that is convergent with what their actual willingness to pay would be if there were a market in which to purchase the non-market good. An estimate of stated willingness to pay in a contingent valuation study, unlike a new-to-market consumer good study, is intractably hypothetical because the market to which the respondents are valuing is also hypothetical. Individuals will likely never have an option to purchase the environmental good in question. Therefore, the debate is over whether preferences can

be sufficiently unbiased in a contingent valuation context, and the proposed solution rests on designing the survey setting to eliminate the bias.

The explanation economists offer for the presence of hypothetical bias is that individuals do not properly consider their budget constraint in a hypothetical setting (Arrow et al. 1993). When they are faced with the real opportunity to purchase a good, they realize they are more income-constrained than they initially accounted for, and therefore they are willing to pay less for the good than they stated in the hypothetical setting. But why think that a budget-constrained willingness to pay is more meaningful than a budget-unconstrained willingness to pay? The typical economic response is that actual willingness to pay is the invariable phenomenon that contingent valuation aims to measure because actual willingness to pay is derived from true preferences. But this does not seem right. If the reason hypothetical and actual willingness to pay diverge is that individuals do not properly consider their budget constraints in a hypothetical setting, then the difference between the two is a function of how stringent the budget constraint is when an agent assesses their willingness to pay.

Willingness to pay is meaningful insofar as it is evidence for welfare gains. It indicates the degree to which a good satisfies the preferences of an agent relative to the other goods that they can purchase, and the satisfaction of preferences is evidence for welfare gains. The budget constraint, however, is a problematic aspect of willingness to pay when willingness to pay is taken to inform of welfare gains. It introduces the marginal value of income into this measure of preference satisfaction. Because of this, two identical reports of willingness to pay from two individuals give no information

about the relative sizes of welfare gains. One individual may reap large welfare gains from a one unit increase in an environmental good, but have a low willingness to pay for the good because they have very little income. The other individual may gain little welfare from the same one unit increase in an environmental good, but have a large willingness to pay because they have a substantial amount of income to allocate across various goods. Since they are wealthier, they have a lower marginal value of income, which means they are willing to pay more for goods from which they derive lower welfare gains. The phenomenon of interest in a contingent valuation study is welfare gains, and welfare gains are masked by the effect of the budget constraint in determining the willingness to pay of individuals.

It is therefore not obvious why economists typically hold that actual, rather than stated, willingness to pay is the target phenomenon in the context of contingent valuation. If stated willingness to pay is less budget constrained than actual willingness to pay, as Arrow et al. (1993) argue, then this might be a desirable feature of willingness to pay when we are interested in garnering information about welfare gains in a policy context.

There are two lessons we can draw from the debate over the presence of hypothetical bias in contingent valuation studies. According to J. Hausman (2012), willingness to pay elicited in a hypothetical setting such as a contingent valuation study is intractably inaccurate. According to Haab et al (2013), there is evidence that the effect of hypothetical bias on stated willingness to pay can be controlled by using particular survey designs or statistical methods. Both sides of the debate therefore agree that the effect of

hypothetical bias on stated willingness to pay is problematic; they differ in terms of whether it must be eliminated (which is impossible in a contingent valuation setting) and whether it must be controlled. For J. Hausman, an individual can never be sufficiently unbiased to report meaningfully on their willingness to pay when they are in a hypothetical context. For Haab et al., it is possible that the contingent valuation study can construct a context with the right incentives for an individual to report an unbiased willingness to pay.

The debate therefore partly concerns what is taken as evidence for the success of contingent valuation in eliciting willingness to pay that approximates actual willingness to pay. But the debate also concerns the standards placed on preferences. J. Hausman rejects the contingent valuation method because it elicits stated preferences, which he argues are never indicative of true preferences; no matter how one alters the survey design or employs statistical methods to correct the bias, the contingent valuation method can never elicit information about true preferences because it can never elicit actual willingness to pay. For J. Hausman, then, an externality is measured by what an individual is willing to pay when they are free from hypothetical bias. This implies that externalities are measured by what individuals are actually willing to pay for a non-market good, not what they actually say they are willing to pay. Haab et al. instead understand an externality to be constituted of what individuals say they are willing to pay when they are sufficiently unbiased. *Sufficiently unbiased* to them means that hypothetical bias has been reasonably controlled for by the survey design or by employing particular statistical methods. Differences in the interpretation of the type of willingness to pay that

constitutes an externality—and therefore differences in the way an externality is interpreted—thus accounts for some of the disagreement among economists about the viability of measuring externalities generated by nonmarket goods.

4.4.2 Willingness to pay and willingness to accept

The second problem J. Hausman (2012) raises is that individuals report a willingness to pay for non-market goods that diverges from their willingness to accept a payment for the same good in contingent valuation studies. That is, their monetary value for the same good differs depending on whether they suppose that they are purchasing the good from another agent or whether they suppose that they own the good and are being paid to give it up. According to economic theory, one agent's willingness to pay to purchase some good should differ from their willingness to accept a payment for that good only by an income effect. That is, willingness to pay assumes the agent does not own the good and shows how much they will pay to acquire it. Willingness to accept assumes the agent owns the good and shows how much they are willing to accept to give up the good. In theory, the dollar value an agent sets on the good should not be affected by whether they own it or not. Willingness to pay therefore should differ from willingness to accept only insofar as accepting a payment of the good changes their real income relative to the scenario in which they are purchasing the good.

Contingent valuation studies find, however, that the two measures diverge by a degree that cannot be explained by an income effect alone. Several studies have found that the minimum payment individuals are willing to accept for someone else to acquire a

good they own is three times higher than what they are willing to pay to acquire the same good (Landry and List 2007). J. Hausman (2012) argues that this shows that individuals are not drawing on stable preferences when they respond to a contingent valuation survey. That is, they are not sufficiently informed in the contingent valuation scenario to reveal their true preferences.

Behavioural economists instead interpret this deviation between willingness to pay and accept to be caused by a psychological mechanism, called the endowment effect, whereby an individual values goods more highly when they own the goods (Kahneman et al. 1991). This effect has been observed in laboratory settings as well as in contingent valuation settings for several types of goods. This indicates that this is not a problem unique to contingent valuation or non-market goods in general. Instead, behavioural economists interpret the endowment effect to reveal a problem with the economic notion of value because the value individuals have for goods appears to be unstable and context-dependent.

Contingent valuation researchers instead argue that individuals are not accustomed to selling goods. This lack of experience causes them to report a willingness to accept that does not reflect their true value of the good. On the contrary, individuals are familiar with purchasing goods in markets where they take prices as given. As such, they have experience with assessing their willingness to pay for a good in a way that reflects their true value for the good. Willingness to pay is therefore a more accurate measure of value than willingness to accept. Therefore, these economists retain the mainstream economic notion of value, unlike behavioural economists, and use

willingness to pay to measure value because it is arguably the more accurate measure of value.

The debate among the economists who retain the mainstream economic notion of value therefore concerns whether individuals are sufficiently informed to reveal their value by reporting on their willingness to pay or willingness to accept. Proponents of contingent valuation think that individuals are sufficiently informed when they report their willingness to pay because they have experience with forming willingness to pay for market goods. Critics of contingent valuation instead think that individuals are never sufficiently informed in a contingent valuation setting to reveal their true preferences either through reporting on their willingness to pay or willingness to accept. The debate thus concerns whether the divergence between willingness to pay and willingness to accept is evidence that individuals are insufficiently informed to report an accurate monetary representation of their value for a good, or whether the divergence can be explained by factors other than a lack of information.

4.4.3 The scope problem

The third and final major problem for contingent valuation, according to J. Hausman (2012), is that stated preferences do not respond appropriately to the scope of non-market goods. That is, an individual's willingness to pay for an environmental good seems to change very little as the quantity of the good increases. For example, one study found that the willingness to pay to clean one lake is roughly equal to the willingness to pay to clean five lakes (Diamond and Hausman 1994). The problem is that it is

reasonable to suppose that individuals should be willing to pay more if they are paying to clean more lakes. One explanation for the phenomenon, offered by behavioural economists, is that individuals are really paying for a *warm glow* feeling when they contribute to causes (Thaler and Sunstein 2008). They pay to feel good about contributing to a good cause, and this feeling is not sensitive to how much of the good cause their payment brings about. Therefore, the extent of the environmental preservation generated by their payment is inconsequential to them; what matters is that they feel good about doing any quantity of good.

J. Hausman (2012, 47) argues instead that the scope problem "demonstrates the nonexistence of preferences in a contingent valuation setting." It shows that stated willingness to pay in a contingent valuation setting is arbitrary because it indicates that individuals construct their preferences during the survey. J. Hausman thus assumes that individuals will report irrational preferences in the absence of fully formed preferences; if they had fully formed preferences, those preferences would respond rationally to the scope of the good in question. He argues that contingent valuation studies are valid only if they do not suffer from the scope effect, and he argues that the test for this is the Hausman-Diamond Adding Up Test. To conduct this test, the researchers must conduct one survey to estimate the willingness to pay for the good A given that a good B is provided, a second survey to estimate the willingness to pay for B given A, and finally another survey to estimate the willingness to pay for both A and B. The sum of the values found in the first and second surveys should equal the value found in the third survey. J. Hausman argues that contingent valuation studies do not pass this test—either

because they fail the test or have not been subjected to the test—and therefore the contingent valuation method has not been proven to successfully elicit preferences.

This test involves strong assumptions about how individuals should value non-market goods (Hanemann 1994; Haab et al. 2013). J. Hausman assumes that an individual's value of goods A and B must be linearly related—the sum of the values for A and B individually must be equal to the value of A and B together. But economists ascribe to the view that most goods, called normal goods, have a diminishing marginal value. For example, the first unit of water a person receives is highly valuable, and as such they are willing to pay a high price for it. Once they have an abundance of water, however, they are willing to pay very little for one extra unit of water. Similarly, if good A is one national park and good B is another national park, and an individual sees the two goods as substitutes even though they are, strictly speaking, distinct goods, then it is reasonable to suppose that an individual would be willing to pay more for the first national park than for they are for the second national park, given that the first is already established (Rollins and Lyke 1998). But if the only option is to pay for either both parks or no parks at all, then parks A and B are treated as a single good. There is little reason to suppose that willingness to pay for the two parks together should be equal to the sum of the willingness to pay for park A and park B.

Nonetheless, the heart of the problem for J. Hausman is that he thinks, as in the other two problems, that the scope problem is evidence that individuals are not reporting on their true preferences in a contingent valuation setting. If they were better informed about the situation, and perhaps more rational or thoughtful about their responses, then

they would report willingness to pay that cohered more closely with the theoretical understanding of preferences. Stated willingness to pay in this context therefore is not an accurate estimate of preference satisfaction or welfare for J. Hausman; therefore, the externality could be measured by what individuals would be willing to pay if they were sufficiently informed, not what they state they are willing to pay. Proponents of contingent valuation like W. Michael Hanemann (1994) and Timothy Haab et al. (2013) instead take these reports of willingness to pay to be meaningful and therefore to constitute the value of an externality. They do not take the scope problem to show that individuals are insufficiently informed or unable to report on their true preferences; instead, they seek to understand why their preferences have an unexpected structure. The economists thus differ in terms of whether they take the scope problem to be evidence that individuals are insufficiently informed to report on their true preferences that are properly related to welfare.

4.5 Lessons

Notice a few important features of the contingent valuation method that help to reveal differing interpretations of externalities among economists. The externality is understood to be measured by the sum of the willingness to pay reports. The demand function constructed from these reports is used to calculate the total value of the externality. But proponents of contingent valuation do not merely take any report of

willingness to pay to be meaningful. Willingness to pay reports are understood to be inaccurate when a survey is conducted poorly, meaning that it does not control for known biases, or it does not inform agents of the problem, or it does not construct a plausible hypothetical market. Reports of willingness to pay are accurate when the survey context does not introduce biases and when the individual's report of willingness to pay is based on the best information about the good in question. This implies that these practitioners do not take the preferences an individual has at any given moment to necessarily determine the value of an externality. Instead, it is only when preferences are informed and unbiased that they determine the value of an externality. Proponents of contingent valuation think that individuals can be sufficiently informed and unbiased to report on their preferences, given a proper survey design. Therefore, they think that actual reports of willingness to pay constitute an externality as long as the contingent valuation study constructs the right context for informed and unbiased willingness to pay to be revealed.

J. Hausman (2012) similarly thinks that reports of willingness to pay are accurate only when respondents are sufficiently informed and unbiased. In contrast, however, he seems to require preferences to be more informed and unbiased than the proponents of contingent valuation require, and therefore for preferences to cohere more closely with the economic understanding of preferences. He takes certain characteristics of the willingness to pay that is elicited through contingent valuation studies to show that the method does not create a context in which individuals are sufficiently informed and unbiased. Indeed, he thinks that reports of willingness to pay elicited in a contingent

valuation context cannot escape hypothetical bias because contingent valuation constructs a hypothetical market in which individuals report hypothetical willingness to pay. He also appears to think the lack of information about and experience with non-market goods prevent the elicitation of true preferences in a survey setting. Harro Maas and Andrej Svorenčík (2017) report that Kenneth Arrow recorded an episode at the hearing on the Exxon Valdez damages estimates in which J. Hausman argued that people are not knowledgeable enough to form accurate preferences in complex policy scenarios. J. Hausman stated, at the hearing, that the damages should not be assessed via contingent valuation, but instead should be left to "people like you [referring to Arrow] or me or other people who are engineers" and not "what I consider to be almost an uninformed opinion by public opinion polls" (Maas and Svorenčík 2017, 334-335).

If J. Hausman thinks that the insufficiently informed preferences present the biggest issue for contingent valuation, rather than the hypothetical survey setting, then other methods for measuring the value of non-market goods are also problematic. One of these methods is hedonic valuation. Individuals indirectly reveal their actual willingness to pay for non-market goods, like clean air, through a market choice. Their willingness to pay for a house in a cleaner neighbourhood as opposed to a polluted neighbourhood can therefore be used to back-out their value for clean air. Individuals lack information that is normally conveyed by market goods, however, for all non-market goods regardless of whether the willingness to pay for this non-market good is revealed through purchases of market goods or in a survey setting. Therefore, if J. Hausman thinks that experience in a market is necessary to be sufficiently informed

about a good, then it is impossible to accurately measure all non-market goods. The implication of this view is that externalities cannot be accurately measured from individual reports of willingness to pay, which in turn implies, on the gains view of externalities, that externalities never arise.

If J. Hausman thinks that hypothetical bias is intractable, on the other hand, then the problem with the contingent valuation method is with the type of value targeted by this method. Non-use value cannot be revealed in market behaviour; information about non-use value therefore must be garnered through surveys. Surveys create a hypothetical market setting in which hypothetical, rather than actual, willingness to pay is revealed. But hypothetical measures of willingness to pay are inaccurate, according to J. Hausman. Therefore, non-use value cannot be ascribed a monetary value, at least one that is determined by willingness to pay. This problem is therefore unique to the contingent valuation method that attempts to estimate the monetary value of non-use values by eliciting hypothetical willingness to pay. This implies that externalities never arise in relation to non-use values, but they might arise for unpriced activities that have a use value to agents, and therefore for which a willingness to pay can be revealed in market behaviour.

The debate over the contingent valuation method partly concerns what counts as evidence that preferences are informed and unbiased. For example, J. Hausman (2012), but not Haab et al. (2013), thinks that the divergence of willingness to pay from willingness to accept is evidence that individuals are reporting on uninformed or biased preferences. But the debate over contingent valuation method also partly stems from

different views on how informed or unbiased individuals must be for their reports of willingness to pay to be taken as evidence of their true preferences.

J. Hausman holds that the willingness to pay that constitutes externalities must be derived from highly informed and highly unbiased preferences, which he infers is impossible for non-market goods with non-use value because of the presence of hypothetical bias. He places high standards on the level of information and lack of bias individuals must have to report accurately on their preferences, and he thinks a survey setting precludes the realization of these standards. Markets hold a privileged position for J. Hausman because they create a context in which sufficient information is conveyed to individuals and biases are minimized.

Haab et al. have less stringent requirements on preferences. The value of an externality is estimated by measuring stated willingness to pay of individuals, which conveys what individuals believe they would be willing to pay for a good if the contingent market constructed in the survey came to fruition. A hypothetical setting does not prevent individuals from reporting informed and unbiased willingness to pay. If this interpretation is correct, then J. Hausman's position is that an externality is generated by an untraded activity for which there is informed and unbiased willingness to pay, and this willingness to pay is what an individual is actually willing to pay for a good. This indicates that non-use value is not a constituent of an externality, or at least that it cannot be accurately measured and included in an estimate of the value of an externality. Haab et al. instead hold that an externality is generated by an untraded activity for which there is sufficiently informed and unbiased willingness to pay, and this willingness to pay is what

an individual states they would be willing to pay if they had the option of purchasing a good.

In this section, I therefore tried to illustrate the differences in how economists interpret externalities. The assessment of the contingent valuation method concerns how informed and unbiased economists think individuals must be to accurately report their willingness to pay, and therefore which reports of willingness to pay can be used to estimate the value of the externality.

They disagree on whether an externality is comprised of actual or hypothetical willingness to pay for a non-market good, and they disagree on whether these are sufficiently similar. Ambiguity over the interpretation of externalities therefore is relevant to how economists measure externalities for policy purposes and whether they think it is possible to measure externalities that involve non-use value.

In the following section, however, I argue that the most problematic feature of contingent valuation is absent in the debate among economists. This is the problem of taking all reports of willingness to pay, even if they are derived from informed and unbiased preferences, to be meaningful measures of welfare in the context of environmental problems.

4.6 Philosophical Criticisms of the Contingent Valuation Method

4.6.1 Daniel Hausman on contingent valuation

Daniel Hausman (2012) argues that the contingent valuation method is misguided because the category of preferences it targets are not properly linked to welfare. Willingness to pay is a measure of preference satisfaction, and preference satisfaction is meaningful only insofar as it is evidence for welfare gains. D. Hausman (2012) argues that preference satisfaction is evidence for welfare gains only when preferences are informed, unbiased, and self-interested. Therefore, willingness to pay is also evidence for welfare gains when it is derived from these "laundered" preferences. As we have seen, proponents and critics of the contingent valuation method agree that willingness to pay is accurately measured when there is reason to believe that it is derived from preferences that are sufficiently informed and unbiased, but disagree on what counts as *sufficiently* informed and unbiased. D. Hausman argues, additionally, that preferences are informative of welfare when they are organized according to what an agent thinks will make them better off. If preferences are directed at making oneself better off, then there is reason to believe that satisfying these preferences will increase welfare. When an agent is informed, unbiased, and forms self-interested preferences, we can presume that their assessment of what will make their lives go better is accurate, and therefore that satisfying these preferences is likely to lead to an increase in their welfare. Therefore, willingness to pay is meaningful insofar as it is derived from this sort of preference.

Individuals can organize their preferences according to considerations other than self-interest. Amartya Sen (1977) famously argues that we have several preference orderings that align with different values or outlooks we might adopt. A consideration of commitment to principles, for example, results in a distinct preference ordering to a preference ordering that is organized by considerations of self-interest. D. Hausman argues that when people act according to preferences that are not self-interested, we cannot take their preferences to be evidence of what will make them better off. This is because their preferences are formed on the basis of benefitting others, for example, rather than benefitting themselves. Satisfying this type of preference therefore might be evidence of a welfare gain for someone other than the individual that holds the preference. Therefore, the stated willingness to pay of an individual for a good that is derived from non-self-interested preferences is not evidence for what will make that individual better off.

D. Hausman argues that contingent valuation attempts to elicit willingness to pay that is derived from preferences that are not self-interested. That is, he argues that individuals primarily value the environment according to considerations other than their personal advantage, such as a concern for justice. Individuals desire environmental preservation not because it will make their lives go better, but because they think they have a duty to be "good stewards of the earth" (D. Hausman 2012, 91). This means that the willingness to pay that is elicited for environmental goods is likely derived from preferences that are not self-interested; more likely, these preferences are formed on the basis of their commitments. D. Hausman argues that the satisfaction of this type of

preference is not evidence for welfare gains and therefore the associated willingness to pay is also not indicative of welfare gains. But willingness to pay is meaningful only insofar as it indicates what will make an individual better off. D. Hausman thus concludes that the willingness to pay for an environmental good, which is elicited using a contingent valuation study, is meaningless.

D. Hausman (2012) argues that preference satisfaction theories of welfare, which hold that welfare is identical to the satisfaction of informed, unbiased, and self-directed preferences, is problematic in part because of its focus on *self-directed* preferences. He argues that it is unclear how to distinguish self-directed preferences from other-directed preferences and, furthermore, it seems like some other-directed preferences are relevant to welfare gains. For example, Jack's preference that his child live a long life appears to be, strictly speaking, other-directed because the preference concerns his child and not himself. Yet, the satisfaction of this preference bears on Jack's well-being. If this preference were to be frustrated, Jack's life would go much worse than if the preference were satisfied. This other-directed preference therefore bears on Jack's welfare. D. Hausman's evidential view of welfare holds instead that self-interested, rather than self-directed, preferences are evidence for welfare gains. Self-interested preferences are preferences that are directed toward increasing one's own welfare, and may be self- or other-directed. Whether Jack's preference concerns his child or himself is not consequential in assessing the relation of this preference to his welfare. If the preference is self-interested, or formed according to what Jack believes will make him better off, then this preference is evidence for what will make him better off. Therefore, Hausman

circumscribes the self-interested, informed, and unbiased preferences as the category of preference that provides evidence for welfare gains. The non-self-interested preferences, even if they are informed and unbiased, do not provide evidence for welfare gains or convey meaningful information for the purposes of economic policy assessment.

It is not convincing, however, that the motivation one has for holding a preference determines whether the satisfaction of that preference bears on one's welfare. Jack's preference that his son live a long life is not likely to be motivated by self-interested considerations. Jack is not motivated by considerations of what will make *his* life go better when he forms the preference that his son live a long life. Instead, he is motivated by what is welfare-improving for his child. But satisfying this preference does, in fact, make Jack's life go better. Furthermore, Jack likely knows that the satisfaction of this preference will make his life go better even though this concern does not motivate the formation of this preference. What matters for using preference satisfaction as evidence for welfare gains, however, is whether the satisfaction of a preference is associated with an increase in welfare. The satisfaction of Jack's preference that his child lives a long life is likely evidence for gains to his welfare because parents are often better off when their children are alive and healthy, even though this preference is not motivated by his assessment of what will make his life go better. If this is correct, then Jack's preference is not self-interested, in the sense that it is not formed on the basis of what Jack thinks will make his life go better, even though it is evidence for what will make his life go better.

The same thing could be said of environmental preferences. D. Hausman claims that preferences that are motivated by, or formed on the basis of, moral principles or benevolence are not good evidence for what will make one's life go better. But this is not necessarily the case. A person might form a preference for environmental preservation based on a motivation to uphold certain moral principles. Nonetheless, satisfying this preference might make the life of this person go better. And there is good reason to think that environmental conservation does make the lives of individuals go better. Having access to nature can improve health and mental well-being, and this can hold even when these considerations do not factor into the formation of preferences for environmental conservation. A preference for environmental conservation therefore might not be organized by a motivation to increase welfare, but the satisfaction of this preference might still increase welfare.

Part of the issue for specifying the nature of the evidential relationship between preference satisfaction and welfare stems from equivocation over the term *self-interest*. David Hume (1738) distinguishes between self-interest associated with the untrammeled pursuit of one's own advantage, such as lying for personal gain, and enlightened self-interested, which often requires resisting immediate gratification in order to secure long-run benefit, such as telling the truth in order to benefit in the future from being believable. D. Hausman identifies self-interest with a desire to make one's life go better, but it is not clear what counts as making one's life go better. If the untrammeled pursuit of one's personal advantage is what constitutes self-interest, then it is plausible that environmental preferences are un-self-interested even if they increase welfare in the long

run; one could secure more immediate personal gain by purchasing consumption goods rather than investing in long-run projects even if these projects increase their long-run welfare. But if self-interest is more akin to enlightened self-interest, in that it concerns making one's life go well in the long run, then it seems that a preference for environmental conservation might be self-interested and contribute to making one's life go better. For example, I might consider that investing in environmental conservation now reduces my consumption of goods that provide me with immediate personal benefit, but it will make my life better twenty years from now because the severity of weather events will be reduced. In one sense, then, my preference for conservation is not self-interested because it does not cohere with a goal of securing personal advantage, and in another sense, it is self-interested because it increases my well-being in the long-run.

D. Hausman (2012) thinks that self-interested preferences are evidence for welfare and that non-self-interested preferences are not evidence for welfare. I have tried to show that this distinction does not circumscribe the preferences that are evidence for welfare. This is because preferences that are not formed on the basis of self-interested, however construed, are sometimes evidence for welfare gains. As such, it is not clear that environmental preferences are irrelevant to welfare claims, as D. Hausman suggests. Therefore, D. Hausman's argument that contingent valuation produces meaningless results does not withstand scrutiny. It is plausible that some environmental preferences, even if motivated by moral considerations, are relevant to welfare gains.

D. Hausman focuses on the relationship between preferences and welfare in his critique of the contingent valuation method. The more problematic relationship,

however, is between willingness to pay and preference satisfaction. It might be the case that an environmental preference is evidence for welfare even though it is based on principles or commitments, but this preference might not be expressible in terms of willingness to pay. For example, Jack has a preference that his child live a long life, and the satisfaction of this preference is evidence for his welfare. Nonetheless, he cannot express this preference in terms of a willingness to pay for his child to live a long life. Similarly, satisfying a preference for environmental conservation might be good evidence for welfare gains, but this preference might also resist a reduction to willingness to pay.

4.6.2 Mark Sagoff on citizen values

Mark Sagoff (2007) denotes the type of preference that cannot be expressed in monetary terms *citizen preferences*. These preferences are directed toward what we take to be right and just. Consumer preferences instead are directed at what will make us personally better off. Sagoff argues that everyone has both a role as a consumer and a citizen, and our preferences correspond to the role we adopt in different contexts. In our roles as consumers, we have self-interested preferences for goods that concern what we want for ourselves. In our roles as citizens, we are directed toward justice and distribution rather than what is good for ourselves as individuals. Sagoff argues that environmental problems are failures to uphold social goals that "represent not goods we choose but values we recognize" (2007, 27). That is, environmental preferences concern our citizen values, not consumer values, and environmental decision-making is a political activity that is assessed according moral principles rather than a judgment of welfare.

Therefore, environmental regulations that control pollution, for example, express society's views about the rights of people and property. Disagreements over these regulations should be resolved via debates over their moral qualities and objective merits, not by weighing their costs and benefits.

Sagoff argues that representing any value of the environment in terms of willingness to pay distorts the way individuals value the environment. This is because only consumer values can be represented in terms of willingness to pay because they represent self-interest desires about what we want for ourselves. Since environmental values are citizen values that concern what we want for society, any willingness to pay that is elicited for environmental goods does not express information about individual welfare, and consequently, it is not meaningful. When willingness to pay is derived from spiritual or political values, for example, it is uninformative of individual welfare. Therefore, Sagoff argues that there is no way to make sense of willingness to pay when the willingness to pay is stated for goods that are valued as citizen goods.

Like D. Hausman, Sagoff therefore distinguishes between self-interested and un-self-interested preferences to argue that willingness to pay for environment entities is a meaningless measure. But instead of focusing on how these two types of preferences bear on welfare gains, he focuses on whether they can be expressed in monetary terms. Sagoff thinks that preferences directed at one's own welfare can be expressed in monetary terms because we are accustomed to satisfying this type of preference in consumer good markets. The observation that consumer preferences are often expressed in monetary terms in markets, however, does not get to the heart of why these

preferences, and not citizen preferences, can be meaningfully expressed in terms of willingness to pay.

Many preferences that seem to be partly based on citizen values are often expressed in terms of a willingness to pay in consumer markets. A preference for fair trade over conventional coffee is plausibly partly based on moral considerations and therefore seems to be, in apart, a citizen preference. Yet, this preference can be expressed in terms of a willingness to pay, and as such is an example of a citizen value that is expressed in markets. There are also markets for charities that express willingness to pay to help save the lives of people and protect endangered species, for example. There are markets for works of art that are culturally invaluable; these pieces are bought and sold in markets even though the buyers and sellers, and many other individuals external to the market exchange, value these pieces beyond their contribution to their welfare. Therefore, consumer values and citizen values are not as distinct as Sagoff supposes. Why is it not possible for individuals to accurately express their preferences for things like environmental conservation in monetary terms but they can express their preference for goods like fair trade coffee and works of art in monetary terms? A foray into the distinction between intrinsic and instrumental value is helpful to understand why Sagoff thinks that willingness to pay derived from environmental preferences is meaningless, and therefore why the project of the contingent valuation method is misguided.

4.6.3 Intrinsic and instrumental value

A prominent position among environmental ethicists is that environmental goods cannot be valued monetarily. This position typically rests on the distinction between instrumental and intrinsic value. Something is instrumentally valuable if it is valued as a means of bringing about some further end, whereas something is intrinsically valuable if it is valuable as an end in itself. For example, a carpenter has instrumental value to a customer insofar as they are valued as a means to renovate one's house. But that carpenter, as a human being, is also intrinsically valuable. Entities that are a part of the natural environment have instrumental value insofar as they contribute to the well-being of humans by improving air quality, filtering ground water, and providing materials for the development of new medicines, for example. A central aim of environmental ethics is to argue that entities that are a part of the natural environment also have value beyond their instrumental value (O'Neill 1992; Jamieson 2001). That is, entities in the natural environment also have intrinsic value.

The core argument against using an economic framework to determine environmental outcomes—and therefore conceiving of environmental problems primarily as externality problems— is that monetary valuation is only meaningful for instrumentally valuable goods. If it can be established that entities in the natural environment indeed have intrinsic value, then it can also be argued that a monetary valuation for these intrinsically valuable environmental goods is not meaningful.

The notion of intrinsic value is notoriously vague in the environmental ethics literature, however, and this weakens the claim that intrinsically valuable things cannot be valued monetarily. There are some goods that seem to have intrinsic value that are

nonetheless regularly traded in markets, such charitable donations and culturally significant pieces of art. There are also some activities or objects that are intrinsically valuable in a way that can be expressed in monetary terms. For example, a person might value hiking a mountain as an end in itself, rather than a means to a further end. But they might also have a price they are not willing to pay to hike that mountain. They would not hike the mountain if they had to pay fifty dollars to do so, but they are willing to pay fifteen dollars to hike that mountain. This expresses how much preference satisfaction they gain from hiking that mountain, even though they value hiking the mountain as an end in itself. It is uncontroversial, however, that some things that are intrinsically valuable cannot be expressed in monetary terms. Individual human beings, for example, are intrinsically valuable in a way that cannot be expressed in monetary terms.

Part of the confusion is a result of equivocation over the term *intrinsic value* (O'Neill 2001). On some interpretations, intrinsic value is non-instrumental value. For example, there are activities, objects, and states of affairs, that are ends in themselves for individuals, rather than a means to some further end. The value one holds for hiking a mountain as an end in itself is thus a non-instrumental value. On other interpretations, however, intrinsic value refers to the Kantian principle of treating individuals as ends in themselves. To claim that something is an end in itself, in this way, is to claim that it has moral standing. The first sense of intrinsic value need not involve any ascription of moral standing to these things; however, John O'Neill (2001, 165) argues that "there is a plausible claim to be made about the relation between them — that if y is of value to x, and x has ethical standing, then there is a prima facie ethical duty for ethical agents not to

deprive x of y." Therefore, the non-instrumental and moral interpretations of intrinsic value are related but distinct interpretations of intrinsic value.

Intrinsic value is used in both ways by environmental ethicists to develop a substantive position according to which some non-human objects have intrinsic value. The project aims to establish that entities in the natural environment are intrinsically valuable in the sense that they have moral standing independent of their usefulness to humans. That is, in environmental ethics literature, intrinsic value typically denotes both non-instrumental value and moral standing. Expressing non-instrumental value in monetary terms is not necessarily problematic, as the value for hiking a mountain suggests. The type of intrinsic value that is problematic for monetary valuation is the type that denotes moral standing. Willingness to pay for a thing cannot adequately capture the intrinsic value for a thing or person insofar as this value expresses moral standing. To minimize confusion, I will call the non-instrumental form of intrinsic value *non-instrumental value* and reserve *intrinsic value* for the type of intrinsic value that connotes moral standing.

It is not difficult to interpret within an economic framework why willingness to pay cannot be meaningfully expressed for things that have intrinsic value. Things that have intrinsic value insofar as they have moral standing do not have substitutes (Venn and Quiggin 2007). You can substitute the services of one contractor for the services of a different contractor insofar as you instrumentally value the services provided by this person. But you cannot substitute the contractor, valued as an individual that has moral standing, for the other contractor. The concept of willingness to pay rests on the

assumption that an individual is indifferent between acquiring a good and retaining a certain amount of money. If the price for a good exceeds their willingness to pay, for example, then they purchase another good instead. If the price of coffee gets too high, for example, then they might purchase tea instead. If something does not have a substitute, then it cannot be expressed in terms of willingness to pay. An expression of willingness to pay for a good without substitutes cannot capture information about welfare because an individual is willing to pay any price for a given quantity of this good; demand is highly inelastic for goods with few substitutes. Willingness to pay in this setting captures wealth constraints rather than welfare gains. Consider a parent who is purchasing life-saving medication for their child and is willing to pay whatever they can afford to acquire this medication. This willingness to pay capture the limits of their wealth, not the extent to which the medication satisfies their preferences and contributes to their welfare. An individual might be only willing to pay one-hundred dollars to conserve an entity in the natural environment, but if they are willing to pay this because they believe it is intrinsically valuable, then this willingness to pay might capture how much wealth they have to spend rather than how much this good contributes to their preference satisfaction. Willingness to pay is informative of wealth constraints, not preference satisfaction, when the willingness to pay is a report on something that is intrinsically value.

Environmental economists do not typically distinguish between intrinsic and instrumental value; instead, they draw a distinction between use and non-use value. Non-use value presents methodological challenges to economists because it is not revealed in

market behaviour. The typical methodology of estimating the value individuals have for goods according to their revealed preferences therefore does not work for values that are not related to the use individuals make of goods. This non-use value can take the form of valuing that something exists, without having any intention or desire of acquiring or using that thing. It cannot be revealed apart from a directed conversation that questions the person about their values.

It might be supposed that use value is a synonym for instrumental value and non-use value is a synonym for non-instrumental value, if not intrinsic value. If something is instrumental in achieving a further end, then it is used to achieve this end. For example, a person might value a soccer ball instrumentally, insofar as they can use it. If something is valued despite not being used or usable, then it is plausible that it is non-instrumentally valuable. A soccer ball that was used in a famous game and which is now kept in a museum is valued not for its usefulness; instead it is valued in and of itself. Therefore, it has a non-use value that seems to be non-instrumental value.

Use value is not identical to instrumental value, however, because it is possible to non-instrumentally value the use of something. For example, the act of playing guitar involves using a guitar, and using a guitar by playing it might be non-instrumentally valuable to a musician; playing the guitar is not valuable as a means to some further end, but rather as an end in itself. Therefore, a musician's value for the act of playing a guitar can be described both as a use value and a non-instrumental value. Use value does not imply instrumental value because something can have a use value that is associated with a non-instrumental value. Similarly, non-instrumental value does not imply non-use value

because the use of some things can be non-instrumentally valuable. One cannot have a non-use value for something that is merely instrumentally valuable, however. This is because something is instrumentally valuable if it is valuable only as a means to some further end; this implies that it is valued insofar as it is useful in achieving this end. Therefore, instrumental value implies use value.

Non-use value does not imply non-instrumental value, however. Krutilla (1967) argues that option values are a type of non-use values. An option value is the value one has of keeping open the option of using something in the future, despite not valuing its current use. For example, someone might value the continued existence of the Grand Canyon just to keep open the option of visiting it in the future. This is not a non-instrumental value because they value the Grand Canyon only as a means to potential benefits they might derive from using it in the future. This value therefore corresponds to an instrumental value one might hold in the future rather than a non-instrumental value for the Grand Canyon as an end in itself.

Krutilla stipulates that existence value is another type of non-use value. Existence value is the value one has for the knowledge that something continues to exist. This is a value for *knowing* that something exists; it is not necessarily a value for the existence of the thing in itself. An individual can have an existence value for an area of the natural environment, for example, without non-instrumentally valuing for this area. They might instrumentally value the area of the natural environment insofar as it is the object of their knowledge, and they value this knowledge in and of itself. Although this might not be a compelling interpretation of existence value, Krutilla nonetheless does not commit to the

natural environment having any intrinsic or non-instrumental value even though he acknowledges that individuals value knowing that things in the natural environment exist.

It is plausible, however, that non-use value is often associated with non-instrumental value. If someone values the continued existence of the Grand Canyon, and thus has a non-use value for it, then it is plausible that they value the Grand Canyon either as an end in itself or insofar as they believe it has moral standing. Environmental ethicists hold that individuals do not value the mere knowledge that entities in the natural environment exist; they value the existence of entities in the natural environment because these things are either non-instrumentally or intrinsically valuable.

Economists and philosophers therefore employ distinct categories of value. Economists distinguish between use and non-use value because they are concerned with whether values are revealed in markets. Something that has use value, regardless of whether this is associated with instrumental or intrinsic value, can be purchased in a market. Someone might purchase a rare painting solely because they want to ensure its preservation, and thus value it non-instrumentally. Nonetheless, they reveal a willingness to pay in a market for this good. Non-use value is problematic because it is not necessarily expressed in markets. Someone might have an existence value for the same painting, and this value is not revealed in a market because someone else purchasing and preserving the painting upholds this non-use value. Revealed preference thus fails to capture this type of value; therefore, individuals must be surveyed to reveal their non-use value.

The contingent valuation method target non-use values, and some non-use values could be interpreted as intrinsic values that denote moral standing. This type of non-use value cannot be ascribed a meaningful monetary value. As I suggested, this can be understood in an economic framework according to the lack of substitute for things that have moral standing because the concept of willingness to pay assumes that there are available substitutes. But this does not mean that *any* report of willingness to pay for an environmental good is meaningless, as D. Hausman and Sagoff suggest. On the contrary, this means that insofar as some portion of environmental entities have intrinsic value, a report of willingness to pay cannot express the entire value individuals hold for these environmental entities and might instead express their wealth constraints. If a measure of willingness to pay is a report on the intrinsic value of the environmental entity, then it is not a meaningful measure of preference satisfaction or welfare and therefore cannot be incorporated into a measure of the value of an externality. If willingness to pay is instead an expression of instrumental value, however, then it is a meaningful measure of preference satisfaction and welfare as long as preferences are properly related to welfare.

If this is correct, then the problem with the contingent valuation method is not that it attempts to estimate a monetary value for environmental entities that expresses individual values for these entities. It is plausible that environmental goods have benefits to human welfare and willingness to pay can capture information about welfare gains. Therefore, the contingent valuation method can capture these benefits by garnering information about willingness to pay for environmental goods. Individuals can be ascribed a monetary value determined by their wage that captures some information

about the instrumental value of their labour, or the benefits their labour brings to the welfare of others. But this monetary value does not capture their value as a human being. Similarly, environmental entities can be ascribed a monetary value that captures the benefits they bring to human welfare, but their intrinsic value cannot be captured by this measure. The contingent valuation method is problematic if it conflates willingness to pay that is evidence for welfare gains with willingness to pay that reflects the intrinsic value an individual has for the environmental entity. The value of an externality is a measure of welfare gains derived by an unpriced good or activity; if a willingness to pay that is divorced from welfare gains is reported by individuals and incorporated by researchers into the measure of the externality, then the contingent valuation method will overstate the value of the externality.

The contingent valuation method is directed at estimating the value of externalities, and the value of externalities are determined by the benefits an unpriced activity generates for human welfare, such as the health benefits derived from cleaner air. Therefore, willingness to pay derived from preferences that are evidence for preference satisfaction constitute the value of an externality, and contingent valuation elicits these values through properly constructed surveys. The contingent valuation method therefore accurately measures an externality insofar as the elicited willingness to pay is properly related to welfare and preference satisfaction. The value of an externality is not the entire benefit of an unpriced activity, and similarly contingent valuation does not elicit willingness to pay for all benefits of the activity. The value of an externality excludes the intrinsic value of a good, and similarly, contingent valuation should do the same.

Determining which type of willingness to pay is reported by individuals during a contingent valuation study, and ensuring that only the meaningful type of willingness to pay is incorporated into an estimate of the benefits of an environmental good, is an empirical problem that must be addressed by theorists who use the contingent valuation method to estimate the value of an environmental externality.

4.7 Conclusion

Economists and philosophers tend to regard the contingent valuation method as a problematic approach to environmental problems. Economists typically mistrust results derived from surveys in part because they think individuals are intractably biased in such a setting. Philosophers have argued the contingent valuation method is incoherent because it claims to set a monetary value on things that cannot be valued monetarily. I have tried to show that these criticisms often concern the interpretation of externalities, rather than the success of the contingent valuation method at measuring externalities.

I argued that economists disagree on what counts as a sufficiently informed and unbiased preference, and therefore they disagree on how to interpret externalities in the world. I also exposed the weakness of the philosophical criticisms of contingent valuation. These criticisms single out the incoherence of attempting to set a monetary value on intrinsically valuable things. But these criticisms are misdirected because contingent valuation is designed to measure externalities, which are a measure of welfare

gains. The conflicting or confused interpretations of an externality therefore drives these criticisms of the contingent valuation method.

If an externality is understood as an untraded spillover effect, then there is a temptation to describe any problem related to a non-market good as an externality problem and to use contingent valuation to estimate its value. As the gains view of externalities illustrates, this is a mistake. Not all untraded spillover effects are externalities because not all untraded spillover effects are relevant to welfare or accurately expressed in terms of willingness to pay. It is therefore important to be clear on the meaning of the concept of an externality to clarify the appropriate context in which the contingent valuation method can be appropriately applied.

CHAPTER FIVE

Regretful Decisions and Climate Change[11]

5.1 Introduction

Experts advise immediate and widespread action to limit climate change, yet we continue to do little to significantly reduce greenhouse gas (GHG) emissions (Intergovernmental Panel on Climate Change [IPCC] 2013). As I have argued in the previous chapters, economists attribute this problem to the presence of externalities. Their recommendation to combat climate change is to institute a carbon tax to shift the costs associated with pollution onto those who create the pollution, thus creating an incentive to reduce pollution by aligning private interests with social interests (Hahn and Stavins 1992; Stavins 2011; Akerlof et al. 2019).

In this chapter, I argue that climate change is not merely an externality problem, regardless of how externalities are interpreted. Even if private and social interests

[11] This chapter is originally published in a slightly different form as Livernois, Rebecca. 2018. "Regretful Decisions and Climate Change" *Philosophy of the Social Sciences* 48, 2: 168-191. Copyright © 2017 (Rebecca Livernois) https://doi.org/10.1177/0048393117741335.

regarding climate change were perfectly aligned, meaning that all externalities were eliminated, we may still end up in a regretful state despite acting rationally. This is because climate change involves intertemporal choices that are uncertain and characterized by marginal costs that are uninformative of the total costs of polluting. The aim of maximizing social well-being under these conditions results in an incentive to over-pollute.

My argument rests on an analogy, first proposed by Chrisoula Andreou (2006), between environmental problems and Warren Quinn's (1990) puzzle of the self-torturer (PST). I argue that the PST, as I interpret it, is a helpful model for understanding one of the problems we face in climate change decisions. To make this case, I first show why Andreou's interpretation of the PST and its relation to environmental problems is flawed. Quinn and Andreou argue that the PST shows that intransitive preferences are sometimes rational. Andreou takes the analogy between the PST and environmental problems to show that intransitive preferences are rational in environmental problems, in particular.[12] Thus, she argues that a unified collective—what I instead call a "social planner," following the tradition in welfare economics—can follow a path of environmental destruction by making choices based on informed, rational, yet intransitive preferences.[13]

[12] Andreou describes the self-torturer's preferences as "reasonable" or "understandable." I interpret these terms to be interchangeable with "rational" when it is understood broadly, as I define below.
[13] A social planner is an agent, imagined in welfare economics, who acts in accordance with the interests of society. This agent only cares about the well-being of all agents and makes all choices in the economy. The choices of the social planner thus demonstrate the outcome that would result if all agents acted in accordance with the social, rather than their private, interests.

In rational choice theory, an agent is rational if and only if her preferences are transitive and complete (Grüne-Yanoff 2012). Clearly, intransitive preferences in the PST cannot be rational in this sense. In this paper, I take a preference relation to be rational if preferences align with the overarching goal of making oneself better off, according to a subjective evaluation of well-being (Sen 1977; D. Hausman 2012). This is to say that a preference relation is irrational if an agent knowingly prefers the option that makes her worse off according to her own evaluation of her well-being. Transitivity is often understood to be a precursor even to this broader notion of rationality. One reason for this is that agents can be made to act as money pumps when they have intransitive preferences, whereby a series of trades leaves them with the original outcome but less wealth (Davidson, McKinsey, and Suppes 1955). Andreou's claim that environmental preferences are rational and intransitive is therefore significant because it denies this understanding of rationality. Indeed, she claims that the self-torturer can invariably prefer the option that makes her better off, thus acting rationally, despite being made into a money pump.

But Andreou's claim is more significant than merely asserting the possibility of rational intransitive preferences in a tightly constrained scenario. She argues that environmental preferences in particular can be rationally intransitive. If correct, this finding is potentially devastating to economic models that attempt to provide solutions to environmental problems like climate change. This is because transitivity is a foundational assumption in economic theory. For example, consumer choice theory envisions a consumer who maximizes her utility subject to her budget constraint. Utility

is merely a mathematical function that represents ordered preferences that are necessarily transitive. Therefore, a utility function that includes environmental preferences could not be constructed if environmental preferences were indeed characteristically intransitive. This means that economic theory, such as the theory that underpins the use of carbon taxes to solve the cost-distribution problem inherent to climate change, could not be accurately applied to environmental problems. Moreover, it would be a mistake to assume environmental preferences were transitive in order to make the environmental problem amenable to an economic solution. If the intransitivity of environmental preferences causes the environmental problem we wish to solve with economic theory, then assuming transitivity in order to apply economic theory would assume away the crux of the problem.

In this paper, I argue that the PST does not show that intransitive preferences are rational. I argue instead that the (properly specified) PST is an informative model of the problems a rational agent with transitive preferences faces when marginal costs are uninformative of total costs in intertemporal and uncertain choices. I first recount the PST and explain Andreou's argument that environmental problems have a comparable structure and therefore comparable result to the puzzle. I then argue that Andreou is right to claim that environmental problems have a similar structure, but reject her claim that intransitive preferences are rational in the PST. I argue that the self-torturer only appears to have intransitive preferences when the PST is underspecified and the self-torturer's evaluation of her well-being is inconsistent.

We must assume the self-torturer possesses certainty to disentangle the effects of intransitivity and uncertainty on choices. When the self-torturer possesses certainty, she possesses full information and perfect foresight in the context of the experiment. This means that she knows *ex ante* everything she would know *ex post* for every possible choice. We must also explicitly consider the way in which the self-torturer evaluates her well-being. I argue that the changes in the self-torturer's costs and benefits—also called the marginal costs and benefits—are uninformative of her total level of well-being, which is determined by total costs and total benefits. Therefore, she ought to consider total costs, not marginal costs, when forming rational preferences over outcomes. I argue that when we assume the self-torturer possesses certainty and that she takes only total measures of her costs and benefits to be informative of her well-being, it becomes clear that it is not rational for her to exhibit intransitive preferences.

I argue that climate change decisions are made under conditions analogous to the PST under uncertainty. In this context, I show how a social planner who has transitive preferences and who aims to maximize net benefit (total benefits net of total costs) faces incentives that will likely lead to a regrettable outcome. This is significant because it avoids the troubling implications of the existence of rational intransitive environmental preference while explaining how rational environmental decisions can lead to a suboptimal outcome even when private and social interests are aligned. This result suggests that when an intertemporal decision problem, like climate change, involves negligible marginal costs and uncertainty, adopting a policy of maximizing social well-

being may lead to a worse outcome than a policy that primarily aims to avoid reaching a regretful state.[14]

5.2 The Puzzle of the Self-Torturer

Consider Warren Quinn's (1990) PST. Suppose a patient, called the self-torturer, agrees to permanently wear a portable electric shocking device. This device has a dial of ordered electricity-level settings from 0 (off) to 1001. Increased settings correspond with increased pain—at level 0 there is no pain while at 1001 the pain is excruciating. For each single turn of the dial, the electricity level increases so slightly that the self-torturer cannot discriminate between adjacent settings. However, she experiences different pain levels at sufficiently distant settings. Before the experiment begins, the self-torturer has a trial week to test and compare different settings. At the end of the week, the dial is returned to 0. From then on, she has two options each week:

A. Stay at the current setting and receive no reward. B. Increase to the next setting and receive $10,000.

If the self-torturer chooses option (A), the experiment ends and she indefinitely remains at the current setting. If the self-torturer chooses (A) in the first week, the shocking device stays turned off at setting 0 and she does not receive a monetary reward. If the self-torturer chooses option (B), then the setting increases by one level, she

[14] For more on the question of adopting a precautionary or a cost-benefit approach to climate change, see Broome (2012), Gardiner (2006), and Steel (2015).

receives a monetary reward, and she is presented with options (A) and (B) the following week. The outcomes of the options are the same each week; she never has the option of returning to a lower pain level, nor does she have the option of reducing her pain level in any way, such as by taking painkillers. She is aware of all these conditions, and the following assumptions are made about her preferences:

I. The self-torturer prefers more money to less money.

II. The self-torturer prefers less pain to more pain.

III. There is some setting s_m such that for any setting s_n where $n \geq m$, the self-torturer prefers setting 0 with no reward to remaining at sn with the associated reward.

IV. No other preferences are relevant for the self-torturer's choices.[15]

The puzzle arises from considering what would be reasonable for the self-torturer to do under these conditions. Given preference (I) and (II), the self-torturer should invariably choose option (B) each week. This is because the only relevant preference each time she chooses between the two options is her preference for more money, given that adjacent settings feel the same. If she chooses (B) each week then she will end up at the last setting where she is in excruciating pain. If her preferences were transitive, then she would prefer setting 1001 to setting 0 because she prefers increasing the setting each week ($s_{j+1} > s_j$). However, understandably, she does not prefer setting 1001 to setting 0 because she prefers no money to being tortured, in line with condition (III). This is taken to show that the self-torturer has intransitive preferences: once she reaches s_n, she has

[15] This format of the PST, slightly different to the original, is taken from Tenenbaum and Raffman (2012).

reason to prefer setting 0. Yet, if she is given the option of returning to setting 0 and resuming the experiment, she also has reason to continue choosing option (B), leading her back to s_n; that is, her preferences are cyclical ($0 < 1 < \ldots < s_{m-1} < s_m < s_n < 0$).

Therefore, the self-torturer is in a tricky situation because there seems to be no optimal stopping point. The self-torturer seems right to think that if she is going to quit, she is always better off quitting at the next setting rather than the current setting. Yet she is also right to think that if she always acts in this way she will end up in excruciating pain. The self-torturer's problem is that any stopping point is suboptimal; even when she is in terrible pain, it is still rational for her to increase settings given that adjacent settings feel the same and her only options are to stay at the current setting or increase by one setting.

Proponents of the PST claim that the result of the puzzle is that the self-torturer's preferences over outcomes, which are formed from a preference for more money and a preference for less pain, are rational yet intransitive. The PST therefore presents a challenge to the principle that transitivity is a prerequisite to rationality. The implication is that, in certain contexts, it is rational for an agent to have preferences that can be exploited as a money pump.

5.3 The Puzzle of Air Pollution

Andreou constructs a simplified case of an environmental problem that has the same structure as the PST. In this puzzle, Andreou rules out the possibility of interpersonal conflicts of interest by assuming the decision maker is a social planner.[16] A social planner is an agent who acts in accordance with the interests of society. Each month, it decides whether to continue consuming pollution-intensive luxury goods. The pollution is known to be carcinogenic but its effect on health each month is negligible. The social planner knows that if it does not reduce pollution, then eventually the health of the community will be irreparably damaged (Andreou 2006, 104). It has the following options each month:

a. Stop consuming luxury goods and thereby reduce pollution.
b. Continue consuming luxury goods and thereby continue polluting.

It has the following preferences:

i. The social planner prefers more consumption of luxuries to less consumption of luxuries.
ii. The social planner prefers less pollution (better health) to more pollution (worse health).
iii. There is some level of health s_m such that for any level of health s_n where $n \geq m$, the social planner prefers setting 0 with no reward to remaining at s_n with the associated reward.
iv. No other preferences are relevant for the social planner's choices.

This scenario is structurally identical to the PST where pain is replaced with poor health, monetary reward is replaced with consuming luxury goods, and increasing voltage

[16] The term Andreou uses for this decision-maker is the "unified collective."

settings is replaced with monthly pollution. The same tricky situation arises for the social planner: since one extra month of pollution makes no noticeable difference to the community's health, the only preference that is relevant each month is the preference for consuming more luxury goods. Therefore, it is always better for the social planner to choose to keep consuming luxuries for one more month. However, if the social planner reasons like this every month, then the community will reach a state where health is so poor that it would prefer to give up all the rewards to return to a better state of health. Hence, the social planner's preferences are intransitive.

I interpret Andreou's argument as follows. If key PST conditions arise in a decision scenario, then intransitive preferences are rational in that scenario; these key PST conditions arise in some environmental problems such as air pollution; therefore, intransitive preferences are rational in some environmental problems such as air pollution. I also interpret Andreou to take individually negligible effects (what I call negligible marginal costs) to be the key PST condition. She states, "where individually negligible effects prompt intransitive preferences, destructive conduct can prevail even if the (individual or collective) agent is guided by a single (or shared) set of stable and informed preferences" (Andreou 2006, 104-105). The PST is supposed to show that intransitive preferences are rational when we face a sequence of choices where the marginal cost of each decision is individually negligible but jointly significant while the marginal benefits are large.

It is reasonable to believe that this asymmetry in marginal effects arises in climate change decisions because the effects of GHG emissions are noticeable only at large time

intervals, such as every decade (Brown 2014, 131). Furthermore, the marginal costs in the puzzle need not be imperceptible to produce the result of rational intransitive preferences (Tenenbaum and Raffman 2012, 11). Even if marginal costs are perceptible yet minuscule in comparison to the marginal benefits, it is still in the self-torturer's best interest to continue to the next setting each time she contemplates stopping the experiment. So, even if it is not believable that the marginal costs of climate change are imperceptible each time we are faced with a climate change decision, it is believable that the perceived short-term costs associated with an extra period of polluting are dwarfed by the benefits we receive from polluting activities.[17] The second premise of Andreou's argument is therefore correct when we assume that the key PST condition is the existence of negligible marginal costs and large marginal benefits. In the next section, however, I argue that when the PST is properly specified, it becomes clear that the first premise is false: intransitive preferences are irrational even when key PST conditions arise.

5.4 The Revised Puzzle of the Self-Torturer

The PST is underspecified in two respects. In section 3.1, I argue that the PST must assume the self-torturer possesses certainty before any conclusions can be drawn

[17] Even if we are not entirely clear how to measure the marginal costs of pollution given the underdeveloped theory of values associated with climate change, it is at least one clear possibility that the marginal costs are minuscule (Gardiner 2011, 224).

about the structure of the self-torturer's preferences. In section 3.2, I argue that the way the self-torturer is presumed to assess her well-being in the PST is misleading. I argue that when the PST is properly specified by resolving these issues, it becomes clear that intransitive preferences are not rational in the PST. I suggest that underspecifying the puzzle in precisely these two ways gives the mere appearance of rational intransitive preferences.

5.4.1 Certainty

Intransitivity can only be exhibited unambiguously when the decision maker acts under certainty. In the PST, we "observe" what we take to be rational choices. Choices are assumed to be revealing of unobservable preferences because it is assumed that agents choose their most preferred option. However, uncertainty disrupts the link between preferences and choices. When an agent is uncertain, her choices are unlikely to be informed by her preferences alone. Instead, her choices are based on her preferences in addition to her beliefs about future outcomes (D. Hausman 2012, 33). As such, observed choices under uncertainty are not informative of preference structure, but rather the combination of preference structure and beliefs about the future. As such, we must assume the self-torturer possesses certainty in the PST in order to draw any conclusions about her preference structure.

Furthermore, it is particularly important to eliminate the confounding effect of uncertainty in the PST because both uncertainty and intransitivity can lead to the regretful state we observe in the PST. In this state, the decision maker regrets finding

herself in a situation where she is worse off than she would have been had she acted differently in the past. The PST is supposed to be a case of regret that results from intransitive preferences in which the self-torturer prefers to return to a lower setting once she rationally reaches an excruciatingly painful one. A decision maker with transitive preferences can also end up in a regretful state because of her ignorance about the future. She preferred every step she took to reach her final state, yet, upon reaching that state, realizes that she is now worse off than she would have been if she had acted on better information.

Although both uncertainty and intransitivity can result in a regretful outcome, a key feature that distinguishes the two causes of regret is how the agent would act given the opportunity to try the experiment again. The agent with intransitive preferences would not act differently on her second attempt of the experiment. When the self-torturer possesses certainty, meaning that she possesses full information and perfect foresight, her stable and informed intransitive preferences cause her to reach the regretful state. Increasing information through learning therefore would not change her choices. As such, upon returning to a lower setting from an excruciating one, she would prefer to increase settings until she is back at the excruciating setting, *ad infinitum*, if she had intransitive preferences. Thus, she would exhibit a stable preference loop. The agent that has transitive preferences and acts under uncertainty, however, would act differently if she were given the chance to try the experiment again. Importantly, she would learn how to avoid a regretful state. She would not (rationally) act as she did in the past

because she would know that this would make her worse off than if she acted differently. This agent learns, through experience, to make choices that improve her well-being.

To determine whether the hypothetical behavior of a decision maker is the result of intransitive preferences, we must therefore eliminate the confounding effect of uncertainty on her behavior by assuming she possesses certainty. If the self-torturer has intransitive preferences, then we should observe a stable preference loop when she is given a second try at the experiment. The original PST assumes the self-torturer has a trial week to test different settings, which gives her some information prior to the start of the experiment. She is nonetheless acting under uncertainty if she only has this limited time to learn about the settings. When the self-torturer possesses certainty, it is as if she had an infinite number of tries at the experiment and as such possesses complete knowledge of all the possible outcomes in the puzzle. So, I add to the PST that the self-torturer has all the knowledge she may require during the experiment as well as perfect foresight.

5.4.2 Marginal and total considerations

The second respect in which the PST is underspecified is in its understanding of how the self-torturer evaluates her well-being. The PST stipulates that the self-torturer has only two preferences: one for more money, and one for less pain. It does not explicitly stipulate, however, how she weighs these competing preferences to form preferences over outcomes that involve both money and pain. What the self-torturer ultimately cares about, however, is making herself better off. Her well-being is

determined by how much wealth she possesses and how much pain she feels. Therefore, evaluating her well-being requires evaluating how pain and wealth combine to determine her overall well-being. Indeed, the main claim of the PST is that intransitive preferences are rational, where rationality is defined in terms of well-being. Therefore, to assess if preferences are rational, we must determine how the self-torturer should evaluate the contributions of wealth and pain to her well-being.

It is implied in the PST that a valid assessment of well-being is one that is grounded in marginal considerations alone. Proponents of the PST claim that it is rational for the self-torturer to invariably prefer higher settings to lower settings because increasing settings is invariably associated with a greater change in wealth than the change in pain. The change in pain from one setting to the next is the marginal cost of increasing settings, while the change in wealth from one setting to the next is the marginal benefit of the action. So, the claim is that it is rational to choose an action when the marginal benefits are greater than the marginal costs of that action because it indicates that the action creates a positive change in well-being. But marginal cost and benefits are relevant to well-being only by way of their relationship to total costs and total benefits. The self-torturer's current pain level (total costs) and her total acquired wealth (total benefits) determine her well-being. Well-being increases when the total benefits increase more than the increase in total costs. This is (usually) the case when marginal benefits are greater than marginal costs. Since marginal costs and marginal benefits are (usually) just the change in total costs and total benefits given a change in settings, when marginal benefits are greater than marginal costs, net benefits (total

benefits net of total costs) increase by moving to the next setting. So, given that every turn of the dial is associated with greater marginal benefits than marginal costs, it is claimed by proponents of the PST that every turn of the dial increases well-being. As such, it is concluded that it is rational for the self-torturer to invariably prefer higher settings.

If marginal effects were the only relevant measure of well-being to the self-torturer, however, then she would invariably prefer higher settings and never prefer lower settings; thus, no intransitivity would be revealed. But proponents of the PST also claim that at some point, the self-torturer reaches a point of such excruciating pain that she rationally prefers to be at some lower setting and return the associated reward. It is rational because her well-being could be increased by decreasing her pain level even if this means reducing her wealth. Notice that this evaluation of well-being is no longer grounded in marginal terms. The self-torturer instead assesses her well-being in totality rather than merely assessing the changes in her costs and benefits. In doing so, she considers her total benefits in comparison to her total costs at her current setting and decides that reducing her total costs and total benefits would increase her well-being. At the point of excruciating pain where the self-torturer exhibits intransitive preferences, the self-torturer realizes that her decisions based on marginal considerations alone led her to a state that she now regrets. Her net benefit is less than it was at a previous setting.

Considering marginal effects alone for the entire experiment thus leads to the conclusion that increasing one more setting invariably improves well-being even at an excruciating level of pain. Considering total effects alone for the entire experiment leads

to the conclusion that, after a certain point in the experiment, decreasing several settings improves well-being. Restricting the self-torturer to consider either marginal or total effects therefore leads to conflicting claims about which actions make the self-torturer better off. If we assume the self-torturer only assesses marginal costs and marginal benefits until she reaches a high level of pain, at which point we assume she assesses only her total costs and total benefits, then we will get the result that she prefers to reach a state that she then regrets reaching. So, the rational self-torturer's apparent intransitivity is driven by shifting from marginal to total considerations of the costs and benefits of her actions. The regretful point in the experiment, which is supposed to be the point at which intransitive preferences are revealed, is located by shifting from a marginal to a total evaluation of well-being.

The standard interpretation of the PST therefore assumes the self-torturer evaluates her well-being in total terms only once she reaches a regretful state. But why should we assume that the self-torturer ignores the total costs and benefits in favor of marginal costs and benefits for most of the experiment only to switch her focus once she has already reached a regretful state? Considering marginal costs alone is justified when they are a reliable indicator of total costs, which is the case when the marginal costs are the changes in total costs. Furthermore, considering marginal costs and benefits in isolation usually provides information about the relationship between total costs and benefits. The point at which total benefits are maximized net of total costs is likely difficult to discern by looking at total costs and benefits alone. It would require a trial and error process by assessing how the size of net benefits compares at various settings

to determine where a maximum is reached. Marginal costs and benefits provide information about where this maximum is reached when they have their usual relationship to total costs and benefits since net benefit is maximized where marginal benefits are equal to marginal costs. In decision scenarios where marginal costs and total costs have their normal relationship, it is therefore consistent for agents to shift from an evaluation of well-being in marginal to total terms because they are essentially evaluating total costs in both cases, merely in different terms.

However, marginal costs are not derivable from, and therefore are not informative of, total costs in the PST. If they were, then total benefits would be invariably greater than total costs because marginal benefits are invariably greater than marginal costs. This is clearly not the case because net benefit does not invariably increase over the course of the experiment. In the PST, we therefore have two distinct measures of costs. The marginal costs that are relevant to the self-torturer are not the changes in total costs. Since marginal costs and total costs are distinct measures of costs, it is not warranted to claim that the self-torturer considers only marginal costs for some decisions and only total costs for other settings. A rational decision maker, who aims to maximize well-being, ought to make choices based on the measure of well-being that is the most informative of her well-being. Well-being is maximized when the total benefits are maximized net of total costs, which does not occur where marginal costs equal marginal benefits in the PST. Since marginal costs are uninformative of the total costs in the PST, and the total costs and total benefits determine well-being, the rational self-

torturer should take only the total costs and benefits to be the relevant measure of well-being at all points in the experiment.

5.4.3 The revised puzzle under certainty

In the previous two sections, I argued that we must assume the self-torturer possesses certainty in order to distinguish between the regretful state of the self-torturer caused by intransitivity and regret caused by uncertainty. I also argued that the self-torturer must base her decisions only on total costs and benefits if she aims to maximize her well-being. In this section, I re-evaluate the PST, incorporating these clarifications in the puzzle. I argue that in this revised version of the PST, the self-torturer's preferences are transitive if she acts rationally.

Consider a revised version of the PST where the self-torturer makes decisions under certainty, meaning that she possesses full information and perfect foresight. The self-torturer also evaluates her well-being in total terms at all points in the puzzle. For the self-torturer's preferences to be rational, they must align with the overarching goal of maximizing her well-being. This means that she cannot rationally prefer an option she knows will make her worse off than an available alternative. Recall the condition on the self-torturer's preferences (III) that states that there is some setting s_m such that for any setting s_n where $n \geq m$, the self-torturer prefers setting 0 with no reward to remaining at s_n with the associated reward. In the original interpretation of the PST, it is argued that it is rational for the self-torturer to prefer to reach s_n because the marginal benefits are

invariably greater than the marginal costs when increasing settings. This claim is false for the self-torturer who evaluates her well-being in total terms. The assumption of certainty means that the self-torturer can locate s_n and she knows that she will be in a terrible state at this setting, as determined by her total costs and benefits. At some lower setting s_l, she is better off than she is at s_n. Therefore, having preferences that lead the self-torturer from s_l to s_n are preferences for states of affairs that she knows will make her worse off than available alternatives. As such, when the self-torturer determines her preferences based on total considerations under certainty, it is no longer the case that increasing settings always makes her better off. The self-torturer knows that reaching *sn* makes her worse off and therefore this setting should be avoided if she is acting rationally.

Furthermore, since it is never rational to knowingly make oneself worse off, the self-torturer should never reach a setting at which she would prefer to give up some money to return to some lower setting. If she can determine that s_n is the level at which she would give up all her rewards to return to level 0, then it is also reasonable to suppose that she can determine other levels at which she would give up none of her rewards and levels at which she would give up some of her rewards to return to a lower setting. We can imagine the self-torturer evaluating the settings by continuously attempting the puzzle by increasing the settings up until the point where she realizes her well-being is lower than it was earlier in the experiment. In doing so, she can eliminate the settings that she regrets reaching, eventually locating some point in the puzzle that is her optimal stopping point (s^*). Under certainty, however, she need not go through the

actual trial and error process to locate s^* since she possesses perfect foresight and full information.

This stopping point may be a single setting or a range of settings. Likely, it is a range of settings since the nature of the puzzle is such that no one setting is obviously different, in terms of costs, to its adjacent settings. Whether it is a single setting or a range is inconsequential to the argument here, which is concerned with whether intransitive preferences are ever rational in the PST. So, long as it is rational to stop before reaching a regretful state, the self-torturer has not exhibited intransitive preferences. Even if the stopping point is a range of settings, it is not rational to proceed past this range to a higher setting because this would be an instance of knowingly making oneself worse off. The self-torturer's stopping point is therefore the last point—either a setting or a range of settings—at which she is not willing to return any of the rewards to return to a lower setting, which marks the point at which her net benefit, or well-being, is maximized. If she proceeded past s^*, then she would end up in a state where she would rather return some of her wealth to return to a lower setting. This is to say that she would regret this state since her well-being is higher at s^*. Therefore, intransitive preferences are not rational because it is never rational for the self-torturer to reach a setting that she knows reduces her well-being.

This may initially sound absurd: how can the self-torturer think she will be worse off at s^*+1 than at s^* (where s^* and s^*+1 are either adjacent settings or the last setting of s^* and the first setting of s^*+1, if they are ranges) if the move from s^* to s^*+1 feels the same to her? And if this is the case, then it seems rational, like at any other setting, for

her to move from s^* to s^*+1 because the only relevant preference to this decision is her preference for more money. If this is the case, then it is rational for the self-torturer to reach a regretful setting. To see why this is incorrect, consider a version of the PST that forces the self-torturer to consider the total effects of her actions. The only difference in this altered PST is that settings are increased once a week like in the original PST but the self-torturer chooses every 5 weeks whether to continue in the experiment or remain at the current setting. The self-torturer cannot feel a difference between adjacent settings but can feel a significant difference in pain between intervals of five setting. Suppose the self-torturer is in a decision-making week at setting 95 and is trying to decide whether to continue the experiment, therefore reaching setting 100 in five weeks, or to stop at setting 95. While the move to setting 96 creates a negligible increase in pain, the pain level she knows she will experience at level 100 will be noticeably different to her current pain level. She evaluates her well-being in terms of her total benefits and total costs at setting 100 and decides that she is better off at setting 95. When evaluating her well-being in marginal terms, the change in her costs from setting 95 to 100 is no longer negligible because the alteration of the puzzle makes the marginal costs informative of the total costs. This means that there is no longer a discrepancy between the relevant total and marginal evaluations. The change in pain level between 90 and 95 is worth the extra $50,000 but she knows that at level 100 the increase in pain is not worth the extra $50,000. She is better off at 95 than she is at 90 and 100. She therefore prefers setting 95 to settings 90 and 100. The negligible marginal costs between adjacent settings no longer give the self-torturer an incentive to continue to the last setting of the experiment

because the relevant marginal costs are large enough that she does not seem to be always better off, in marginal terms, if she stops at the next setting. Hence this scenario is not a puzzle; it is a normal situation where a tradeoff is made between the costs and benefits of an action and the marginal considerations are informative of the total considerations. If preferences in this case are informed and rational, then they are transitive.

If the self-torturer possesses certainty, it is not reasonable for her to act differently in the one-week (original PST) and five-week (altered PST) cases. If setting 100 is not worthwhile in the altered PST, then it is also not worthwhile in the original PST. Therefore, if the self-torturer acts rationally, then she should never reach setting 100 because doing so would make her knowingly worse off. In the altered PST, it is apparent to the self-torturer that somewhere in the range of 95 to 99 her pain level reaches a point such that the extra monetary reward is no longer worth the extra pain, so reaching that point would make her worse off. The decision is more difficult in the one-week case because the self-torturer must choose which setting to stop at in the range of 95 to 99 when adjacent settings feel the same. Unlike the altered PST, the point at which her level of pain shifts from "worthwhile" to "not-worthwhile" is likely difficult to discern. However, if the self-torturer knows that setting 95 is worthwhile and 100 is not worthwhile, then there must be some pair of settings or pair of ranges of settings between 95 and 100 where, if asked what her pain level is at each setting, her answer shifts from "worthwhile" to "not worthwhile."[18] That is, there is a stopping point between 95 and 100.

[18] A similar argument is made by Shelly Kagan (2011, 132-33).

The first premise in Andreou's argument is therefore false when the key PST conditions include negligible marginal costs and certainty. If the self-torturer's level of certainty as well as the way in which she assesses her well-being is explicitly considered, then the key PST conditions do not imply rational yet intransitive preferences. It is not rational for the self-torturer to proceed to a point where she is worse off than a previous setting, according to an evaluation of her total costs and total benefits.

Frank Arntzenius and David McCarthy (1997) offer a different solution to the puzzle. They amend the PST such that the trial period before the start of the experiment offers better information to the self-torturer than was assumed in the original PST. In their version of the puzzle, settings are administered to the self-torturer at random during the trial period. The self-torturer reports her level of pain at each setting, which is recorded by the experimenters. The frequencies of her pain reports at each setting are then given to her and she uses these frequencies as evidence for her pain level at each setting during the experiment. Arntzenius and McCarthy thus describe how the self-torturer can increase her information in order to distinguish between adjacent settings. Like Arntzenius and McCarthy, I also suggest that there is an informational problem in the PST. However, I argue that we need not devise ways for the self-torturer to gain information in order to solve the puzzle. It is necessary to assume that the self-torturer already has full information in order for the puzzle to show what proponents of the PST claim it shows—the existence of rational intransitive preferences.

Furthermore, contrary to what I argue, Arntzenius and McCarthy argue that when the self-torturer has full information, decisions ought to be determined by the marginal

value of the action. The self-torturer uses the pain reports as evidence for her changing pain level at different settings. She then assigns a value to the change in pain level from one setting to the next and she assigns a value to the extra wealth she receives by increasing settings. As such, she assigns a marginal value to both the change in pain and the change in wealth. Therefore, she measures her well-being in terms of the marginal utility (the marginal value of money) and marginal disutility (the marginal value of pain) she gains by moving from one setting to the next. It is further assumed that the marginal value of pain is increasing over settings and the marginal value of money is diminishing. That is, money is valued more highly the less she has of it and pain is valued more highly the more pain she is already experiencing. The self-torturer maximizes her well-being by stopping at the point where marginal utility just outweighs marginal disutility, since this indicates the point at which total utility is maximized.

In Arntzenius and McCarthy's solution, the existence of an optimal stopping point is guaranteed by the diminishing marginal value of money and the increasing marginal value of pain. This is a problematic solution, however, because the concept of diminishing or increasing marginal value presupposes transitivity. Since transitivity is exactly what the PST is calling into question, this is an unwarranted assumption for a solution to the puzzle. Even if we grant Arntzenius and McCarthy the assumption of diminishing and increasing marginal value, however, their solution is problematic particularly in the context of climate change decisions. They purport to solve the puzzle by denying that marginal costs are truly negligible in the PST. They argue that when the self-torturer has enough information, she sees that marginal costs are not negligible

despite their appearance, and therefore she ought to treat them as non-negligible. Therefore, their solution rests on rejecting the key assumption of the PST that a change in costs, or pain, really can be negligible if the total costs are accumulating. They claim that the apparent paradoxical nature of the PST is founded on an illegitimate assumption; changes in pain cannot be truly negligible if pain accumulates throughout the experiment. I maintain, on the contrary, that the assumption of negligible marginal costs is legitimate. Marginal costs that are relevant to the decision maker are negligible because the increase in pain is negligible from one setting to the next. This is compatible with the claim that total costs are increasing when we understand that the marginal costs relevant to the self-torturer are uninformative of total costs. This cost structure occurs in decision problems like climate change where the marginal costs that are relevant to our decisions are negligible even though the total costs are accumulating. The marginal costs are experienced as negligible, and therefore, our model of climate change decisions should treat them as negligible as well. In doing so, we can better understand the implications of this cost structure for decisions like climate change. Therefore, the key feature of the PST that informs our understanding of climate change decisions is just the feature that Arntzenius and McCarthy deny: the disconnect between marginal and total costs.

5.4.4 The revised puzzle under uncertainty

The previous sections isolate the features that make decisions in PST-like settings unusual. When marginal costs are negligible, they are uninformative of the total costs that are nonetheless increasing. Decisions on the margin are therefore not rational in this

context. But, when the decision maker possesses certainty, she knows this and bases her decisions on total costs and benefits. As such, the self-torturer acts rationally by maximizing total benefits net of total costs by stopping at s^*, which is located by total rather than marginal evaluations. In this section, however, I argue that aiming to maximize well-being is likely to result in a regretful outcome when the self-torturer faces uncertainty.

Consider the revised PST under uncertainty. The self-torturer faces uncertainty about the relationship between her pain level and the setting level. Recall that s^* is the highest setting or range of settings at which the self-torturer does not prefer to return to a previous setting by giving back some of her wealth. It is located under certainty by evaluating the net benefit at different settings and choosing the setting where it is maximized. Under uncertainty, the self-torturer knows how much wealth she will have at each setting because each subsequent setting is always associated with an extra $10,000; however, she does not know how much pain she will feel at each setting. She is not uncertain about how to weigh her pain against monetary reward—if she did not know how to evaluate her well-being, then there would be little hope of maximizing well-being in the first place—but she does not know which pain level will occur at each setting. This means that she does not know what her total level of wealth will be when she feels a given level of pain. Since the pain level she is willing to withstand (total costs) depends on how well she is compensated for it (total benefits), if she does not know which pain level she will feel for each possible level of wealth, then she cannot locate the point in the experiment at which net benefit is maximized (s^*).

Furthermore, she cannot "back out" this relationship as the experiment progresses. This could be the case if she thinks that the increase in electricity associated with increased settings is not constant. In some cases, the self-torturer could feel a difference in pain after three settings, but in other cases, the increase in electricity for each setting could be so minute that she only feels a difference after 20 settings. So, for example, the self-torturer knows that setting 10 is associated with a total benefit of $100,000, but she does not know if she will experience any pain or moderate pain at this setting. Depending on how quickly pain accumulates, setting 10 might be s^*. However, when she is at setting 10 she cannot know whether her well-being has reached a maximum because she does not know how her pain will accumulate after setting 10. She might reason that, although well-being is quite high at this setting, it is probable that it will continue to increase after setting 10. As such, it is rational to increase settings. In doing so, however, she might find herself at setting 15 only to realize that her well-being has declined. She now knows that setting 10 is, in fact, s^*. Since her choices are irreversible, however, she does not have the option of returning to setting 10, so she must remain indefinitely in a regretful state.

This regretful outcome is not the result of intransitive preferences, as proponents of the PST argue. Instead, uncertainty combined with negligible marginal costs that are uninformative of total costs prevent the self-torturer from accurately locating her stopping point. Since the self-torturer does not have access to marginal costs as information about how her net benefit is changing, she cannot know where her net benefit is maximized until after it has already begun to decline. Hence, she must proceed

past s^* in order to recognize that s^* is in fact her optimal stopping point, only to find herself in a regretful state in which she would rather give up some monetary reward to return to a lower setting. Furthermore, she might reason that although she passed her global maximum, there could be another local maximum further along in the experiment and therefore aim for a second-best outcome. However, there might not be another local maximum, in which case she would reach an even lower level of well-being. Reaching a regretful state in the attempt to maximize net benefit is particularly problematic in this decision scenario because it is irreversible, thus precluding the opportunity to learn from her regretful decisions.

5.5 The Problem of Climate Change

The revised PST under uncertainty is a useful model for thinking about the challenges we face in making optimal climate change decisions even after private and social interests are aligned. The type of uncertainty faced by the self-torturer is analogous to the uncertainty we face in climate change decisions because precise quantitative predictions of the greenhouse effect are uncertain (Broome 2012, 29). We are not uncertain that if our polluting habits do not change, then we will eventually reach a costly state of environmental damage. However, we do not know precisely when we will reach this level of environmental damage because we do not know the precise quantity of GHGs in the atmosphere that will cause a given temperature change, nor do we know

the precise amount of harm this temperature increase will create. Therefore, the relationship between the amount of GHG pollution (s) and the amount of harm we experience as a result of the environmental damage is uncertain. This means that we cannot accurately locate our optimal stopping point where net benefits are maximized (s^*). Since uncertainty is an essential component of climate change decisions, the inclusion of uncertainty into the PST is a better model for climate change decisions than the puzzle of air pollution that assumes certainty.

Consider the problem of climate change where private interests are aligned with social interests. That is, a social planner chooses annually whether to continue or to stop emitting GHGs with the aim of maximizing social net benefit. The social planner accurately evaluates and compares the benefits of GHG pollution and the harm it causes. The change in harms felt each year caused by the additional quantity of GHG emissions in the atmosphere is negligible and the benefits are huge since their economy is dependent on fossil fuel consumption. It does not know which year, or quantity of GHG emissions, maximizes its well-being because the relationship between the amount of pollution and the harm experienced as a result of the pollution is uncertain. However, there are three expert predictions for how long it will take to reach a certain temperature increase (and the associated amount of harm) that causes a catastrophe (s_n), where each prediction has an associated optimal stopping point (s^*). Well-being is maximized at s^*, determined by the total costs and total benefits of polluting. Accordingly, the social planner believes that s^* will occur in 60 years with a probability of 0.3, in 80 years with a

probability of 0.6, and 200 years with a probability of 0.1. It thus must decide, given these predictions, when to stop polluting.

Since year 80 is the most probable optimal stopping point, it plans to stop polluting in 80 years. When it reaches year 60, it recognizes that its well-being is high but expects the next 20 years of polluting to feel similar in terms of total harms but to significantly increase total benefits. Therefore, it expects net benefit to increase over the next 20 years of polluting, so it chooses to continue polluting. This choice is rational given its uncertainty about the relationship between the quantity of pollution and the associated harms, and given that the negligible marginal costs do not provide information about when net benefit is maximized. However, in year 70, it realizes that its well-being has decreased compared to year 60. It now has the information necessary to recognize that year 60 was in fact its optimal stopping point; however, since the harms of climate change are not reversible (at least given current technology and a long but finite time frame), it must remain indefinitely in a regretful state. It ended up in a regretful and irreversible state because it expected well-being to increase after year 60 when in fact its well-being was maximized in year 60. Negligible marginal costs and uncertainty inhibited its ability at year 60 to recognize that it should stop polluting since it reasoned that an extra 20 years may not feel very different to the harm it was currently experiencing. Despite the social planner basing its decision on the best information available, it nevertheless ended up in a regrettable state because it lacked information about the future and faced marginal costs that were uninformative of its total costs. Additionally, it

had no opportunity to learn about its total costs and total benefits from experience given the irreversibility of climate change decisions.

The series of decisions that leads the self-torturer and the social planner into a regretful state might have the appearance of rational intransitive preferences: decisions that aim to maximize well-being nonetheless lead the decision maker to a regretful state where they would prefer to be in a state that they previously preferred to leave. However, it merely looks like they have rational intransitive preferences when it is not recognized that they are acting under uncertainty without the information about their net benefit that is normally provided by marginal costs. The challenge of this decision scenario results from acting under conditions that make it difficult to locate and recognize when one has reached an optimal stopping point. Given the preferences for more money and less pain as well as the negligible marginal costs, there is incentive to keep polluting in order to capture greater net benefit. At the same time, the damaging effects of pollution accumulate insidiously. Therefore, we should replace Andreou's first premise with the premise that if key conditions arise in the PST, which include uncertainty and individually negligible effects, then a regretful state can understandably arise. These key conditions arise in environmental problems like climate change; therefore, an understandable yet regretful state can arise in environmental problems like climate change even when private and social interests are aligned.

This model offers insight into the problem of climate change. The prominent explanations of climate change treat the problem as primarily caused by interpersonal and intergenerational conflicts of interest. The outcome of independently optimizing

agents does not result in the best outcome for society because maximizing private net benefit does not maximize social net benefit in the presence of an externality. The model of climate change decisions offered here suggests that this is not the only cause of the climate change problem. Even if climate change decisions were made by a social planner who aims to maximize social well-being, meaning that all externalities were eliminated, then we may still end up in a suboptimal state. When we think about climate change decisions as a sequence of intertemporal choices, we see that its cost structure is unusual: the marginal costs are uninformative of total costs. The result of this cost structure is that the social planner cannot know well-being is at a maximum until it has already declined.

In this type of scenario, the aim of maximizing well-being is counterproductive because, in search of a maximum, the social planner must first identify a point of declining well-being. But by the time it has located such a point, it is too late to maximize well-being because climate decisions are irreversible. Aiming to maximize well-being is therefore a strategy that is likely to result in a regretful outcome in which well-being is not maximized. We ought to instead aim to maximize well-being while also aiming to avoid reaching a regretful state. This requires us to stop polluting before we are sure that we have reached a maximum level of well-being, but when we know we have not yet reached a regretful state. This means that we may need to forgo some of the benefits of polluting in order to avoid reaching a regretful state of the environment.

5.6 Conclusion

In this chapter, I argued that the fact that environmental problems tend to be associated with negligible marginal costs does not imply that the social planner has intransitive environmental preferences. I suggested that the PST only appears to reveal rational intransitive preferences when the level of certainty possessed by the self-torturer and the appropriate measure of well-being are not explicitly considered. I argued that an uncertain decision scenario coupled with negligible marginal costs can lead the rational self-torturer or social planner into a regretful state despite having transitive preferences.

This conclusion is less troubling than Andreou's because it does not undermine existing solutions to environmental problems that require the assumption of transitive preferences. For example, my argument indicates that using an optimal tax to internalize an environmental externality is appropriate to the extent that its assumption of characteristically transitive environmental preferences is accurate. Furthermore, it is not the case that we rationally prefer to reach a regretful state, as the original PST suggested. Instead, I argued that aiming for optimality, in the sense of maximizing well-being by balancing the costs and benefits of an activity, can result in a regretful state in scenarios such as climate change. This is because the optimal pollution stopping point at which well-being is maximized cannot be located until well-being has already declined. This is because the social planner faces uncertainty while the marginal costs of polluting are uninformative of the total costs of polluting.

Overshooting the optimal stopping point is therefore likely when the social planner aims to maximize well-being in a scenario that has this peculiar cost. This implies that expected value theory, which is often considered to be the best way to deal with the uncertainties inherent in climate change decisions, may not be the best strategy for problems like climate change. It is not sufficient to merely aim to maximize social well-being by internalizing an externality using a Pigovian tax. In a scenario like climate change, aiming to maximize well-being can lead to a regretful state. Therefore, climate change is not merely an externality problem; we may need to forgo some of the social benefits of polluting in order to avoid reaching an irreversible and regretful state.

CHAPTER SIX

Conclusion

In this dissertation, I developed the gains view of externalities, which is a novel characterization of externalities as they are treated in prominent mainstream microeconomic models. On this view, an externality arises when an untraded activity generates untapped gains from exchange that are associated with untapped welfare gains. This concept does not readily transfer to the actual world, however, because gains from exchange and welfare gains are often divergent.

There are two ways to interpret an externality in the world, both of which are problematic. First, an externality arises when there are actual untapped gains from exchange over an untraded activity that are informative of untapped welfare gains. This interpretation characterizes externalities in such a way that they may rarely arise in the world, depending on how informed and unbiased preferences must be to inform of welfare gains.

Second, an externality arises when hypothetical untapped gains from exchange can be posited for a given untapped welfare gain that is generated by an untraded activity. This interpretation characterizes externalities in such a way that they are

unobservable and ubiquitous. All untraded activities that have a welfare effect on other agents could be interpreted as an externality, depending on the judgment of the researcher. This interpretation of an externality thus characterizes a Pigovian tax as a paternalistic policy. This is because the tax addresses what individuals should prefer and what they should be willing to pay, rather than what they actually prefer and what they are actually willing to pay for an untraded activity. Therefore, the policymaker must know better what is good for citizens than the citizens themselves. Depending on one's political leaning, this could be a strength or a weakness of this interpretation of an externality.

I thus initially focused on economic models of externalities to show how the two-agent constrained optimization framework establishes the features of an externality that, in turn, establish the appropriate policy response to unpriced activities. I then showed that it is unclear how to interpret an externality in the world. Subsequently, I showed that the contingent valuation method, which aims to estimate the magnitude of an externality, suffers from the ambiguity inherent to the concept of an externality. Finally, I argued that climate change is not merely an externality problem. Even if all externalities were eliminated, there would still be an incentive for a social planner to over-pollute. This incentive is caused by the uncertainty of climate change coupled with its peculiar cost structure in which the marginal costs of polluting are uninformative of the total costs.

It is not clear how to interpret externalities in the world, especially since it is plausible that individuals often hold false beliefs or are biased in policy-relevant scenarios like climate change. Addressing the climate change problem is therefore not as simple as

internalizing an externality with a Pigovian tax, as economists suggest (Akerlof et al 2019). It is unclear what constitutes an externality and, on any interpretation of an externality, the climate change problem is more than an externality problem. Economists could thus better serve policymakers by addressing the ambiguity of their policy recommendations and by addressing the additional challenges the climate change problem presents for policymakers.

As it stands, I have argued that the concept of an externality does not provide policy guidance; this means that a Pigovian tax or trading schemes that aim to internalize an externality, and thus achieve economic efficiency, are not well-supported policies. Nonetheless, a tax on carbon emissions or a carbon trading scheme might be useful not to internalize an externality but instead to achieve a desired quantity of emissions reductions. The difference is whether the aim of the policy is to internalize an externality and thus to achieve the efficient level of emissions, or whether the aim is to reduce emissions to a pre-specified level using a cost-effect market-based policy instrument. My argument concerns the former, but not the latter, type of tax. This argument thus calls into question the coherence of a Pigovian carbon tax, not any tax on carbon emissions in general.

There are several aspects of the concept of an externality that require further attention. As I noted in the introductory chapter, in this dissertation I set aside the issue of estimating the present value of costs that accrue to future generations. This is an issue for interpreting externalities in the world, however, because it implies that the untapped gains from exchange that constitute the externality are between agents whose lives do

not overlap. It is not clear that gains from exchange are a meaningful estimate of welfare gains in this context. Therefore, it is plausible that this intergenerational problem cannot be interpreted as an externality problem or that an intergenerational externality has distinct constituents from an intra-generational externality. John Broome (2018), for example, develops a notion of economic efficiency that can accommodate intergenerational problems and thus develops a slightly modified version of the concept of an intergenerational externality.

Furthermore, the method some economists use to estimate the social cost of carbon, which is the value of the externality generated by carbon emissions, is unanalyzed in this dissertation. Economists such as William Nordhaus (2014; 2017) and Nicholas Stern (2006; 2014a; 2014b) estimate the social cost of carbon using integrated assessment models. These models attempt to link economic models with climate science models to produce an estimate of the social cost of carbon. It is unclear, however, that integrated assessment models are estimating an externality, at least as it is characterized in economic theory. This is because integrated assessment models estimate the value of the externality in the macro-economy and assume that this estimate is identical to the sum of the willingness to pay of individuals for a reduction in carbon dioxide emissions. It is not clear, however, that this reductionist assumption is warranted (Hoover 2010; Nelson 1984). That is, it is not clear that the microeconomic theory that justifies the use of a carbon tax is relevant to the macroeconomic estimates that are generated by integrated assessment models.

Furthermore, there are significant differences between markets, missing markets, and policies that mimic the outcomes of markets that I have not yet examined. I hypothesize that the justification for taking as given unequal income distributions when analyzing the benefits of markets does not extend to missing markets or market-based policies. This justification rests on the claim that competition between firms results in the provision of goods by markets at their lowest possible price. However, missing markets lack competition. Therefore, they also lack an important justification for using markets to allocate unpriced goods in society. If this is right, then the differences between markets, missing markets, and market-based policies are sufficient to warrant differential treatment in economics.

Lastly, I have not yet developed a positive account of externalities. How should externalities be understood given the problems with the concept? Externalities seem to do little conceptual work in addressing environmental problems apart from identifying that there is a social issue that markets do not adequately address. Market-based policies therefore should not be used with the intention of internalizing an externality, the size of which determines the socially optimal level of pollution. Instead, these policies should be used as a cost-effective way to create incentives to reduce pollution to a level that is determined independently of economic views of Pareto efficiency. These are some of the problems that will guide my research in the coming years.

Bibliography

Akerlof, George A. 1970. "The Market for "Lemons": Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84: 488-500.

Akerlof, George, Robert Aumann, Angus Deaton, Peter Diamond, Robert Engle, Eugene Fama, Lars Peter Hansen, Oliver Hart, Bengt Holmström, Daniel Kahneman, Finn Kydland, Robert Lucas, Eric Maskin, Daniel McFadden, Robert Merton, Roger Myerson, Edmund Phelps, Alvin Roth, Thomas Sargent, Myron Scholes, Amartya Sen, William Sharpe, Robert Shiller, Christopher Sims, Robert Solow, Michael Spence, Richard Thaler, Paul Volcker, Martin Baily, Michael Boskin, Martin Feldstein, Jason Furman, Austan Goolsbee, Glenn Hubbard, Alan Krueger, Edward Lazear, N. Gregory Mankiw, Christina Romer, Harvey Rosen, Laura Tyson, Ben Bernanke, Alan Greenspan, Janet Yellen, George Shultz and Lawrence Summers. 2019. "Economists' Statement on Carbon Dividends: Bipartisan agreement on how to combat climate change." *Wall Street Journal,* Jan 16, 2019. Accessed on May 6, 2019. https://www.wsj.com/articles/economists-statement-on-carbon-dividends-11547682910.

Aldy, Joseph E., Matthew J. Kotchen, and Anthony A. Leiserowitz. 2012. "Willingness to Pay and Political Support for a U.S. National Clean Energy Standard." *Nature Climate Change* 2(5): 596-599.

Alexandrova, Anna. 2008. "Making Models Count." *Philosophy of Science* 75, no. 3: 383–404.

———. 2015. "Well-Being and Philosophy of Science." *Philosophy Compass* 10, no. 3: 219–31.

Andreou, Chrisoula. 2006. "Environmental Damage and the Puzzle of the Self-Torturer." *Philosophy & Public Affairs* 34, no. 1: 95–108.

Angner, Erik. 2013. "Is It Possible to Measure Happiness?" *European Journal for Philosophy of Science* 3, no. 2: 221–40.

Anthoff, David, and Richard S. J. Tol. 2010. "On International Equity Weights and National Decision Making on Climate Change." *Journal of Environmental Economics and Management* 60, no. 1: 14–20.

———. 2013. "The Uncertainty About the Social Cost of Carbon: A Decomposition Analysis Using Fund." *Climatic Change* 117, no. 3: 515–30.

Arntzenius, Frank, and David McCarthy. 1997. "Self Torture and Group Beneficence." Erkenntnis 47, no. 1: 129–44.

Arrow Kenneth J. 1969. "The organization of economic activity: issues pertinent to the choice of market versus non market allocation." In *The Analysis and Evaluation of Public Expenditures: The PPB System*. Vol. 1*,* 59-73. Congress of the United States.

———. 1970. "Political and Economic Evaluation of Social Effects and Externalities." In *The Analysis of Public Output*, edited by Julius Margolis, 1-31. New York: Columbia University Press.

Arrow, Kenneth J., and Anthony C Fisher. 1974. "Environmental Preservation, Uncertainty, and Irreversibility." *The Quarterly Journal of Economics* 88, no. 2: 312–19.

Arrow, Kenneth J., Robert Solow, Paul R. Portney, Edward E. Leamer, Roy Radner, and Howard Schuman. 1993. "Report of the NOAA Panel on Contingent Valuation," *Federal Register* 58, no. 10: 4601-14.

Aslanbeigui, Nahid, and Steven G. Medema. 1998. "Beyond the Dark Clouds: Pigou and Coase on Social Cost." *History of Political Economy* 30, no. 4: 601-619.

Aslanbeigui, Nahid, and Guy Oakes. 2012. "On Pigou's Theory of Economic Policy Analysis." Œconomia 2, no. 2: 123–50.

Auffhammer, Maximilian. 2018. "Quantifying Economic Damages from Climate Change." *Journal of Economic Perspectives* 32, no. 4: 33–52.

Aufrecht, Monica. 2011. "Climate Change and Structural Emissions: Moral Obligations at the Individual Level." *International Journal of Applied Philosophy* 25, no. 2: 201–13.

Backhouse, Roger E. 2005. "The Rise of Free Market Economics: Economists and the Role of the State Since 1970." *History of Political Economy* 37 (Suppl. 1): 355–92.

———. 2015. "Economic Power and the Financial Machine: Competing Conceptions of Market Failure in the Great Depression." *History of Political Economy* 47 (Suppl. 1): 99–126.

Backhouse, Roger E, and Matthias Klaes. 2009. "Applying Economics, Using Evidence." *Journal of Economic Methodology* 16, no. 2: 139–44.

Banerjee, Simanti, Timothy N Cason, Frans P de Vries, and Nick Hanley. 2017. "Transaction Costs, Communication and Spatial Coordination in Payment for Ecosystem Services Schemes." *Journal of Environmental Economics and Management* 83: 68–89.

Banzhaf, H. Spencer. 2011. "Consumer Sovereignty in the History of Environmental Economics." *History of Political Economy* 43, no. 2: 339–45.

———. 2017. "Constructing Markets." *History of Political Economy* 49: 213–39.

Bartha, Paul, and C Tyler DesRoches. 2016. "The Relatively Infinite Value of the Environment." *Australasian Journal of Philosophy* 95, no. 2: 1–26.

Bator, Francis M. 1958. "The Anatomy of Market Failure." *The Quarterly Journal of Economics* 72, no. 3: 351–79.

Batterman, Robert W. 2002. "Asymptotics and the Role of Minimal Models." *The British Journal for the Philosophy of Science* 53, no. 1: 21–38.

Baumol, William J., and Wallace E. Oates. 1975. *The Theory of Environmental Policy, Second Edition.* Cambridge: Cambridge University Press.

Beckerman, Wilfred, and Joanna Pasek. 1997. "Plural Values and Environmental Valuation." *Environmental Values* 6: 65–86.

Berta, Nathalie, and Elodie Bertrand. 2014. "Market Internalization of Externalities: What Is Failing?" *Journal of the History of Economic Thought* 36, no. 3: 331–57.

Berta, Nathalie. 2017. "On the Definition of Externality as a Missing Market." *The European Journal of the History of Economic Thought* 24, no. 2: 287–318.

Bhatia, M. R. and J. A. Fox-Rushby. 2003. "Validity of Willingness to Pay: Hypothetical Versus Actual Payment." *Applied Economics Letters* 10, no. 12: 737–40.

Bicchieri, Cristina. 2016. *Norms in the Wild.* Oxford: Oxford University Press.

Binmore, Ken. 1997. "Rationality and Backward Induction." *Journal of Economic Methodology* 4, no. 1: 23–41.

Blaug, Mark. 1962. *Economic Theory in Retrospect*. Portsmouth, NH: Heinemann Educational Books Ltd.

Boumans, Marcel. 2005. "Measurement Outside the Laboratory." *Philosophy of Science* 72 (5): 850–63.

Breffle, William S, Mark E Eiswerth, Daya Muralidharan, and Jeffrey Thornton. 2015. "Understanding How Income Influences Willingness to Pay for Joint Programs:

A More Equitable Value Measure for the Less Wealthy." *Ecological Economics* 109: 17–25.

Brennan, Geoffrey. 2008. "Market Failure: Compared to What?" *Ethics and Economics* 6, no. 1: 1-6.

Broome, John. 1991. *Weighing goods*. Oxford: Blackwell.

———. 1994. "Discounting the Future." *Philosophy & Public Affairs* 23, no. 2: 128–56.

———. 2000. "Cost-Benefit Analysis and Population." *The Journal of Legal Studies* 29, no. 2: 953–70.

———. 2012. *Climate Matters: Ethics in a Warming World*. New York: W. W. Norton & Co.

———. 2013. "A Small Chance of Disaster." *European Review* 21 (Suppl. 1): S27–S31.

———. 2018. "Efficiency and Future Generations." *Economics and Philosophy* 34, no. 2: 221-241.

Brown, Mark B. 2014. "Climate Science, Populism, and the Democracy of Rejection." In *Culture, Politics and Climate Change: How Information Shapes Our Common Future*, edited by Deserai A. Crow and Maxwell T. Boykoff, 129-45. London: Routledge.

Buchanan, James M. 1962. "Politics, Policy, and the Pigovian Margins." *Economica* 29, no. 113: 17–28.

———. 1969. "External Diseconomies, Corrective Taxes, and Market Structure." *American Economic Review* 59, no. 1: 174–77.

Buchanan, James M, and William C. Stubblebine. 1962. "Externality." *Economica* 29, no. 116: 371–84.

Caldari, Katia, and Fabio Masini. 2011. "Pigouvian Versus Marshallian Tax: Market Failure, Public Intervention and the Problem of Externalities." *The European Journal of the History of Economic Thought* 18, no. 5: 715–32.

Caney, Simon. 2014. "Climate Change, Intergenerational Equity and the Social Discount Rate." *Politics, Philosophy & Economics* 13, no. 4: 320–42.

Carley, Sanya. 2011. "Normative Dimensions of Sustainable Energy Policy." *Ethics, Policy & Environment* 14, no. 2: 211–29.

Carson, Richard T. 2012. "Contingent Valuation: A Practical Alternative When Prices Aren't Available." *Journal of Economic Perspectives* 26, no. 4: 27–42.

Carson, Richard T., Robert C. Mitchell, W. Michael Hanemann, Raymond J. Kopp, Stanley Presser, Paul A. Ruud. 1992. "A Contingent Valuation Study of Lost Passive Use Values Resulting from the Exxon Valdez Oil Spill." *A Report to the Attorney General of the State of Alaska,* November 10, 1992.

Carson, Richard T., Leanne Wilks, and David Imber. 1994. "Valuing the Preservation of Australia's Kakadu Conservation Zone." *Oxford Economic Papers,* New Series, Special Edition on Environmental Economics 46: 727-749.

Carson, Richard T., and Theodore Groves. 2007. "Incentive and Informational Properties of Preference Questions. *Environmental and Resource Economics* 37, no. 1: 181-210.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.

———. 2007. *Hunting Causes and Using Them: Approaches in Philosophy and Economics.* Cambridge: Cambridge University Press.

———. 2012. “Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps.” *Philosophy of Science* 79, no. 5: 973–89.

Cartwright, Nancy, and Eleanora Montuschi, eds. 2014. *Philosophy of Social Science: A New Introduction*. Oxford: Oxford University Press.

Chakravartty, Anjan. 2017. "Scientific Realism." In *Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta. Accessed on May 21, 2019. https://plato.stanford.edu/archives/sum2017/entries/scientific-realism/.

Christie, Michael. 2007. “An Examination of the Disparity Between Hypothetical and Actual Willingness to Pay Using the Contingent Valuation Method: The Case of Red Kite Conservation in the United Kingdom.” *Canadian Journal of Agricultural Economics* 55: 159–69.

Ciracy-Wantrub, S. V. 1947. “Capital Returns from Soil Conservation Practices.” *Journal of Farm Economics* 29: 1181-96.

Claassen, Rutger. 2016. “Externalities as a Basis for Regulation: A Philosophical View.” *Journal of Institutional Economics* 12, no. 3: 541–63.

Clark, Judy, Jacquelin Burgess, and Carolyn M Harrison. 2000. “‘I Struggled with This Money Business’: Respondents’ Perspectives on Contingent Valuation.” *Ecological Economics* 33: 45–62.

Coase, R H. [1960] 2013. “The Problem of Social Cost.” *The Journal of Law and Economics* 56, no. 4: 837-877.

Colander, David. 2015. “Framing the Economic Policy Debate.” *History of Political Economy* 47 Suppl. 1): 253–66.

Cookson, Richard. 1996. "Welfare Economic Dogmas: A Reply to Sagoff." *Environmental Values* 5: 59–74.

Cropper, Maureen L., and Wallace E. Oates. 1992. "Environmental Economics: A Survey." *Journal of Economic Literature* 30, no. 2: 675-740.

Dahlman, Carl J. 1979. "The Problem of Externality." *The Journal of Law and Economics* 141, no. 162: 1–23.

Davidson, Donald, J. C. C. McKinsey, and Patrick Suppes. 1955. "Outlines of a Formal Theory of Value, I." *Philosophy of Science* 22: 140-60.

Davidson, Donald. 1980. *Essays on Action and Events*. Oxford: Clarendon Press.

Davis, Robert K. 1963. "Recreation Planning as an Economic Problem. *Natural Resources Journal* 3: 239-49.

Deblonde, Marian K. 2000. "Environmental Economics: The Meaning of an 'Objective' Policy Science." *Environmental Values* 9, no. 2: 235–48.

Debreu, Gerard. 1959. *Theory of Value: An Axiomatic Analysis of Economic Equilibrium.* New York: John Wiley & Sons.

Demsetz, Harold. 1996. "The Core Disagreement Between Pigou, the Profession, and Coase in the Analyses of the Externality Question." *European Journal of Political Economy* 12, no. 4: 565–79.

Diamond, Peter, and Jerry A. Hausman. 1994. "Contingent Valuation: Is Some Number Better Than No Number?" *The Journal of Economic Perspectives* 8, no. 4: 45–64.

Dietsch, Peter. 2008. "Does Market Failure Justify Redistribution?" *Ethics and Economics* 6, no. 1: 1-7

Eatwell, John, Murray Milgate, and Peter Newman, eds. 1989a. *Allocation, Information, and Markets*. New York: W. W. Norton.

———, eds. 1989b. *The Invisible Hand*. London: Macmillan Reference.

European Commission. n.d. "EU Emissions Trading System (EU ETS)." *European Commission: Policies, Information and Services*. Accessed on April 3, 2019. https://ec.europa.eu/clima/policies/ets_en.

Egan, Patrick J., and Megan Mullin. 2017. "Climate Change: US Public Opinion." *Annual Review of Political Science* 20: 209-227.

Elster, Jon. 1979. *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

———. 1983. *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Fankhauser, Samuel, Richard S. J. Tol, and David W. Pearce. 1997. "The Aggregation of Climate Change Damages: A Welfare Theoretic Approach." *Environmental and Resource Economics* 10: 249–66.

Field, Barry C., and Martha K. Field. 2001. *Environmental Economics*. New York: McGraw-Hill Higher Education.

Fletcher, Guy. 2016. *The Routledge Handbook of the Philosophy of Well-being*. London: Routledge.

Fontaine, Philippe. 2014. "Free Riding." *Journal of the History of Economic Thought* 36, no. 3: 359–76.

Friedman, Milton. 1953. "The Methodology of Positive Economics" in *Essays in Positive Economics,* 3-43. Chicago: Chicago University Press.

Frodeman, Robert. 2006. "The Policy Turn in Environmental Ethics." *Environmental Ethics* 28, no. 1: 3–20.

Fullerton, Don, and Robert Stavins. 1998. "How Economists See the Environment." *Nature* 395, no. 6701: 433–34.

Gardiner, Stephen M. 2001. "The Real Tragedy of the Commons." *Philosophy & Public Affairs* 30, no. 4: 387–416.

———. 2006. "A Core Precautionary Principle." *Journal of Political Philosophy* 14, no. 1: 33-60

———. 2011. *A Perfect Moral Storm: The Ethical Tragedy of Climate Change*. Oxford: Oxford University Press.

Gaus, Gerald. 2008. *On Philosophy, Politics, and Economics*. Belmont CA: Thomson Wadsworth.

Gollier, Christian, and Nicolas Treich. 2003. "Decision-Making Under Scientific Uncertainty: The Economics of the Precautionary Principle." *Journal of Risk and Uncertainty* 27, no. 1: 77–103.

Government of Canada. 2017. "Pricing Carbon Pollution in Canada: How it will work." *Environment and Climate Change Canada.* Last modified June 21, 2017. Accessed on June 16, 2019. https://www.canada.ca/en/environment-climate-change/news/2017/05/pricing_carbon_pollutionincanadahowitwillwork.html.

Grüne-Yanoff, Till. 2008. "Learning from Minimal Economic Models." *Erkenntnis* 70, no. 1: 81–99.

———. 2012. "Paradoxes of Rational Choice Theory." In *Handbook of Risk Theory Epistemology, Decision Theory, Ethics, and Social Implications of Risk*, edited by Sabine

Roeser, Rafaela Hillerbrand, Per Sandin, and Martin Peterson, 499-516. London: Springer.

Guala, Francesco. 2001. "Building Economic Machines: the FCC Auctions." *Studies in History and Philosophy of Science* 32, no. 3: 453–77.

Haab, T.C., M.G. Interis, D.R. Petrolia, and J.C. Whitehead. 2013. "From Hopeless to Curious? Thoughts on Hausman's "Dubious to Hopeless" Critique of Contingent Valuation." *Applied Economic Perspectives and Policy* 35, no. 4: 593–612.

Hahn, Robert W., and Robert N. Stavins. 1992. "Economic Incentives for Environmental Protection: Integrating Theory and Practice." *American Economic Review* 82, no. 2: 464-68.

Hale, Benjamin. 2011. "Nonrenewable Resources and the Inevitability of Outcomes." *The Monist* 94, no. 3: 369–90.

Hanemann, W. Michael. 1991. "Willingness-to-pay vs. Willingness-to-accept: How Much Can They Differ?" *American Economic Review* 81, no. 3: 635-47.

———. 1994. "Valuing the Environment Through Contingent Valuation." *The Journal of Economic Perspectives* 8 (October): 19–43.

Hansen, Fredrik. 2011 "The Stern Review and its Critics: Economics at Work in an Interdisciplinary Setting." *Journal of Economic Methodology* 18, no. 3: 255–270

Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162: 1243–48.

Hare, R.M. 1952. *The Language of Morals*, Oxford: Clarendon Press.

Hart, Rob. 2001. "The Indifference Curve, Motivation, and Morality in Contingent Valuation." *Environmental Values* 10: 225–42.

Hausman, Daniel M. 1992a. "When Jack and Jill Make a Deal." *Social Philosophy and Policy* 9, no. 1: 95–113.

———. 1992b. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

———. 2008. "Market Failure, Government Failure, and the Hard Problems of Cooperation." *Ethics and Economics* 6, no. 1: 1-6.

———. 2010. "Hedonism and Welfare Economics." *Economics and Philosophy* 26, no. 3: 321–44.

———. 2012. *Preference, Value, Choice, and Welfare.* Cambridge: Cambridge University Press.

———. 2016. "On the Econ Within." *Journal of Economic Methodology* 23, no. 1: 26-32.

———. 2018. "Behavioural Economics and Paternalism." *Economics and Philosophy* 34, no. 1. 53–66.

Hausman, Daniel M., and Brynn Welch. 2010. "Debate: to Nudge or Not to Nudge." *Journal of Political Philosophy* 18, no. 1: 123–36.

Hausman, Daniel M., and Michael S. McPherson. 2006. *Economic Analysis, Moral Philosophy, and Public Policy.* 2nd ed. Cambridge: Cambridge University Press.

Hausman, Daniel, Michael McPherson, and Debra Satz. 2017. *Economic Analysis, Moral Philosophy, and Public Policy.* 3rd ed. Cambridge: Cambridge University Press.

Hausman, Jerry A., ed. 1993. *Contingent Valuation: A Critical Assessment.* Amsterdam: North Holland.

———. 2012. "Contingent Valuation: From Dubious to Hopeless." *Journal of Economic Perspectives* 26, no. 4: 43–56.

Heath, Joseph. 2013. "The Structure of Intergenerational Cooperation." *Philosophy & Public Affairs* 41, no. 1: 31-66.

Heller, Walter P, and David A. Starrett. 1976. "On the Nature of Externalities." In *Theory and Measurement of Economic Externalities*, edited by Steven A. Y. Lin. New York: Academic Press.

Hiller, Avram. 2011. "Climate Change and Individual Responsibility." *The Monist* 94, no. 3: 349–68.

Hodgson, Geoffrey M. 2007. "Meanings of Methodological Individualism." *Journal of Economic Methodology* 14, no. 2: 211–26.

Holland, Stephen P, Erin T Mansur, Nicholas Z Muller, and Andrew J Yates. 2016. "Are There Environmental Benefits from Driving Electric Vehicles? The Importance of Local Factors." *American Economic Review* 106, no. 12: 3700–3729.

Holton, Richard. 1999. "Intention and Weakness of Will." *Journal of Philosophy* 96: 241-262.

Hoover, Kevin D. 2010. "Idealizing Reduction: the Microfoundations of Macroeconomics." *Erkenntnis* 73, no. 3: 329–47.

Hume, David. [1738] 2000. *A Treatise of Human Nature*. Edited by David Fate Norton and Mary J. Norton. Oxford: Oxford University Press.

Infante, Gerardo, Guilhem Lecouteux, and Robert Sugden. 2016. "'On the Econ Within': A Reply to Daniel Hausman." *Journal of Economic Methodology* 23, no. 1: 33-37.

Intergovernmental Panel on Climate Change. 2013. "Summary for Policymakers." In *Climate Change 2013—The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by

Thomas F. Stocker, Dahe Qin, Gian-Kasper Plattner, Melinda M. B. Tignor, Simon K. Allen, Judith Boschung, Alexander Nauels, Yu Xia, Vincent Bex, and Pauline M. Midgley, 3-29. Cambridge: Cambridge University Press.

Jamieson, Dale, ed. 2001. *A Companion to Environmental Philosophy.* Malden MA: Blackwell Publishing

———. 2014. *Reason in a Dark Time*. Oxford: Oxford University Press.

Jeuland, Marc, Marcelino Lucas, John Clemens, and Dale Wittington. 2009. "A Cost-Benefit Analysis of Cholera Vaccination Programs in Beira, Mozambique." *World Bank Economic Review* 23, no. 2: 235-67.

Jiang, Yi, Leshan Jin, and Tun Lin. 2011. "Higher Water Tariffs for Less River Pollution—Evidence from the Min River and Fuzhou City in China." *China Economic Review* 22, no. 2: 183-95.

Johansson-Stenman, Olof. 1998. "The Importance of Ethics in Environmental Economics with a Focus on Existence Values." *Environmental and Resource Economics* 11: 429–42.

Johnson, Bruce K., and John C. Whitehead. 2007. "Value of Public Goods from Sports Stadiums: The CVM Approach." *Contemporary Economic Policy* 18, no. 1: 48-58.

Johnson, Marianne. 2015. "Public Goods, Market Failure, and Voluntary Exchange." *History of Political Economy* 47 (Suppl. 1): 174–98.

Kagan, Shelly. 2011. "Do I Make a Difference?" *Philosophy & Public Affairs* 39, no. 2: 105–41.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler. 1991. "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias." *The Journal of Economic Perspectives* 5, no. 1: 193–206.

Kahneman, Daniel, and Jack L. Knetsch. 1992. "Valuing Public Goods: Purchase of Moral Satisfaction." *Journal of Environmental Economics and Management* 22, no. 1: 57-70.

Kahneman, Daniel, and Alan B. Krueger. 2006. "Developments in the Measurement of Subjective Well-Being." *Journal of Economic Perspectives* 20, no. 1: 3–24.

Kelleher, J P. 2015. "Is There a Sacrifice-Free Solution to Climate Change?" *Ethics, Policy & Environment* 18, no. 1: 68–78.

Koning, Koen de, Tatiana Filatova, and Okmyung Bin. 2017. "Bridging the Gap Between Revealed and Stated Preferences in Flood-Prone Housing Markets." *Ecological Economics* 136: 1–13.

Krutilla, John V. 1967. "Conservation Reconsidered." *The American Economic Review* 57, no. 4: 777–86.

Lagueux, Maurice. 2010. "The Residual Character of Externalities." *The European Journal of the History of Economic Thought* 17, no. 4: 957–73.

Landry, Craig E, and John A. List. 2007. "Using Ex Ante Approaches to Obtain Credible Signals for Value in Contingent Markets: Evidence from the Field." *American Journal of Agricultural Economics* 89, no. 2: 420–29.

Lawford-Smith, Holly. 2014. "Benefiting from Failures to Address Climate Change." *Journal of Applied Philosophy* 31, no. 4: 392–404.

Lester, Richard A. 1946. "Shortcomings of Marginal Analysis for Wage-Employment Problems." *American Economic Review* 36: 62-82.

Lewis, David. 1973. "Causation" *Journal of Philosophy* 70: 556–567.

Livernois, Rebecca. 2018. "Regretful Decisions and Climate Change." *Philosophy of the Social Sciences* 48, no. 2: 168-191.

Loomis, John, Armando Gonzalez-Caban, and Robin Gregory. 1994. "Do Reminders of Substitutes and Budget Constraints Influence Contingent Valuation Estimates?" *Land Economics* 70, no. 4: 499-506

Maas, Harro, and Andrej Svorenčík. 2017. "'Fraught with Controversy': Organizing Expertise against Contingent Valuation." *History of Political Economy* 49, no. 2: 315-45

Macleod, Colin M. 2008. "Market Failure, Justice, and Preferences." *Ethics and Economics* 6, no. 1: 1-7.

Mäki, Uskali. 2002. *Fact and Fiction in Economics*. Cambridge: Cambridge University Press.

Marciano, Alain, and Steven G Medema. 2015. "Market Failure in Context: Introduction." *History of Political Economy* 47 (suppl. 1): 1–19.

Marciano, Alain. 2011. "Buchanan on Externalities: An Exercise in Applied Subjectivism." *Journal of Economic Behavior & Organization* 80, no. 2: 280–89.

Marshall, Alfred. [1890] 1961. *Principles of Economics*. Edited by Claude William Guillebaud. London: Macmillan.

Mas-Colell, Andreu, Michael Dennis Whinston, and Jerry R Green. 1995. *Microeconomic Theory*. Oxford: Oxford University Press.

Matthews, Yvonne, Riccardo Scarpa, and Dan Marsh. 2017. "Stability of Willingness-to-Pay for Coastal Management: A Choice Experiment Across Three Time Periods." *Ecological Economics* 138: 64–73.

Meade, J E. 1952. "External Economies and Diseconomies in a Competitive Situation." *The Economic Journal* 62, no. 245: 54–67.

Medema, Steven G. 2007. "The Hesitant Hand: Mill, Sidgwick, and the Evolution of the Theory of Market Failure." *History of Political Economy* 39, no. 3: 331–58.

———. 2009. *The Hesitant Hand*. Princeton: Princeton University Press.

———. 2014. "The Curious Treatment of the Coase Theorem in the Environmental Economics Literature, 1960-1979." *Review of Environmental Economics and Policy* 8, no. 1: 39–57.

Meng, Kyle C. 2017. "Using a Free Permit Rule to Forecast the Marginal Abatement Cost of Proposed Climate Policy." *American Economic Review* 107, no. 3: 748–84.

Mitchell, Robert Cameron, and Richard T. Carson. 1989. *Using Surveys to Value Public Goods: The Contingent Valuation Method.* Washington DC: Resources for the Future.

Montuschi, Eleanora. 2014. "Scientific Objectivity." In *Philosophy of Social Science: A New Introduction,* edited by Nancy Cartwright and Eleanora Montuschi. Oxford: Oxford University Press.

Morgan, Mary S. "Models, Stories, and the Economic World." In *Fact and Fiction in Economics: Models, Realism and Social Construction*, edited by Uskali Maki, 178-201. Cambridge: Cambridge University Press.

———. 2012. *The World in the Model*. Cambridge: Cambridge University Press.

Morgan, Mary S, and Margaret Morrison, eds. 1999. *Models as Mediators*. Cambridge: Cambridge University Press.

Moscati, Ivan. 2013. "Were Jevons, Menger, and Walras Really Cardinalists? On the Notion of Measurement in Utility Theory, Psychology, Mathematics, and Other Disciplines, 1870-1910." *History of Political Economy* 45, no. 3: 373–414.

Munda, Giuseppe. 1997. "Environmental Economics, Ecological Economics, and the Concept of Sustainable Development." *Environmental Values* 6, no. 2: 213–33.

———. 2016. "Beyond Welfare Economics: Some Methodological Issues." *Journal of Economic Methodology* 23, no. 2: 185-202.

Nagel, Thomas. 1986. *The View from Nowhere.* Oxford: Oxford University Press.

Nelson, Alan. 1984. "Some Issues Surrounding the Reduction of Macroeconomics to Microeconomics." *Philosophy of Science* 51, no. 4: 573-594.

Nordhaus, William. 2014. "Estimates of the Social Cost of Carbon: Concepts and Results From the DICE-2013R Model and Alternative Approaches." *Journal of the Association of Environmental and Resource Economists* 1, no. 1/2: 273–312.

———. 2017. "Revisiting the Social Cost of Carbon." *Proceedings of the National Academy of Sciences* 114, no. 7: 1518–23.

Nyborg, K. 2000. "Homo Economicus and Homo Politicus: Interpretation and Aggregation of Environmental Values." *Journal of Economic Behavior & Organization* 42, no. 3: 305–22.

O'Neil, John. 1992. "The Varieties of Intrinsic Value." *The Monist* 75, no. 2: 199-137.

———. 2001. "Meta-ethics." In *A Companion to Environmental Philosophy,* edited by Dale Jamieson, 163-176. Malden MA: Blackwell Publishing.

Papandreou, Andreas A. 1994. *Externality and Institutions.* New York: Clarendon Press.

———. 2003. "Externality, Convexity and Institutions." *Economics and Philosophy* 19, no. 2: 281–309.

Pareto, Vilfredo. [1909] 1971. *Manual of Political Economy*. Edited by. A. S. Schwier and A. N. Page. New York: Augustus M. Kelley.

Parfit, Derek. 1984. *Reasons and Persons.* Oxford: Oxford University Press.

Parker, Wendy. 2014. "Climate Change." In *Philosophy of Social Science: A New Introduction,* edited by Nancy Cartwright and Eleonora Montuschi, 31-47. Oxford: Oxford University Press.

Pearce, David. 1991. "The Role of Carbon Taxes in Adjusting to Global Warming." *The Economic Journal* 101, no. 407: 938-948.

———. 2002. "An Intellectual History of Environmental Economics." *Annual Review of Energy and the Environment* 27, no. 1: 57–81.

Perman, Roger, Yue Ma, Michael Common, David Maddison, and James Mcgilvray. 2013. *Natural Resource and Environmental Economics*. 4th ed. London: Pearson Education.

Pigou, Arthur. [1920] 2017. *The Economics of Welfare*. London: Routledge.

Plott, Charles R. 1966. "Externalities and Corrective Taxes." *Economica, New Series* 33, no. 129: 84-87.

Portney, Paul R. 1994. "The Contingent Valuation Debate: Why Economists Should Care." *Journal of Economic Perspectives* 8, no. 4: 3–17.

———. 2000. "Environmental Problems and Policy: 2000-2050." *The Journal of Economic Perspectives* 14: 199–206.

Prior, Michael. 1998. "Economic Valuation and Environmental Values." *Environmental Values* 7, no. 4: 423–41.

Quinn, Warren S. 1990. "The Puzzle of the Self-Torturer." *Philosophical Studies* 59, no. 1: 79–90.

Railton, Peter. 1986. "Facts and Values." *Philosophical Topics* 24: 5-31.

Read, Daniel. 2005. "Monetary Incentives, What Are They Good For?" *Journal of Economic Methodology* 12, no. 2: 265–76.

Reiss, Julian. 2012. "Idealization and the Aims of Economics: Three Cheers for Instrumentalism." *Economics and Philosophy* 28, no. 3: 363–83.

———. "The Explanation Paradox." *Journal of Economic Methodology* 19, no. 1: 43–62.

Rendall, Matthew. 2019. "Discounting, Climate Change, and the Ecological Fallacy." *Ethics* 129: 441–63.

Rollins, Kimberly and Audrey Lyke. 1998. "The Case for Diminishing Marginal Existence Values" Journal of Environmental Economics and Management 36: 324-344.

Rolston, Holmes, III. 2006. "Intrinsic Values on Earth: Nature and Nations." In *Environmental Ethics and International Policy*, edited by Henk A.M.J ten Have. Paris: UNESCO Publishing.

Roth, Alvin E., Tayfun Sonmez, and M. Utku Unver. 2004. "Kidney Exchange." *The Quarterly Journal of Economics* 119, no. 2: 457–88.

Ryan, Anthony M, and Clive L Spash. 2011. "Is WTP an Attitudinal Measure? Empirical Analysis of the Psychological Explanation for Contingent Values." *Journal of Economic Psychology* 32, no. 5: 674–87.

Sagoff, Mark. 1994. "Four Dogmas of Environmental Economics." *Environmental Values* 3, no. 4: 285–310.

———. 2004. *Price, Principle, and the Environment*. Cambridge: Cambridge University Press.

———. 2007. *The Economy of the Earth*. Cambridge: Cambridge University Press.

———. 2010. "The Poverty of Economic Reasoning About Climate Change." *Philosophy and Public Policy Quarterly* 30 (3/4): 8–15.

———. 2011. "The Quantification and Valuation of Ecosystem Services." *Ecological Economics* 70 (3): 497–502.

Samuelson, Paul A. 1954. "The Pure Theory of Public Expenditure." *The Review of Economics and Statistics* 36, no. 4: 387–89.

Samuelson, William, and Richard Zeckhauser. 1988. "Status Quo Bias in Decision Making." *Journal of Risk and Uncertainty* 1: 7-59.

Samuelsson, Lars. 2010. "Reasons and Values in Environmental Ethics." *Environmental Values* 19, no. 4: 517–35.

Sandel, Michael. 2012. *What Money Can't Buy*. London: Penguin Books.

Sandmo, Agnar. 1980. "Anomaly and Stability in the Theory of Externalities." *The Quarterly Journal of Economics* 94, no. 4: 799-807.

Sarkar, Sahotra. 2012. *Environmental Philosophy: From Theory to Practice*. Malden MA: Wiley-Blackwell.

Satz, Debra. 2012. *Why Some Things Should Not Be for Sale*. Oxford: Oxford University Press.

Schabas, Margaret. 1990. *A World Ruled by Number: William Stanley Jevons and the Rise of Mathematical Economics.* Princeton: Princeton University Press.

———. 2005. *The Natural Origins of Economics.* Chicago: University of Chicago Press.

Schelling, Thomas C. 1968. "The Life You Save May Be Your Own." In *Problems in Public Expenditure Analysis,* edited by Samuel B. Chase Jr. Washington DC: Brookings Institute.

———. 1997. "The Cost of Combating Global Warming: Facing the Tradeoffs." *Foreign Affairs* 76, no. 6: 8.

Searle, John. 2001. *Rationality in Action*. Cambridge: MIT Press.

Sen, Amartya K. 1977. "Rational Fools: A Critique of the Behavioral Foundations of Economic Theory." *Philosophy & Public Affairs* 6, no. 4: 317–44.

———. 1995. "Environmental Evaluation and Social Choice: Contingent Valuation and the Market Analogy." *The Japanese Economic Review* 46, no. 1: 23–37.

———. 2002. *Rationality and Freedom*. Cambridge: Harvard University Press.

Sklar, Lawrence. 2000. *Theory and Truth: Philosophical Critique within Foundational Science*, Oxford: Oxford University Press.

Smith, V. Kerry. 1993. "Nonmarket Valuation of Environmental Resources: An Interpretive Appraisal." *Land Economics* 69, no. 1: 1–26.

Spash, Clive L. 2008. "How Much Is That Ecosystem in the Window? The One with the Bio-Diverse Tail." *Environmental Values* 17, no. 2: 259–84.

Staaf, Robert J., and Francis X. Tannian, eds. 1973. *Externalities: Theoretical Dimensions of Political Economy*. New York: Dunellen.

Stavins, Robert N. 2011. "The Problem of the Commons: Still Unsettled After 100 Years." *American Economic Review* 101, no. 1: 81–108.

Steel, Daniel. 2015. *Philosophy and the Precautionary Principle: Science, Evidence, and Environmental Policy*. Cambridge: Cambridge University Press.

Stern, Nicholas. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge: Cambridge University Press.

———. 2014a. "Ethics, Equity and the Economics of Climate Change Paper 1: Science and Philosophy." *Economics and Philosophy* 30, no. 3: 397–444.

———. 2014b. "Ethics, Equity and the Economics of Climate Change Paper 2: Economics and Politics." *Economics and Philosophy* 30, no. 3: 445–501.

Stroud, Sarah. 2014. "Weakness of Will." In *Stanford Encyclopedia of Philosophy,* edited by Edward N. Zalta. Accessed on November 19, 2018. https://plato.stanford.edu/archives/spr2014/entries/weakness-will/.

Sugden, Robert. 2002. "Credible Worlds: The Status of Theoretical Models in Economics." In *Fact and Fiction in Economics: Models, Realism and Social Construction*, edited by Uskali Maki, 107-136. Cambridge: Cambridge University Press.

Sumner, Jenny, Lori Bird, and Hillary Dobos. 2011. "Carbon Taxes: A Review of Experience and Policy Design Considerations." *Climate Policy* 11, no. 2: 922–43.

Sumner, Larry. W. 1996. *Welfare, Happiness, and Ethics.* Oxford: Clarendon Press.

Sunstein, Cass R. 2006. "Irreversible and Catastrophic." *Cornell Law Review* 91: 841–98.

Tappolet, Christine. 2003. "Emotions and the Intelligibility of Akratic Action." In *Weakness of Will and Practical Irrationality*, edited by Sarah Stroud and Christine Tappolet, 97-120. Oxford: Clarendon Press.

Tarsney, Christian. 2017. "Does a Discount Rate Measure the Costs of Climate Change?" *Economics and Philosophy* 33, no. 3: 337–65.

Tenenbaum, Sergio, and Diana Raffman. 2012. "Vague Projects and the Puzzle of the Self-Torturer." *Ethics* 123, no. 1: 86–112.

Thaler, Richard H., and Shlomo Benartzi. 2004. "Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving." *Journal of Political Economy* 112, no. 1/2: S164-S187.

Thaler, Richard H., and Cass R. Sunstein. 2003. "Libertarian Paternalism." *American Economic Review* 93, no. 2: 175–79.

———. 2008. *Nudge*. London: Penguin Books.

Tietenberg, Thomas H, and Lynne Lewis. 2009. *Environmental and Natural Resource Economics*. London: Routledge.

Tversky, Amos. 1969. "Intransitivity of Preference." *Psychological Review* 84: 31-48.

Varian, Hal R. 1994. "A Solution to the Problem of Externalities When Agents are Well-Informed." *The American Economic Review* 84, no. 5: 1278-1293.

Venn, Tyron J, and John Quiggin. 2007. "Accommodating Indigenous Cultural Heritage Values in Resource Assessment: Cape York Peninsula and the Murray–Darling Basin, Australia." *Ecological Economics* 61, no. 2-3: 334–44.

Voorhoeve, Alex, and Ken Binmore. 2006. "Transitivity, the Sorities Paradox, and Similarity-Based Decision-Making." *Erkenntnis* 64: 101–14.

Wang, Tuo, R. Venkatesh, and Rabikar Chatterjee. 2007. "Reservation Price as a Range: An Incentive-Compatible Measurement Approach." *Journal of Marketing Research* XLIV: 200-213.

Weitzman, Martin L. 1974. "Prices vs. Quantities." *The Review of Economic Studies* 41, no. 4: 477-91.

———. 2009. "On Modeling and Interpreting the Economics of Catastrophic Climate Change." *Review of Economics and Statistics* 91, no. 1: 1–19.

———. 2013. "A Precautionary Tale of Uncertain Tail Fattening." *Environmental and Resource Economics* 55, no. 2: 159–73.

Wellisz, Stanislaw. 1964. "On External Diseconomies and the Government-Assisted Invisible Hand." *Economica* 31, no. 124: 345-362.

Westra, Laura. 2000. "The Disvalue of 'Contingent Valuation' and the Problem of the 'Expectation Gap.'" *Environmental Values* 9: 153–71.

Woodward, James F. 2002. "What Is a Mechanism? A Counterfactual Account." *Philosophy of Science* 69: S366-S377.

———. 2011. "Data and Phenomena: A Restatement and Defense." *Synthese* 182, no. 1. Springer: 165–79.

World Bank and Ecofys. 2018. *State and Trends of Carbon Pricing 2018*. Washington, DC: World Bank. Accessed on April 10, 2019. https://openknowledge.worldbank.org/handle/10986/29687.