ESSAYS ON ISSUES IN CLIMATE CHANGE POLICY

Marc Daube

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of
PhD (Economics)
at the
University of St Andrews

October 2016

Declaration

1. Candidate's declarations:

I, Marc Daube, hereby certify that this thesis, which is approximately **50,000** words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student in September, 2012 and as a candidate for the degree of Doctor of Philosophy in September, 2012; the higher study for which this is a record was carried out in the University of St Andrews between 2012 and 2016.

Date: .... Signature of Candidate: ....

2. Supervisor's declaration:

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: .... Signature of Supervisor: ....

3. Permission for publication: *(to be signed by both candidate and supervisor)*

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

PRINTED COPY

a) Embargo on all or part of print copy for a period of **three** years (maximum five) on the following ground(s):
- Publication would preclude future publication

Supporting statement for printed embargo request if greater than 2 years:
I request an embargo under the above reason because I wish to publish future papers based on the work in this thesis. This embargo would avoid work being rejected on the basis that it is already published in the thesis.

ELECTRONIC COPY

a) Embargo on all or part of print copy for a period of **three** years (maximum five) on the following ground(s):
- Publication would preclude future publication

Supporting statement for electronic embargo request if greater than 2 years:
I request an embargo under the above reason because I wish to publish future papers based on the work in this thesis. This embargo would avoid work being rejected on the basis that it is already published in the thesis.

ABSTRACT AND TITLE EMBARGOES

An embargo on the full text copy of your thesis in the electronic and printed formats will be granted automatically in the first instance. This embargo includes the abstract and title except that the title will be used in the graduation booklet.

If you have selected an embargo option indicate below if you wish to allow the thesis abstract and/or title to be published. If you do not complete the section below the title and abstract will remain embargoed along with the text of the thesis.

a)	I agree to the title and abstract being published	YES
b)	I require an embargo on abstract	NO
c)	I require an embargo on title	NO

Date **15 May 2017** signature of candidate _______________ signature of supervisor _______________

Please note initial embargos can be requested for a maximum of five years. An embargo on a thesis submitted to the Faculty of Science or Medicine is rarely granted for more than two years in the first instance, without good justification. The Library will not lift an embargo before confirming with the student and supervisor that they do not intend to request a continuation. In the absence of an agreed response from both student and supervisor, the Head of School will be consulted. Please note that the total period of an embargo, including any continuation, is not expected to exceed ten years.
Where part of a thesis is to be embargoed, please specify the part and the reason.

Acknowledgements

My greatest thanks and appreciation goes to my family, and in particular my mother, who has been a friend, guide and inspiration my entire life.

I would like to express special thanks to my supervisor, Professor David Ulph. Without his continuous support and advice this thesis would not have been possible.

I would also like to thank the academic staff at the School of Economics & Finance, University of St Andrews, for providing useful feedback during the annual PhD Conference, and the various individuals who provided feedback following presentations at the WCERE 2014 conference in Istanbul (in particular my discussant Prof. Stefan Baumgärtner), the EAERE 2015 conference in Helsinki, and the EAERE 2016 conference in Zurich (in particular my discussant Dr. Gerard van der Meijden). In addition I would like to thank the support staff at the School of Economics & Finance, University of St Andrews, and in particular Caroline Moore, Angela Hodge and Eliana Wilson, for all their support and kindness over the years.

Finally, my gratitude goes to the Economic and Social Resource Council who funded most of my research. Furthermore, I thank the School of Economics & Finance, University of St Andrews for providing financial support for the last nine months of my PhD.

Abstract

This thesis addresses three themes relating to climate change. The first is which types of fossil fuel to leave in the ground when they can differ in both their extraction cost and emissions rate. The analysis shows that without resource constraints there will always be use of at least one fossil fuel in the steady-state. With exhaustion constraints, any fossil fuel that has a lower extraction cost than the marginal cost of the backstop will be extracted in finite time regardless of the emissions rate. The only environmental consideration is the timing of extraction rather than leaving fossil fuel stock in the ground forever. The second theme is how altruistic concern of individuals for the well-being of others influences the socially optimal consumption levels and optimal emissions tax in a global context. If individuals have altruistic concern but believe that their consumption is negligible, they will not change their behaviour. However, non-cooperative governments maximising domestic welfare will internalise some of the damage inflicted on other countries depending on the level of altruistic concern individuals have and the cooperative optimum also changes as altruism leads individuals to effectively experience damage in other countries as well as the direct damage to them. Still, for behaviour to change, individuals need to make their decisions in a different way. The third chapter develops a new theory of moral behaviour whereby individuals balance the cost of not acting in their own self-interest against the hypothetical moral value of adopting a Kantian form of behaviour, asking what would happen if everyone else acted in the same way as they did. If individuals behave this way, then altruism matters and it may induce individuals to cut back their consumption. But nevertheless the optimal environmental tax is exactly the same as the standard Pigovian tax.

Contents

Introduction

The Climate Change Challenge

Climate change is one of the most complex problems facing our civilisation today. As Brunner et al. (2012) describe it, the climate change challenge is characterised by ”deep uncertainties, many interdependencies and complex social dynamics” (p. 260). One of the central aspects of the climate change problem is that is it is both a global intra-temporal externality and an inter-temporal externality.[1] In addition, the activities leading to climate change (mostly the production of energy) are central to modern economic life. Energy production is not a sector that can be replaced by some other activity and so governments have to find ways that reduce emissions but don't impact economic activity and the competitiveness of the economy too greatly.

The fact that climate change is a global intra-temporal externality presents a problem because there is no global government. If all countries were relatively similar in terms of the ratio of how much they cause the externality and how much they suffer from it, this would be less of a problem. But there are large asymmetries between countries. Some countries contribute a lot to the externality but are unlikely to suffer significantly from it, while other countries contribute virtually nothing to the externality but can expect to suffer a lot. This means there are very varied incentives and abilities of different countries to cut back on greenhouse gas (GHG) emissions. It is a widely held view that in order to control the GHG emissions externality, binding international agreements are necessary. However, developing such agreements has proved very difficult in the past. Countries face not just uncertainty over costs and benefits of such agreements but also uncertainty about whether other parties will keep their side of the promise.[2]

The second problem is that climate change is an inter-temporal externality. Since it is

[1] IPCC (2014) provides an excellent overview of the inter- and intra-generational issues associated with the climate change challenge.

[2] The global aspect of the climate change challenge will be discussed in more detail in Chapter 2.

the cumulative stock of GHG emissions that creates the damage, any harm suffered today is the consequence of action taken in the past and similarly action taken today will only lead to harm at some point in the future, potentially only affecting people generations away. Therefore, those people who will suffer the most from today's emissions have no role in current policy making. This naturally leads to the question of how policy actions (such as emission taxes) should vary over time. The use of the appropriate discount rates has been a long-standing issue in the economic analysis of the impact of climate change and the cost associated with mitigation (for example Nordhaus (2007), Stern (2008), Dasgupta (2008), Smith (2010), IPCC (2014)) This also relates to the debate about the trade-off between inter-generational welfare and intra-generational welfare.

Another issue is one of uncertainty and learning. We may not fully understand the potential damage caused by GHG emissions at the moment but we we will learn more about it as time passes. The question is therefore whether we should act now or delay some potentially tough action until we know more. In addition, there is a big issue of commitment in climate change policy. Current government are not able to credibly tie the hands of future governments. However, this leads to major uncertainty over future policy which is a particularly significant problem since actions to reduce GHG emissions require large-scale lump-sum investments that will only yield return over long time horizons.[3]

Thesis Outline

This thesis addresses three themes relating to the greenhouse gas emissions externality leading to climate change. These are optimal fossil fuel extraction, global environmental taxes and moral behaviour.

Chapter 1 analyses which types of fossil fuel to leave in the ground under optimal emissions taxation. It builds on the previous literature on optimal resource extraction with an emissions externality, developing a model of multiple fossil fuels that can differ in both their extraction cost and emissions rate. The aim is to demonstrate under which conditions with positive emissions decay certain fossil fuels may not be extracted at all or will be extracted regardless of their emissions rate. The analysis shows that without resource constraints there will always be use of at least one fossil fuel in the steady-state. With exhaustion constraints on all fossil fuels any fossil fuel that has a lower extraction cost than the marginal cost of the backstop will be extracted in finite time regardless of

[3] See for example Brunner et al. (2012), Ulph and Ulph (2013).

their emissions rate. The only environmental consideration is the timing of extraction rather than whether some of the fossil fuel stock should be left in the ground forever.

Chapter 2 then looks at how altruistic concern for the well-being of others influences the socially optimal consumption level and optimal emissions tax in a global context. A global externality like the GHG emissions leading to climate change are due to both free-riding at the individual level as well as the government level. A government aiming to maximise domestic social welfare may make individuals internalise the damage within their country, but will free-ride on the damage caused to other countries. Only a global cooperative solution could internalise global damage entirely. If individuals have altruistic concern for others but continue to believe that their consumption is negligible relative to the total, they will not change their behaviour. However, this paper shows that in a multi-country setting the global equilibrium levels of consumption for both the non-cooperative and cooperative solutions are affected by altruism. The key results are (a) that non-cooperative governments maximising domestic welfare will internalise some of the damage inflicted on other countries depending on the level of altruistic concern individuals have, but (b) the cooperative global optimum also changes as altruism leads individuals to effectively experience damage in other countries as well as the direct damage to them. Since altruistic concern for others may vary across countries, global welfare then becomes a function of the relative levels of altruistic concern between countries.

Chapter 3 develops a theory of alternative moral behaviour and how this may affect consumption choices and the optimal emissions tax in a single country context. Free-riding is often associated with self-interested behaviour. However, if there is a global pollutant, free-riding will arise if individuals calculate that their emissions are negligible relative to the total, so total emissions and hence any damage that they and others suffer will be unaffected by whatever consumption choice they make. In this context consumer behaviour and the optimal environmental tax are independent of the degree of altruism. For behaviour to change, individuals need to make their decisions in a different way. The model developed in Chapter 3 is based on Daube and Ulph (2016) and proposes a new theory of moral behaviour whereby individuals recognise that they will be worse off by not acting in their own self-interest, and balance this cost off against the hypothetical moral value of adopting a Kantian form of behaviour, that is by calculating the consequences of their action by asking what would happen if everyone else acted in the same way as they did. The analysis shows that: (a) if individuals behave this way, then altruism can matter and, depending on the choice function, the greater the degree of altruism the more individuals cut back their consumption of a 'dirty' good; (b) nevertheless the optimal environmental tax is exactly the same as that emerging from classical analysis where individuals act in self-interested fashion.

The first part of Chapter 3 is based on the published paper (Sections 3.1 to 3.3.3), but contains a more extensive literature survey which also gives an overview of the literature on social norms and crowding of intrinsic motivation (Section 3.2.2 and Section 3.2.3). The second part of Chapter 3 sets out some significant extensions to the basic model that I have undertaken (Sections 3.3.4 to 3.6). These include: a generalised (non-linear) choice function; the case of multiple dirty goods; as well as the case of heterogeneous preferences. Finally, Chapter 3 also develops a model where individuals may exhibit a desire for conformity in addition to their propensity to act morally (Section 3.7).

This thesis does not contain a concluding chapter. Because the topics analysed in each of the three chapters are quite different in nature it is not effective to attempt to compare the results of the individual chapters within the context of this thesis and develop ideas for future research. Instead each of the three chapters has its own concluding remarks where the main findings of each chapter are discussed and suggestions for future research are developed.

To give some background information the remainder of the Introduction will provide a brief overview of the key concepts in traditional environmental economics and behavioural economics. While Chapter 1 is based on traditional approaches in the analysis of natural resource and environmental economics, Chapter 2 and Chapter 3 are based on concepts of pro-social behaviour in behavioural economics.

Overview of Environmental Economics

According to Perman et al. (2003), there are three main themes in the analysis of natural resources and environmental issues. These are efficiency, optimality and sustainability. Economics concerns itself particularly with allocative efficiency. An efficient allocation of resources ensures that the net benefit to all of society is maximised and resources are not wasted on inefficient activities. The concept of social optimality is achieved when a resource allocation maximises the overall objective of society. Efficient resource allocation is a necessary condition for optimality but it is possible that an efficient resource allocation is not socially optimal. The third concept is that of sustainability which means the use of resources has to take account of posterity. Some economists may argue that social optimality would internalise any need for sustainability but as Perman et al. (2003) argue, optimality as it is pursued by economics does not always take into account posterity to an adequate degree. The idea of sustainability is usually driven by moralist views which dictate that sustainability is a moral obligation regardless of economic optimality.

Market failure is a central problem in environmental economics. As Hanley et al. (2013) describe it, market failure for environmental goods and resources means that "benefits and costs cannot be allocated with precision across and within nations and generations" (p. 15). This usually arises because property rights over environmental resources are not well defined or cannot be transferred. Market failure in environmental issues generally occurs due to the public good problem as well as externalities, open-access property, and hidden information.[4] GHG emissions leading to climate change is a classic example of a negative externality problem as the emitters do not carry all the costs of the pollution. The approach to dealing with such market failures can vary widely. In some instances regulation is used (a command-and-control approach) where a regulator conducts a cost-benefit analysis of the environmental problem and enforce these regulations through fines. Other approaches aim to correct market failure through market-based instruments such as taxes or tradable permits. A tax is the standard method to correct for a negative externality in economic theory, generally known as a Pigovian tax.[5] This tax increases the private marginal cost of the activity that creates the externality to the level of of the social marginal cost. If the tax is set to the correct level, this reduces the level of activity to the socially optimal level. However, because it is very difficult to measure the social cost of a negative externality such as pollution, setting the level of tax is subject to a lot of uncertainty. Another approach to dealing with an externality are tradable emissions permits (as used in the EU-ETS). These cap the total amount of emissions within the system, but the price of each unit of emissions is determined through a market for permits. In theory this guarantees an efficient allocation of emissions within the system given the constraint on total emissions. One downside of this approach is, however, that a central regulator still has to determine the cap on total emissions, which is a non-market intervention and can therefore lead to an inefficient or sub-optimal results. Due to the difficulties with valuation, measurement and implementation when it comes to standard economic instruments, interest in the field of behavioural economics has increased significantly in recent years as a way to find methods and policies that could potentially enable

[4] Hanley et al. (2013)

[5] This concept was first developed by Pigou (1932).

us to deal with environmental problems without only relying on extrinsic incentives.

Environmental policy is also often subject to strong ideological views. As Frey (1999) summarises, there are two main approaches to thinking about environmental issues. On one hand, there is the moralist view which is mostly a normative stance on the environment. The moralist views nature with unique value and believes that humans should protect the environment for ethical reasons and because they have the capability to so. This means that a country should reduce GHG emissions even if no other country does so and if it is not in the best economic interest of the country to do so. In addition, those who do not have the same moral views must be forced to protect the environment nevertheless, with little or no regard to any trade-offs. On the other hand is the rationalist or utilitarian viewpoint which is more in line with the approach of neoclassical economics already described. Over time there has been more and more recognition by both the moralist and rationalist camps that an appreciation and understanding of both extrinsic and intrinsic motivation is needed in order to design effective environmental policy. It is therefore important to understand how behavioural economics can complement the traditional way of thinking about externalities.

Overview of Behavioural Economics

The field of behavioural economics has become increasingly popular as it aims to explain observations that deviate from predictions of traditional economic theory and furthermore intends to extend the classic framework of economic theory to account for the observed behavioural failures. The field of behavioural economics started to make significant advances when psychologists such as Kahneman and Tversky began comparing their cognitive models of decision making to economic models in the 1960s.[6]

There are four key themes in behavioural economics. These are used in a range of issues relating to energy and environmental policy (see for example Cabinet Office Behavioural Insights Team (2011), Pollitt and Shaorshadze (2011), IPCC (2014) for an overview of the various policy issues). First, Prospect Theory implies the importance of reference points in assessing welfare changes and is used to explain observed phenomena such as loss aversion (individuals value losses more than gains), the endowment effect (individuals place additional value on goods they already own), and the status-quo bias (individuals

[6] For overviews of behavioural economics see for example Tirole (2002), Camerer et al. (2004), Sobel (2005), Bernheim and Rangel (2007)

tend to stay with default options chosen for them).[7] The second theme is the principle of time-varying discount rates. While neoclassical economics assumes that individuals maximise their utility with exponential discounting, a number of studies have provided evidence that hyperbolic discounting - higher discount rates for short time periods and lower discount rates for longer time horizons - is more applicable in practice.[8] Third is the notion of bounded rationality. This refers to the idea that individuals have cognitive constraints that render their decision making ability sub-optimal compared to the homo economicus assumed by neoclassical theory. Observed phenomena that are explained by bounded rationality include choice overload (individuals have difficulty making a choice when there are too many options), heuristics (shortcuts to decision making), and the failure to assess statistical probabilities (individuals often base their decisions on 'vivid and salient information' rather than actual probabilities).[9]

The final key theme in behavioural economics is prosocial behaviour (see for example Kahneman et al. (1986) for an early contribution in this area). Prosocial behaviour is the key theme of Chapter 3 and various aspects of it will be discussed in more detail in that chapter. Behind this concept is the observation that individuals appear to often not just maximise their consumption and monetary payoff, but act in specifically prosocial ways. The provision of public goods through voluntary contributions is a particularly important concept in the area of behavioural-environmental economics as it explores whether public goods (such as clean air) can be provided through an individual's sense of prosocial behaviour rather than through a centralised mechanism. This is also important in the area of climate change where it is very difficult to establish the extrinsic incentives to internalise the GHG emissions externality due to its global nature. In addition, as will be discussed in more detail in Chapter 3, there is some evidence to suggest that the existence of extrinsic incentives can crowd out or crowd in the intrinsic motivation for prosocial behaviour individuals may have.

[7] Kahneman and Tversky (1979), Thaler (1980), Samuelson and Zeckhauser (1988) Shogren and Taylor (2008), among others.

[8] For example Thaler (1981), Benzion et al. (1989), Holcomb and Nelson (1992), Laibson (1997), Camerer et al. (2004).

[9] Contributions on bounded rationality include Simon (1986),Camerer et al. (2004), Thaler (1999), among others.

Chapter 1

An Analysis of Which Types of Fossil Fuel to Leave in the Ground

1.1 Introduction

The Paris Agreement negotiated at the 2015 United Nations Climate Change conference is regarded as one of the most important steps in global efforts against climate change. It was agreed by 195 countries and represents a framework for the reduction in carbon dioxide emissions and a global transition from fossil fuels to renewable energy. However, the agreement leaves room for use of fossil fuels in the long term as long as corresponding emissions are in line with natural absorption rates or balanced by carbon capture and storage technologies.[1] This raises the question if it is optimal to leave some fossil fuels in the ground forever, and if so under which conditions, or if we should extract it all. The topic is of course not a new one, and has already received a lot of attention in natural resource economics. Analysis of optimal use of non-renewable resources dates back to Hotelling (1931). The Hotelling rule states that the marginal net rent - the market price net of marginal extraction cost - of an exhaustible resource has to grow at the same rate as the interest rate which is assumed to be exogenous and constant. This is a result of the arbitrage condition which dictates that that the return from not extracting a unit of the resource has to equal the return of extracting it. Based on Hotelling's insights, a major step in exhaustible resource economics was the development of the Dasgupta-Heal-Solow-Stiglitz (DHSS) model, which is a result of the three seminal articles Dasgupta and Heal (1974), Solow (1974) and Stiglitz (1974). The DHSS model describes an economy with

[1] https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf

two assets, man-made capital and a non-renewable resource stock.[2]

Although at the time not related to the climate change issue, D'Arge and Kogiku (1973) made the case for combining the analysis of the exhaustible resources problem with the pollution problem and famously asked the question: "which should we run out first, air to breathe, or fossil fuels to pollute the air we breathe?" (p. 68). Early contributions in this includes Schulze (1974) who develops a finite horizon model to determine optimal extraction with a stock externality and looks at cases of degrading resource quality, recycling of the resource and the impact of technological change on resource prices in the long-run. Hoel (1978) models optimal extractions paths under a variety of conditions with a stock externality effect where "harmful residuals" (p. 222) from resource extraction cause damage. In addition, the individual can consume both the resource and a recycled version of the resource which contributes less to the emissions stock. While his analysis doesn't explicitly look at the optimal path of a tax to internalise the externality, the model setup is very similar to the later models used to determine the optimal carbon tax path and also used in this chapter. Forster (1980) uses a finite horizon approach to develop a stock externality model of optimal resource use which includes the possibility of "antipollution activities" (p. 326) that in turn require use of the resource. He finds that such activities will not be used and while initial consumption of the resource should be reduced, it then should increase over time.

Analysis of the classic exhaustible resource problem had established that a constant tax on a costless exhaustible resource will not change the depletion of the resource (e.g. Dasgupta and Heal 1979). This in turn would imply that a constant tax, regardless of the level, would have no effect on the emissions stock. This suggests that the time-path of the tax rather than the level is the central aspect to reducing the amounts of the resource extracted and the corresponding emissions. As the understanding of the impact of carbon dioxide emissions generated from burning fossil fuels improved (see for example Nordhaus 1994), this was reflected in the literature on optimal resource extraction and a focus on the optimal time path of a carbon tax developed. A fundamental model combining the exhaustible resource problem with the negative emissions externality was developed by Sinclair (1992) and argued that since the question for resource owners is mainly about the timing of extraction rather than the total amount to be extracted, the aim of an emissions tax has to be to delay some extraction to the future. His analysis of an optimal carbon tax suggested that the tax should be falling over time. While expectation of a falling

[2] Benchekroun and Withagen (2011) provide a closed form solution to the DHSS model which characterises a path of all variables in the model from all possible inital values.

carbon tax would incentivise producer to delay extraction, a rising carbon tax would achieve the opposite. Another early model in this area is Withagen (1994) who compares the optimal time at which to exhaust a resource with and without a stock emissions externality. Following this, Ulph and Ulph (1994) developed a model that uses a convex damage function to reflect the idea that the marginal damage increases as the stock of emissions in the atmosphere increases. Assuming linear-quadratic functional forms they find that the emissions tax should increase as the emissions stock increases and later decrease again as the rising scarcity rent increases the price of the resource. At the same time extraction rates decline and the emissions stock falls to natural levels. Hoel and Kverndokk (1996) build on this analysis by modelling extraction costs that increase as the stock of the remaining resource declines. This approach models "economic exhaustion (zero long-term Hotelling rent)" (p.116) rather than physical exhaustion of the resource which is in line with the approach used by Heal (1976). The authors also model the effect of a backstop technology, a perfect substitute to the fossil fuel resource that is available at constant marginal costs, infinite supply and causes no emissions externality. The main finding is that these factors do not influence the total amount of fossil fuel extraction, but will shift some of the resource extraction to a distant time period where the emissions stock is declining. Tahvonen (1997) develops these two models further and shows that for a model without extraction costs and without a backstop, the optimal emissions tax (and the time path of the emissions stock) follows an inverted U-shape form without having to assume any specific functional form for utility and damage. He then also analyses the case of a backstop technology and extraction costs that increase with resource depletion. This shows that due to the assumptions that a certain proportion of the emissions stock is decaying naturally, a sustained period of simultaneous consumption of both the fossil fuel and the backstop can occur.

All of the analysis mentioned above looks at the first-best solution where the emissions externality is internalised optimally through the Pigovian tax. However, among others, Sinn (2008) argued that it is not politically feasible to implement such a carbon tax, especially on a global scale. His analysis therefore focusses on ways to replicate the first-best outcome in the absence of the required carbon tax. This argument also puts the focus on the green paradox. This principle says that a rising carbon tax or subsidies for green technologies may actually increase current use of fossil fuels (and increase emissions) as resource owners anticipate the fall in demand over time. This could potentially lead to a lower level of overall welfare than if no tax or subsidy had been imposed.[3] Gerlagh (2011)

[3] Of course this is only possible if the instrument used is not the first-best Pigovian tax since this is by definition the social welfare maximising optimum.

defines two types of the green paradox. A 'weak' green paradox arises if the prospect of a cheaper clean backstop leads to an increase in current emissions. This, however, could still be welfare improving if future emissions decrease sufficiently. The 'strong' green paradox on the other hand implies that the cheaper backstop leads to an increase in cumulative damage of the emissions, evaluated at the NPV. As such it reduces total welfare. Gerlagh (2011) analyses a variety of cases of different resource and backstop cost structures, as well as an imperfect backstop. Based on this van der Ploeg and Withagen (2012a) also make an evaluation of the green paradox under different circumstances but first develop the first-best optimum as counterfactual. In that they use a similar setup to Tahvonen (1997) but assume no decay of the emissions stock. This leads to the result that it is never optimal to simultaneously use the fossil fuel resource and the backstop, and the optimal emissions tax will never decrease. They find that if the initial level of the emissions stock is low, there will be use of the fossil fuel resource only to start with but as the tax increases, there will eventually be a switch to the backstop where it will stay forever. Their approach also models extraction costs that are a function of the remaining fossil fuel stock.

The literature discussed so far generally models one exhaustible resource and, in some cases, a backstop technology. However, based on the Hotelling rule Herfindahl (1967) developed a partial equilibrium model of multiple resources which differ only in their marginal extraction cost. This resulted in the least cost first principle which states that extraction has to occur in order of cost, starting with the cheapest one. This has also been shown by Solow and Wan (1976). Since this analysis assumed that the resources were identical in everything but extraction cost, the literature has since made an effort to evaluate if the least cost first principle still holds under a variety conditions. For example, Kemp and Long (1980) show that the principle may not hold in a general equilibrium setting with Ricardian techniques of extraction, and if the marginal extraction cost is constant for each deposit of the resource in terms of a perfect substitute for the resource. Following this, Lewis (1982) shows that even in that framework the least cost first principle holds as long as the resource can be converted into capital and that "a sufficient condition for the strict sequencing of extraction to be optimal is that stored capital be productive so that it can be used to produce additional capital" (p. 1081). While these models characterise the difference between resources through extraction cost, Chakravorty and Krulce (1994) analyse the case where resources have quality differences which are characterised through heterogeneous demand for the resources. Specifically their model assumes that two resources are of the same quality for the generation of electricity, but differ in quality for the use in transport. They show that this would lead to simultaneous extraction of both resources for some period. Further work on the validity of the least cost first principle includes the case where the extraction capacity is constrained (Amigues et al. 1998; Holland 2003) and the existence of setup costs (Gaudet et al. 2001). Furthermore, assuming no decay of emissions, van der Ploeg and Withagen (2012b) develop a model of two fossil fuels where the dirtier one, coal, acts as a backstop. Using simulations they show that with that setup it is optimal to use more of the cleaner fossil fuel (oil in their model) and less coal. They develop the optimal carbon tax path to achieve this order where initially only oil is used, then both oil and coal are used simultaneously and finally only coal is used. They then add a clean renewable backstop to show that this impacts on transition times as well the optimal carbon tax. Chakravorty et al. (2008) use a model of two fossil fuels that have the same extraction costs but different emissions rates to show that the Herfindahl rule does not necessarily have to apply in their setup.

There are three main rationales for arguing that fossil fuels should be left in the ground. First, there are environmental concerns which predict that increasing carbon dioxide levels in the atmosphere could at some point lead to catastrophic damage. Second, there is the reasoning that the presence of a backstop technology will make fossil fuels redundant. And third, increasing extraction costs for fossil fuels may make it uneconomic to extract certain fossil fuels at some point. Most of the literature on exhaustible resources alongside a backstop technology models the extraction cost either as constant or as a function of the remaining resource stock, i.e. the lower the remaining stock the higher the extraction cost. At the same time, however, the emissions rate remains constant. The model developed in this chapter is based on previous models but aims to extend the analysis with regard to extraction costs and emissions rate by modelling multiple fossil fuels and a backstop technology. While the extraction cost of each of the fossil fuels is not dependent on the resource stock, the extraction costs as well as the emissions rate can vary across each fossil fuel. Furthermore, the model assumes that there is positive decay of emissions. This chapter will first analyse the case where none of the fossil fuels are exhaustible and then contrast this to the case with exhaustion constraints. The aim is not to explicitly solve for the optimal extraction path, but to demonstrate under which conditions with positive emissions decay certain fossil fuels may not be extracted at all or will be extracted regardless of their emissions rate. The analysis shows that without the exhaustion constraints the resulting steady-state has to involve a fossil fuel, it cannot be possible to end up with just the backstop.[4] Furthermore, if there are exhaustion

[4] This is assuming that there is at least one fossil fuel with marginal extraction cost lower than the marginal cost of the backstop.

constraint, any fossil fuel with lower extraction costs than the backstop costs will be exhausted in finite time regardless of the emissions rate.

The chapter is structured as follows. Section 1.2 develops the basic model setup and starts the analysis by assuming that fossil fuels can be supplied infinitely and therefore are not subject to exhaustion constraints. Section 1.3 builds on this and analyses the case where the fossil fuels are exhaustible. Furthermore, Section 1.3.3 will look at some of the dynamics in the case of exhaustion constraints and establish conditions for simultaneous consumption of multiple resources. Finally, Section 1.4 provides a brief discussion and concluding remarks.

1.2 Non-Exhaustible Fossil Fuels

1.2.1 Model Setup

In this section we develop the basic model setup and analyse the case where none of the fossil fuels are subject to an exhaustion constraint. Start by assuming that there are $1 \leq n < \infty$ different fossil fuels. The amount of each fossil fuel extracted in period t is denoted by x_t^i $(1 \leq i \leq n)$ and each fuel can be extracted at constant marginal cost c_i.The extraction cost may be the same for multiple fossil fuels but has to be different for at least some of the fuels. Use of the clean backstop technology is denoted z_t and it has a constant marginal cost γ. The fossil fuels and the backstop are perfect substitutes and therefore the flow rate of utility of consumption in period t is given by $u(\sum_{i=1}^{n} x_t^i + z_t)$. Furthermore we have $u(0) = 0$; $u'(.) > 0 \; \forall x^i, z \geq 0$; $u'(.) \to 0$ as $x^i, z \to \infty$; $u''(.) < 0$; and $0 < c_i, \gamma < u'(0)$. The stock of emissions in the atmosphere at time t is denoted M_t with the initial stock of emissions M_0 taken as given. The damage caused by the emissions stock in period t is captured by the damage function $D(M_t)$. This is a strictly increasing and convex function where we have $D(0) = 0$, $D'(0) = 0$ and $D'(M_t) > 0$, $D''(M_t) > 0$ $\forall \; M_t > 0$.[5] Since an increasing and convex damage function will approach infinite marginal damage as the emissions stock increases to very high levels, this representation is consistent with the idea that carbon dioxide levels in the atmosphere may reach a critical level at some point. Next, consumption of the fossil fuels add to the emissions stock at a rate of $\alpha_i > 0$ for each unit of the corresponding fossil fuel used.[6] The emissions

[5] This means that none of the damage is irreversible, and that the marginal damage for a zero emissions stock is neglibile.

[6] If any of the fossil fuels had a zero emissions rate then this would of course make it a backstop technology.

stock, however, also decays at a rate of $\delta > 0$. While it is allowed that different fossil fuels may either have the same extraction cost or the same emissions rate as another fossil fuel, it is assumed that none of the fossil fuels have exactly the same extraction cost and emissions rate as any other fossil fuel, as this would make those fuels identical for the purpose of this analysis. Finally, the rate of time preference is given by $r > 0$. Given this setup we can formulate the optimisation problem as[7]

$$\begin{aligned} Max \quad & \int_0^{\infty} e^{-rt}\Big[u(\sum_{i=1}^{n} x_t^i + z_t) - \sum_{i=1}^{n} c_i x_t^i - \gamma z_t - D(M_t)\Big] dt \\ \text{s.t.} \qquad & \dot{M}_t = \sum_{i=1}^{n} \alpha_i x_t^i - \delta M_t, \end{aligned} \tag{1.1}$$

where the constraint is the equation determining the time-path of the emissions stock. The emissions stock in period t increases at the rate of the aggregate use of all the fossil fuels multiplied by their emissions rate (i.e. the total emissions caused by fossil fuel use) minus the natural decay of the emissions stock.

1.2.2 Model Analysis

The current value Hamiltonian for the optimisation problem in (1.1) is given as

$$H_c = \Big[u(\sum_{i=1}^{n} x_t^i + z_t) - \sum_{i=1}^{n} c_i x_t^i - \gamma z_t - D(M_t)\Big] - \sigma_t\Big[\sum_{i=1}^{n} \alpha_i x_t^i - \delta M_t\Big], \tag{1.2}$$

where the parameter σ_t represents the shadow cost of the emissions stock constraint in period t. From here it is straightforward to determine that the corresponding first-order conditions are

$$u'(\sum_{i=1}^{n} x_t^i + z_t) \leq c_i + \alpha_i \sigma_t, \quad x_t \geq 0, \quad \forall\ 1 \leq i \leq n, \tag{1.3}$$

$$u'(\sum_{i=1}^{n} x_t^i + z_t) \leq \gamma, \quad z_t \geq 0, \tag{1.4}$$

$$\dot{\sigma}_t = (r + \delta)\sigma_t - D'(M_t), \tag{1.5}$$

[7] Dots over variables denote time derivatives.

$$\dot{M}_t = \sum_{i=1}^{n} \alpha_i x_t^i - \delta M_t, \tag{1.6}$$

$$\lim_{t \to \infty} e^{-rt} \sigma_t = 0, \tag{1.7}$$

where (1.3) and (1.4) hold with complementary slackness and (1.7) is the transversality condition. To begin the analysis, we see that (1.3) and (1.4) in combination with the complementary slackness condition ensure efficiency in consumption of the fossil fuels and the backstop. If consumption of any of the fossil fuels or the backstop is positive, then for that fuel the marginal benefit of consumption has to be equal to the marginal social cost. This includes both the marginal extraction cost as well as the effective marginal shadow cost of emissions σ_t, adjusted for the emissions rate. Note that we can interpret σ_t as the required emissions tax. Since the fossil fuels (and the backstop) are perfect substitutes it is also straightforward to see that for any two fuels f and g where $c_f \leq c_g$ and $\alpha_f \leq \alpha_g$ with one of the two holding with strict inequality, then fuel g will always be more expensive than fuel f and therefore never be used regardless of the emissions tax.[8] It is of course also easy to see from (1.3) and (1.4) that any fossil fuel with an extraction cost $c_i > \gamma$ will never be used at all as it is always more expensive compared to the backstop regardless of the emissions rate and emissions tax.

Next, differential equation (1.5) describes the development optimal emissions tax. It is straightforward to determine that $\dot{\sigma}_t > 0$ if $\sigma_t > \frac{D'(M_t)}{r+\delta}$ and vice versa. Furthermore, taking the second time derivative we also see that $\ddot{\sigma}_t > 0$ if $\dot{\sigma}_t > \dot{M}_t \frac{D''(M_t)}{r+\delta}$. As done in Hoel and Kverndokk (1996), we can use (1.5) and the transversality condition (1.7) to derive that the optimal emissions tax at time t is given by

$$\sigma_t = \int_t^\infty \left[e^{-(r+\delta)(\tau - t)} D'(M_\tau) \right] d\tau \geq 0. \tag{1.8}$$

The above shows that the optimal emissions tax is determined by the discounted future marginal damage caused by the emissions externality. This is of course what we would intuitively expect and will help us understand the development of the emissions tax in the analysis. Finally, differential equation (1.6) describes the time path of the emissions stock in the atmosphere which is simply the constraint from the maximisation problem

[8] If there are exhaustion constraints as in Section 1.3 then this fuel g will never be used before fuel f is exhausted.

re-stated.

Ignoring the backstop for the moment, suppose that with a given emissions tax σ_t a particular fossil fuel j is extracted (and none other) and therefore $u'(x_t^j) = c_j + \alpha_j \sigma_t$. Furthermore assume the emissions stock and emissions tax are increasing ($\dot{M}_t > 0$ and $\dot{\sigma}_t > 0$). This increases the marginal cost of fuel j and reduces consumption. As the tax continues increasing, it may at some point be optimal to switch to another fossil fuel k where $c_j \leq c_k$ but $\alpha_j \geq \alpha_k$. Although fossil fuel j has a lower extraction cost, the impact of the emissions rate on the total marginal cost through the tax offsets the extraction cost difference at some point. The tax level that induces a switch from one to the other is given by

$$\sigma_t^{jk} = \frac{c_k - c_j}{\alpha_j - \alpha_k}. \tag{1.9}$$

As we can see, the parity tax level is given by the difference in the extraction costs relative to the difference in emissions rates. This also confirms that fuel k has to have lower emissions rate but higher extraction cost for it to be the next in line with an increasing emissions tax. If we had $c_j \geq c_k$ but $\alpha_j \leq \alpha_k$ then fuel k would already have been more expensive than j at the prevailing tax rate. And similarly we have already established that if either fuel had both a lower extraction cost and emissions rate it would always be the preferable fuel at any level of the tax. Note also that from (1.9) we can see that $\dot{\sigma}_t^{kj} = 0$, and therefore we cannot have simultaneous consumption of multiple fossil fuels if the tax is either increasing or decreasing.[9]

If we order the fossil fuels according to their marginal costs at any time t given the corresponding emissions tax σ_t, we can establish the order in which the switching from one fossil fuel to the next will occur as the tax increases. However, at some point we will reach a point where rather than switching to the next fossil fuel, the tax will lead to marginal cost parity with the backstop technology. Suppose that the fuel just before the backstop is reached is fossil fuel k. Then the tax level that puts fossil fuel k and the backstop at marginal cost parity (i.e. $c_k + \alpha_k \sigma_t = \gamma$) is defined by

$$\sigma_t^{kz} = \frac{\gamma - c_k}{\alpha_k}. \tag{1.10}$$

[9] Note that it is possible that more than two fossil fuels are at parity at any single point in time, but as the tax continues to increase there will only be one fossil fuel that will be optimal to use as time continues.

Of course there is a tax level that puts any of the fossil fuels that have an extraction cost lower than the cost of the backstop at parity with the backstop, but these are of lesser importance as for any fuel other than k there will be another fossil fuel that is cheaper at that tax level. Only fossil fuel k transitions to consumption of the backstop.

So far we have looked at an increasing emissions tax and of course the reverse dynamics would be true for a decreasing emissions tax. Let us now see if there is a steady-state in the system and if so, what this looks like. In the steady-state there is no change over time in the emissions stock ($\dot{M}_t = 0$), emissions tax ($\dot{\sigma}_t = 0$), and extraction and consumption levels ($\dot{x}_t^i = 0 \ \forall \ 1 \leq i \leq n$, $\dot{\gamma}_t = 0$). Thus from (1.5) and (1.6) we get

$$\sigma_t = \frac{D'(M_t)}{r + \delta}, \tag{1.11}$$

and

$$\delta M_t = \sum_{i=1}^{n} \alpha_i x_t^i. \tag{1.12}$$

In order to draw this in the $\sigma_t - M_t$ space we need to express the extraction levels of the fossil fuels in (1.12) in terms of the emissions tax σ_t. The level of extraction is defined by the inverse of the utility function and therefore we know that, assuming only fossil fuel j is consumed at time t, we have $x_t^j = f(c_j + \alpha_j \sigma_t)$ given $x_t^k = 0 \ \forall \ k \neq j$. Thus we can write (1.12) as

$$M_t = \frac{\alpha_j}{\delta} f(c_j + \alpha_j \sigma_t) \qquad \text{if} \quad x_t^k = 0 \quad \forall \quad k \neq j. \tag{1.13}$$

This of course applies to any of the n fossil fuels consumed on their own at time t.[10] Furthermore, we also know that when $\gamma > 0$ and $x_t^i = 0 \ \forall \ 1 \leq i \leq n$, then we require $M_t = 0$ to have $\dot{M}_t = 0$. Note that, using the above example of fossil fuels j and k where $c_j \leq c_k$ and $\alpha_j \geq \alpha_k$, we still have $\frac{\alpha_j}{\delta} f(c_j + \alpha_j \sigma_t) > \frac{\alpha_k}{\delta} f(c_k + \alpha_k \sigma_t)$ if $c_j + \alpha_j \sigma_t^{jk} = c_k + \alpha_k \sigma_t^{jk}$. This means there is a range of consumption levels between those two points which can

[10] There can of course be consumption of more than one resource in the steady state as is demonstrated later in this section. The purpose of (1.13) is to develop a component in the full description of the conditions for a constant emissions stock as shown in (1.15) and illustrated in Figure 1.1.

involve both fossil fuels simultaneously and is consistent with $\dot{M}_t = 0$.

We can now describe the level of the emissions stock consistent with no change in the stock in terms of the emissions tax σ_t. Although this can be done for all n fuels and the backstop, for simplicity and illustration purposes we will describe this for the case of two fossil fuels, j and k, as well as the backstop where $c_j < c_k < \gamma$ and $\alpha_j > \alpha_k$. Furthermore, to ensure that both of these fossil fuels are used before the backstop comes into play we assume that M_0 is sufficiently low and that

$$\frac{\alpha_j}{\alpha_k} > \frac{\gamma - c_j}{\gamma - c_k}. \tag{1.14}$$

This condition ensures that the relative emissions rates and relative cost differences to the backstop for the two fossil fuels are such to ensure that, with an increasing emissions tax, it is fossil fuel k that is used just prior to the backstop coming into play. Given this, the consumption levels consistent with a constant emissions stock (i.e. $\dot{M}_t = 0$) for any emissions tax are defined by

$$M_t = \frac{1}{\delta}\begin{cases} \alpha_j f(c_j + \alpha_j \sigma_t) & \forall \quad \sigma_t < \frac{c_k - c_j}{\alpha_j - \alpha_k} \\ \alpha_k f(\frac{\alpha_j c_k - \alpha_k c_j}{\alpha_j - \alpha_k}) \leq \alpha_j x_t^j + \alpha_k x_t^k \leq \alpha_j f(\frac{\alpha_j c_k - \alpha_k c_j}{\alpha_j - \alpha_k}) & \forall \quad \sigma_t = \frac{c_k - c_j}{\alpha_j - \alpha_k} \\ \alpha_k f(c_k + \alpha_k \sigma_t) & \forall \quad \frac{c_k - c_j}{\alpha_j - \alpha_k} < \sigma_t < \frac{\gamma - c_k}{\alpha_k} \\ 0 \leq \alpha_k x_t^k \leq f(\gamma) & \forall \quad \sigma_t = \frac{\gamma - c_k}{\alpha_k} \\ 0 & \forall \quad \sigma_t > \frac{\gamma - c_k}{\alpha_k} \end{cases}. \tag{1.15}$$

In Figure 1.1 the red line shows equation (1.11) and the blue line shows equation (1.15).[11] Note that the blue line extends down all the way to the x-axis. If we order the fuels in order of their extraction costs, $c_1 \leq c_2 \leq .... \leq c_n$, and assume $\sigma_t = 0$, then there will be some level of the emissions stock consistent with steady-state consumption of the cheapest fuel, i.e. fuel 1 with cost c_1.

Next we see that the black bold line depicts the saddle path for the tax rate that leads to the steady-state level of the emissions stock $\hat{M}$. At this point the tax on emissions also remains constant forever and therefore $\hat{\sigma}_t = \sigma_t^{kz}$. Fossil fuel k and the backstop are

[11] For n fossil fuels there is a curved and flat segment for each fuel that has a lower marginal cost than the backstop for any $\sigma_t < \hat{\sigma}$.

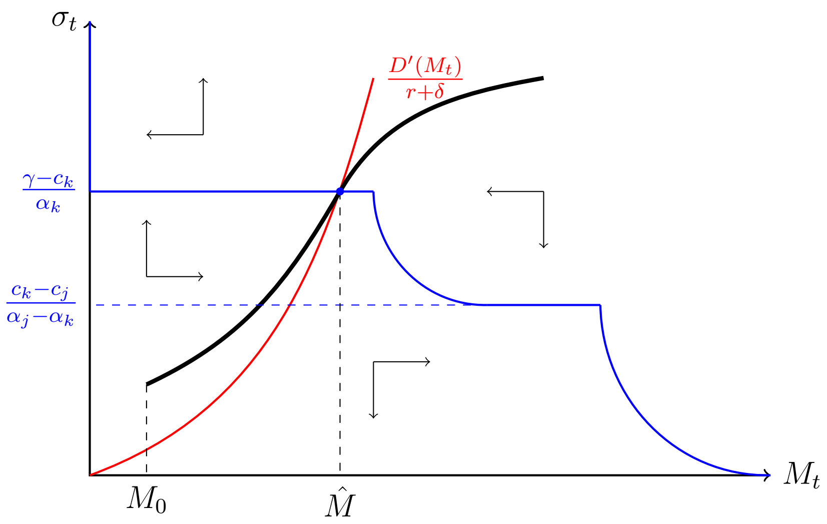

Figure 1.1: Steady-State Emissions Stock and Tax with Non-Exhaustible Fossil Fuels

consumed simultaneously forever (at constant levels), but there is no consumption of any other fossil fuel. It is assumed here that the cost of the backstop relative to the cost of the fossil fuels as well as the damage function are such that (1.11) and (1.15) intersect at a point somewhere on the horizontal part of the blue line and therefore there is consumption of the backstop in the steady-state. The further the intersection moves to the right along the horizontal part the lower the share of the backstop in overall consumption. If the intersection point were at the very right edge of the horizontal part there would be no consumption of the backstop.[12] And if the intersection were somewhere along the curved segments, the steady-state would involve only the corresponding fossil fuel at a level that just offsets the natural decay in each period.[13]

As depicted in Figure 1.1 let us assume we have a low initial level of the emissions stock, $M_0 < \hat{M}$ and initial tax level $\sigma_0 > 0$. Then at time t=0 we start with fossil fuel j and an increasing tax and emissions stock. At that point we have $c_j + \alpha_j \sigma_0 < \gamma$ and $c_j + \alpha_j \sigma_0 < c_i + \alpha_i \sigma_0 \; \forall \; 1 \leq i \leq n, i \neq j$. Furthermore, it also means that regardless of the extraction cost, any fuel h with an emissions rate higher than fuel j that is not cheaper than fuel j at $t = 0$ will never be extracted at all.[14] This means that depending on the

[12] Only if the intersection were all the way at the y-axis (i.e. at $\hat{M} = 0$), would the steady-state use only the backstop. However, given that we have $D'(0) = 0$, this is not possible.

[13] It is also noteworty that, as explained earlier and indicated by the second line in (1.15), if the steady-state tax happens to be equal to a parity tax level between two fossil fuels (e.g. $\hat{\sigma} = \sigma_t^{kj}$), then the steady-state may involve simultaneous consumption of two fossil fuels forever. There is a range of emissions stocks consistent with such a steady state (the horizontal segment between the two curves), consisting of different combinations of the two fossil fuels.

[14] Also any fuel with both a higher extraction cost and a higher emissions rate relative to another fossil

initial level of the emissions stock, it is possible that there may be a whole range of the dirtiest fossil fuels that would never be used at all.

Proposition 1 *If there are $n \geq 1$ non-exhaustible fossil fuels and the initial tax rate is below the steady-state level, any fossil fuel h with a higher emissions rate and higher total marginal cost at the initial tax level compared to another fossil fuel, will never be extracted at all, i.e. $x_t^h = 0$ for any $\alpha_h > \alpha_j$ and $c_h + \alpha_h \sigma_0 > c_j \alpha_j \sigma_0$, $1 \leq j, h \leq n \; \forall \; t \in (0, \infty)$.*

Proof: Suppose that for fossil fuel h at the initial tax level $\sigma_0 < \hat{\sigma}$ we have $x_t^h > 0$ given $\alpha_h > \alpha_j$ and $c_h + \alpha_h \sigma_0 > c_j \alpha_j \sigma_0$. Then (1.3) and the complementary slackness condition - which ensure efficiency - imply that $c_h + \alpha_h \sigma_o = u'(.) > c_j + \alpha_j \sigma_o$. However, this is a contradiction as it violates (1.3) for fossil fuel j and therefore we must have $x_t^h = 0$ at the initial tax level. Of course the same argument also holds for any $\sigma_t > \sigma_0$. Since $\sigma_0 < \hat{\sigma}$ the tax has to be increasing over time and therefore $x_t^h = 0$ up to and including the steady-state.

As the tax increases, the marginal cost of fossil fuel j rises and extraction (and consumption) falls (i.e. $\dot{x}_t^j < 0$). This continues until the tax has reached the level that induces a switch to the next fossil fuel, i.e. the level shown in (1.9). At this point consumption switches from fossil fuel j to the next one. The tax continues to increase in line with the emissions stock and there continues to be a fossil fuel switch depending on which has the lowest marginal cost given the current emissions tax. There could be any number of fossil fuels in that order before the switch to fossil fuel k occurs, the one that will remain cheapest until parity with the backstop is reached. Once fossil fuel k is consumed the tax continues to increase to the backstop parity level given by (1.10). This is of course the steady-state where it will remain forever and there is no further switch to any other fossil fuel. This also implies that any fossil fuel with a higher marginal cost at any tax level lower or equal to the steady-state tax will never be used.

Proposition 2 *If there are $n \geq 1$ non-exhaustible fossil fuels and a backstop, any fossil fuel l with a higher total marginal cost than the backstop cost at all tax levels lower or equal to the steady-state tax level will never be extracted at all, i.e. $x_t^l = 0$ if $c_l + \alpha_l \sigma_t > \gamma$ for all $\sigma_t \leq \hat{\sigma}$, $1 \leq l \leq n \; \forall \; t \in (0, \infty)$.*

Proof: Suppose that for fossil fuel l at any tax level $\sigma_t \leq \hat{\sigma}$ we have $x_t^l > 0$ given $c_l + \alpha_l \sigma_t > \gamma$. Then (1.3) and the complementary slackness condition would imply that $c_l + \alpha_l \sigma_t = u'(.) > \gamma$. However, this is a contradiction as it violates (1.4) and therefore

fuel will never be used at all because it always has a higher marginal cost relative to that fuel regardless of the tax.

we must have $x_t^l = 0$. Of course the same also holds for any $\sigma_t > \hat{\sigma}$.

Now let us suppose that we started with an initial emissions stock higher than the steady-state, $M_0 > \hat{M}$. Then the tax would start at a level $\sigma_0 > \hat{\sigma}$ and decrease over time ($\dot{\sigma}_t < 0$). The high tax would mean that only the backstop is consumed initially ($c_i + \alpha_i \sigma_0 > \gamma$ $\forall\ 1 \leq i \leq n$). With only use of the backstop the emissions stock will decreases ($\dot{M}_t < 0$) until the steady-state is reached where the backstop and fossil fuel k are consumed forever. This means that with a high initial emissions stock only one fossil fuel would ever be extracted. However, this also implies that if there is at least one fuel that has a lower extraction cost than the backstop cost, the steady-state has to involve a fossil fuel, it cannot use only the backstop. If there were only consumption of the backstop the tax and the emissions stock would fall towards zero, and this would again make it optimal to use some of the fossil fuels.

Proposition 3 *If there are $n \geq 1$ non-exhaustible fossil fuels as well as a backstop, and there is at least one fossil fuel with $c_j < \gamma$, $1 \leq j \leq \gamma$, the steady-state consumption and emissions levels cannot involve only use of the backstop, there has to be consumption of a fossil fuel.*

Proof: Suppose this were not the case and we have $z_t > 0$ and $x_t^i = 0\ \forall\ 1 \leq i \leq n$ while there is one fossil fuel j with $c_j < \gamma$. Then we know from (1.6) that $\dot{M}_t = -\delta M_t < 0$. The falling emissions stock in turn implies from (1.5) and (1.7) that the tax is falling, i.e. $\dot{\sigma} < 0$. Of course this is not a steady-state. At the same time, it implies that $\gamma = u'(.) < c_i + \alpha_i \sigma_t\ \forall\ 1 \leq i \leq n$. However, as σ_t decreases and approaches zero, at some point $\hat{t}$ we will have $\gamma = u'(.) = c_j + \alpha_j \hat{\sigma}$ which in turn requires both $z_{\hat{t}} > 0$ and $x_{\hat{t}}^j > 0$ in conjunction with both $\dot{M}_{\hat{t}} = 0$ and $\dot{\sigma}_{\hat{t}} = 0$. Therefore the steady-state has to involve consumption of a fossil fuel.

Note that it is possible that at the steady-state tax level $\hat{\sigma}$ there is not only parity between fossil fuel k and the backstop but it could happen that this point also has parity with one or more other fossil fuels, for example fuel $k + 1$. Then the steady-state would involve simultaneous consumption of all the fossil fuels that are at parity at that point as well as the backstop. The analysis has demonstrated that the steady-state will involve use of at least one fossil fuel that is consumed to some degree forever. This means that if there are exhaustion constraints on the fossil fuels we know that this has to be binding on at least one of them.

For simplicity, the analysis in this chapter assumes that the marginal cost of the backstop is constant over the entire time horizon. However, in reality one would expect the cost of the backstop to change over time, and in particular it would be expected to fall. Depending on the point in time at which the backstop cost decreases, this could change the types of fossil fuels that are not extracted at all, or the amount of extraction for those that are used. Of course eventually there will be a steady state involving the backstop at the cheaper cost as well as a fossil fuel. Since a lower cost of the backstop would mean that a lower tax is required to achieve the steady state, it could mean that some of the fossil fuels with a higher extraction cost but lower emissions rate would not be extracted at all compared to the case without a decrease in the backstop cost. While the final steady state would involve a fossil fuel with a higher emissions rate and lower extraction cost, the steady state emissions stock would be lower because the cumulative fossil fuel use before the steady state is lower and the emissions stock remains constant once the steady state is reached.

We have shown a number of results for multiple fossil fuels and a backstop when there are no exhaustion constraints. We will now explore the implications of finite resource availability and therefore extend the model to include exhaustion constraints.

1.3 Exhaustible Fossil Fuels

1.3.1 Model Setup

Following on from the results in the previous section, let us now develop the model where only a finite stock of each fossil fuel is available for extraction. This also means that if an exhaustion constraint is binding (i.e. the stock is exhausted in finite time), the fuel will be subject to a scarcity rent (also called Hotelling rent). Formally the exhaustion constraints state that cumulative extraction over the entire (infinite) time horizon cannot exceed the initial level of the stock, or $\int_0^\infty x_t^i dt \leq S_0^i \ \forall\ 1 \leq i \leq n$.[15] The initial stock S_0^i is given for all fossil fuels. We can then formulate the new maximisation problem as

$$\begin{aligned} Max \quad & \int_0^\infty e^{-rt}\Big[u(\sum_{i=1}^n x_t^i + z_t) - \sum_{i=1}^n c_i x_t^i - \gamma z_t - D(M_t)\Big]dt \\ \text{s.t.} \quad & \dot{M}_t = \sum_{i=1}^n \alpha_i x_t^i - \delta M_t, \qquad (1.16) \\ & \int_0^\infty x_t^i dt \leq S_0^i \quad \forall\ 1 \leq i \leq n, \end{aligned}$$

[15] Alternatively this can also be expressed as $\dot{S}_t^i = -x_t^i$, which means that the change in the fossil fuel stock in period t is equal to the flow rate of extraction at time t.

This means there are n additional constraints, one for each fossil fuel.

1.3.2 Model Analysis

The current value Hamiltonian for this problem then becomes

$$\begin{aligned} H_c =& \Big[u(\sum_{i=1}^{n} x_t^i + z_t) - \sum_{i=1}^{n} c_i x_t^i - \gamma z_t - D(M_t)\Big] \\ & - e^{rt} \sum_{i=1}^{n} \Big\{\mu_i \Big[S_0^i - x_t^i\Big]\Big\} - \sigma_t \Big[\sum_{i=1}^{n} \alpha_i x_t^i - \delta M_t\Big], \end{aligned} \tag{1.17}$$

with the first order conditions given by

$$u'(\sum_{i=1}^{n} x_t^i + z_t) \leq c_i + \alpha_i \sigma_t + e^{rt}\mu_i, \quad x_t \geq 0, \quad \forall\ 1 \leq i \leq n, \tag{1.18}$$

$$u'(\sum_{i=1}^{n} x_t^i + z_t) \leq \gamma, \quad z_t \geq 0, \tag{1.19}$$

$$\dot{\sigma}_t = (r + \delta)\sigma_t - D'(M_t), \tag{1.20}$$

$$\dot{M}_t = \sum_{i=1}^{n} \alpha_i x_t^i - \delta M_t, \tag{1.21}$$

$$\int_0^\infty x_t^i dt \leq S_0^i, \quad \mu_i \geq 0, \quad \forall\ 1 \leq i \leq n, \tag{1.22}$$

$$\lim_{t \to \infty} e^{-rt}\sigma_t = 0, \tag{1.23}$$

where (1.18), (1.19) and (1.22) hold with complementary slackness. The parameter μ_i is the present value scarcity rent for each fossil fuel and this is constant over time, while $e^{rt}\mu_i$ can be thought of as the user cost. It is straightforward to see that the user cost increases at the rate of time preference. From the complementary slackness condition in (1.22) it is also easy to see that the scarcity rent for a particular fossil fuel is strictly positive if and only if the exhaustion constraint is binding. If the stock of a particular fuel is not fully exhausted then the scarcity rent is zero.

The main purpose of this chapter is to analyse which types of fossil fuels may be left in the ground. It is not required to solve the extraction path explicitly in order to determine that any fossil fuel with an extraction cost lower than the backstop cost will be extracted fully in finite time and therefore the exhaustion constraint is binding and the scarcity rent positive. On the other hand, any fossil fuel that has an extraction cost higher than the backstop cost will never be extracted at all. Note that this also implies that any fuel that is extracted to any degree from $t = 0$ will eventually be exhausted fully. This holds regardless of the emissions rate of the fossil fuels.

Proposition 4 *Any fossil fuel with extraction cost lower than the backstop cost will be exhausted fully in finite time regardless of its emissions rate, i.e. $\mu_i > 0$ for any $c_i < \gamma$, $1 \leq i \leq n$. Any fossil fuel with extraction cost higher than the backstop will never be extracted at all, i.e. $x_t^j = 0$, $\mu_j = 0$ for any $c_j > \gamma$, $1 \leq j \leq n \; \forall \; t \in (0, \infty)$.*

Proof: First it is straightforward to see from (1.18) and (1.19) that for any fossil fuel j where $c_j > \gamma$, $1 \leq j \leq n$, we always have $c_j + \alpha_j \sigma_t + e^{rt}\mu_j > \gamma$ even if both $\sigma_t = 0$ and $\mu_j = 0$. Therefore it is never optimal to use such a fossil fuel at any time regardless of the emissions tax or exhaustion constraints. To demonstrate why fossil fuels with lower extraction cost than the backstop have to be exhausted in finite time, suppose first that none of the n fossil fuels are exhausted, none of the exhaustion constraints are binding and therefore $\mu_i = 0 \; \forall \; 1 \leq i \leq n$. Then we are back in exactly the model analysed in Section 1.2. From the analysis there we know that the system will converge to a steady-state that involves at least one fossil fuel k with $c_k < \gamma$. This fossil fuel will be consumed alongside the backstop forever and therefore an infinite amount of the fossil fuel will be extracted. This in turn means that with finite resource availability fossil fuel k will have to be exhausted in finite time. Denote the time at which fuel k is exhausted T_k. Then this leaves us with $n - 1$ fuels at T_k. We can think of T_k as the initial time for the case of $n - 1$ fossil fuels with non-binding exhaustion constraints. That of course means that another fuel $k + 1$ with $c_{k+1} < \gamma$ will be used in a new steady-state and therefore has to have a binding exhaustion constraint as well and will be exhausted at some point T_{k+1} where $T_k < T_{k+1} < \infty$. By iteration we then know that if there are $0 \leq l \leq n$ fuels with extraction cost lower than the backstop, all l fuels will have to be exhausted in finite time.

Another way to think of this is by assuming that there is only one fossil fuel k to begin with where $c_k < \gamma$, there is a low initial level of the emissions stock and emissions tax and therefore from (1.18) we have $u'(x_0^k) = c_k + \alpha_k \sigma_0 + \mu_k < \gamma$. Suppose the scarcity rent is positive. As the tax increases consumption of the fossil fuel decreases and eventually it will either be exhausted or the tax will rise to a level that puts it at marginal cost parity with the backstop. If at that point there were a switch to just the backstop, this would mean that the tax decreases as the emissions stock falls and crucially it also means that $\mu_k = 0$ as the stock is not full exhausted. However, if $\mu_k = 0$ then it would be optimal to re-start use of the fossil fuel k and this implies the backstop and fossil fuel k have to be consumed simultaneously until the fossil fuel is exhausted. In that period the emissions tax is falling, offsetting the continually increasing user cost (i.e. $e^{rt}\mu_k$) and maintaining parity between fuel k and the backstop. While both the backstop and fossil fuel k are consumed simultaneously during this period, the amount extracted in each period falls and use of the backstop increases, which means the emissions stock of course also decreases. Then at some point T_k fossil fuel k will be exhausted and if there is another fossil fuel then this fossil fuel will also have to be exhausted within finite time and so on until all of the fossil fuels with extraction cost lower than the backstop are exhausted. This argument is consistent with the findings in the literature on single fossil fuels with a backstop and stock-dependent extraction costs, for example Tahvonen (1997). Note also that this result holds regardless of the emissions rates of the fossil fuels. A higher emissions rate will mean that extraction may be delayed and spread over a larger time horizon but still the entire fuel stock will be exhausted eventually. This is because when the emissions stock falls with use of the backstop when no fossil fuel is used anymore, the tax will approach zero and therefore reach a level where even use of a fossil fuel with a very high emissions rate will become optimal.

1.3.3 Simultaneous Consumption of Two Resources

In Section 1.2 the emissions tax was the factor driving switches from one fossil fuel to the other and the backstop. Analogous to (1.9) the emissions tax that ensures marginal cost parity between two fossil fuels j and k is now given by

$$\sigma_t^{jk} = \frac{c_k - c_j}{\alpha_j - \alpha_k} + \frac{(\mu_k - \mu_j)e^{rt}}{\alpha_j - \alpha_k}. \tag{1.24}$$

The above shows that the parity tax level is higher compared to (1.9) if $\mu_k > \mu_j$ and vice versa if $\mu_k < \mu_j$. Therefore it depends on the difference in the scarcity rent whether a switch will occur at a lower or higher tax level relative to that which induces a switch in the absence of exhaustion constraints. Intuitively the scarcity rent (if positive) increases the marginal cost of a fossil fuel, but if the next fossil fuel in line has an even higher scarcity rent, the tax required for parity between the two fuels will be higher still. This also influences the time at which the switch occurs. Of course it is entirely possible that the order of marginal costs at a particular tax level that prevailed in Section 1.2 will change completely depending on the relative sizes of the various scarcity rents the fossil fuels will earn. For simplicity let us continue the analysis for just two fossil fuels. In Section 1.2 we saw that the parity level between two fossil fuels cannot continue for any positive length of time and therefore there will simply be a switch from one to the next. However, this may now be different. To see whether the parity level can be sustained over a positive length of time we take the first and second time derivative of (1.24). This gives

$$\dot{\sigma}_t^{jk} = \frac{(\mu_k - \mu_j)e^{rt}}{\alpha_j - \alpha_k} r. \tag{1.25}$$

and

$$\ddot{\sigma}_t^{jk} = \frac{(\mu_k - \mu_j)e^{rt}}{\alpha_j - \alpha_k} r^2 = r\dot{\sigma}_t^{jk}. \tag{1.26}$$

First, note that $\dot{\sigma}_t^{jk} > 0$ only if $\mu_k > \mu_j$. Therefore, assuming that both the emissions stock and the emissions tax are increasing, we also know that simultaneous consumption between two fossil fuels can only be maintained if $\mu_k > \mu_j$. If we have $\mu_k < \mu_j$ then we would require the tax to decrease in order to maintain parity, which is not in line with assumption taken regarding the starting point. Let us suppose that we do have $\mu_k > \mu_j$. Then in order for parity between two fossil fuels to be maintained the tax rate has to increase at an increasing rate and the rate of increase has to be exponential at the rate of time preference. We have shown in Section 1.2 that $\ddot{\sigma}_t > 0$ if $\dot{\sigma}_t > \dot{M}_t \frac{D''(M_t)}{r+\delta}$. At the same time, taking the second derivative of (1.20) with respect to time and plugging this into (1.26) we can then see that for parity to be maintained we need

$$\dot{\sigma}_t^{jk} = \frac{1}{\delta}\dot{M}_t D''(M_t) > 0. \tag{1.27}$$

Combining those two conditions we see that sustaining parity between the fossil fuels requires

$$\frac{1}{\delta}\dot{M}_t D''(M_t) > \frac{1}{r+\delta}\dot{M}_t D''(M). \tag{1.28}$$

It is straightforward to see that this always holds if the emissions stock is increasing (i.e. $\dot{M}_t > 0$). Since this is the case given our assumption of a low level of the initial emissions stock M_0, this is consistent with the requirements for maintaining parity between the two fossil fuels. Plugging this condition back into (1.20) we can then see that the emissions tax in period t has to be

$$\sigma_t^{jk} = \frac{1}{\delta(r+\delta)}\Big[D'(M_t) + \delta\dot{M}_t D''(M_t)\Big]. \tag{1.29}$$

Therefore we have shown that once parity between the two fossil fuels is achieved under an increasing emissions stock and emissions tax, and we have $\mu_k > \mu_j$, it is possible that both of them are consumed at the same time for some periods. Simultaneous consumption can continue until either one of the fossil fuels is exhausted, but a switch to the next fossil fuel may of course occur as the tax continues to increase. The argument presented assumes that both the tax and the emissions stock is increasing, which means there is only consumption of the fossil fuels and no use of the backstop.

Now suppose again that it is fossil fuel k that will be consumed just before the tax rises to a level that brings marginal cost parity with the backstop and there is no simultaneous consumption with any other fossil fuel at that point. The emissions tax increases until the parity level is reached, which is defined as

$$\sigma_t^{kz} = \frac{1}{\alpha_k}[\gamma - c_k - e^{rt}\mu_k]. \tag{1.30}$$

For a binding exhaustion constraint on fuel k this tax level is lower than for the case of no exhaustion constraint. Since the scarcity rent in general increases the marginal cost of a fossil fuel, a lower tax is required to achieve parity with the backstop. Potentially this could also mean that the backstop starts to be used earlier than with no exhaustion constraint and there is less overall extraction of fossil fuels before the backstop is used. As for the case of the two fossil fuels let us see what is required for parity to be maintained. Formally we have

$$\dot{\sigma}_t^{kz} = -\frac{r}{\alpha_k}e^{rt}\mu_k < 0. \tag{1.31}$$

The condition is analogous to (1.25) but requires the emissions tax to be decreasing. Similarly we can then also show that for parity to be maintained we require

$$\dot{\sigma}_t^{kz} = \frac{1}{\delta}\dot{M}_t D''(M_t) < 0. \tag{1.32}$$

This is exactly the same as (1.27) except that with the backstop being consumed the emissions stock will start to decrease as does the tax. For the emissions stock to be decreasing we require $x_t^k < \frac{\delta}{\alpha_k} M_t$ and $\dot{x}_t^k < \frac{\delta}{\alpha_k} \dot{M}_t$. Therefore the decrease in extraction of the second fossil fuel has to be sufficiently strong and vice versa the increase of the backstop use sufficiently large ($\dot{z} > 0$).

While we have seen that it is possible that both fossil fuel k and the backstop will be consumed at the same time, the question is whether this will happen or not. Suppose that it does not, there is no other fossil fuel, and after parity with the backstop the tax does decrease but not sufficiently to outweigh the increase in the scarcity rent. However, this would mean that there would be no further extraction of fuel k and the stock would not be full exhausted. This in turn would mean that $\mu_k = 0$. But then it would become beneficial to re-start extraction of the fossil fuel.[16] Therefore, once parity with the backstop has been reached, both will be consumed simultaneously until either it is exhausted or a switch to other fossil fuels may occur for some time.

Next, suppose that parity with the backstop is reached while we have simultaneous consumption of the two fossil fuels. This means that at that point we must have parity between both of the fossil fuels and the backstop. Therefore we need the tax levels described in (1.24) and (1.30) to be the same, i.e. $\sigma_t^{x^j x^k} = \sigma_t^{x^k z}$. This is achieved when

$$e^{rt}\mu_k = [\gamma - c_k] - \frac{\alpha_k}{\alpha_j}[\gamma - c_j] - \frac{\alpha_k}{\alpha_j} e^{rt}\mu_j \tag{1.33}$$

However, since this leads to $re^{rt}\mu_k = -\frac{\alpha_k}{\alpha_j} re^{rt}\mu_j < 0$ and we know that $\mu_k > 0$, parity between the three fuels cannot be maintained for any positive length of time. Therefore we can only continue with one fossil fuel and the backstop from that point onwards.

While the literature on exhaustible resources without the stock externality problem suggested that it would never be optimal to simultaneously use two resources (e.g. Dasgupta and Heal 1979), the later literature on optimal emissions taxes, in particular Tahvonen (1997), has shown that simultaneous consumption of both a fossil fuel resource and the clean backstop technology is feasible with the stock externality present. This section has also demonstrated that it is feasible that multiple fossil fuels of different stock sizes, extraction cost and emissions rates can be consumed simultaneously for some period of

[16] This argument is analogous to the one made by Tahvonen (1997).

time, although it requires the emissions tax to be at the right level and change at very particular rates.

1.4 Concluding Remarks

The analysis has shown that without resource constraints, if there is at least one fossil fuel with extraction cost lower than the backstop cost, the resulting steady-state has to involve consumption of a fossil fuel forever. Furthermore, any fossil fuel that has a higher marginal cost at the initial tax level relative to another fossil fuel and any fossil fuel that has a higher marginal cost than the backstop at all levels of the emissions tax, will never be extracted at all. Based on the finding that without resource constraints there will be infinite consumption of at least one fossil fuel, the analysis has then shown that with finite resource availability any fossil fuel that has a lower extraction cost than the backstop will be exhausted fully in finite time and any fossil fuel with a higher extraction will not be extracted at all. Finally, while Tahvonen (1997) shows that it is feasible to have simultaneous consumption of the fossil fuel and the backstop, the analysis in Section 1.3.3 has demonstrated that it is also possible that for some time there is simultaneous consumption of two fossil fuels when there is no use of the backstop.

The results of this chapter imply that the consideration of whether to extract a fossil fuel entirely or leave some in the ground is not an environmental consideration but purely based on extraction cost. It is, however, noteworthy that while all of those fossil fuels will be exhausted fully, the environmental consideration is in the timing of extraction. This is consistent with the previous literature on this topic. The optimal emissions tax in conjunction with the scarcity rents will potentially delay use of some of the fossil fuels to the very distant future. Without solving for the paths explicitly it is not possible to make more precise statement but one of the key factors suggesting large time-scales until exhaustion is that the results of this model crucially depend on the assumption that there is positive decay of the emissions stock. Among others, van der Ploeg and Withagen (2012a,b) argue that the time scales for natural decay of carbon dioxide emissions are so large that for all practical purposes one should abstract from it. If we were to assume that $\delta = 0$ we know that the tax will never fall (e.g. van der Ploeg and Withagen 2012a). This also means that once the tax has reached a level that makes it optimal to use the backstop, there will only be consumption of the backstop from that point onwards and any number of fossil fuels may not be exhausted. While it may be argued that it is more realistic to assume no decay of emissions, the result for that case is fairly straightforward and well established in the literature. The contribution of this chapter lies in establishing what types of fossil fuels may be left in the ground in the presence of positive decay, even though the result may be useful for theoretical purposes only.

Another important issue in the analysis of exhaustible resources is that of a carbon leakage. When only some countries participate in measures to reduce emissions, the non-participating countries may emit more than before the participants took action to reduce emissions. An important contribution in this area is Harstad (2012) who looks at this issue from the supply-side perspective, and in particular models the extraction of fossil fuel deposits in a setting where some countries form a coalition to reduce emissions and take into account the damage caused by the global externality, while non-participating countries ignore the damage in their policy choices. He shows that non-participants emit more than in the optimum, may extract the dirtiest types of fossil fuel and invest too little in renewable technologies. However, when countries can trade fossil fuel deposits, the participants will buy the deposits that have the highest extraction costs. This enables the participants to reduce emissions without the non-participants increasing theirs. The analysis by Harstad (2012) complements the analysis in this chapter and the wider literature on optimal fossil fuel extraction. The model in this chapter looks at the first best optimal carbon tax in a single country (global) context. The issue of carbon leakage becomes relevant when a globally optimal carbon tax cannot be implemented. A key conclusion of this chapter is that with exhaustion constraints all fossil fuel reserves will be extracted in finite time as long as they have a marginal extraction cost lower than the marginal backstop cost. This is also consistent with the results of Harstad (2012) if analysed in an infinite horizon context. While initially participants may find it optimal to purchase deposits to avoid increases in the non-participants' emissions, at some point it will become optimal to extract these deposits as well. However, the timing of extraction is the key environmental consideration and extraction may be delayed to the very distant future. This leads back to the assumption of positive decay which is crucial to the results of this chapter. If $\delta = 0$, the traded deposits may never be extracted at all once the backstop is used and fossil fuel consumption stops forever.

It may also be interesting to extend the analysis to include Carbon Capture and Storage technologies (CCS). These would effectively reduce the emissions rate of (some) fossil fuels and this could have significant impact on the extraction levels and which fuel is used at which point time compared to the case without CCS. Such a model would have to involve modelling of the R&D requirements and costs for such a technology relative to R&D for a renewable backstop.

Chapter 2

Altruism & Global Environmental Policy

2.1 Introduction

As Stern (2007) and Galarraga and Markandya (2009) point out, climate change represents one of the largest externalities that society has had to face. As is the case with all types of externalities, the need for policy intervention arises from free-riding behaviour. In standard economic theory, free-riding occurs because individuals are purely self-interested. And since individuals have to carry all the cost of changing their behaviour while only getting a small benefit from the reduction in harm experienced through the externality, there is little incentive for individuals to change their behaviour without some form of intervention. In the case of a global externality like climate change resulting from GHG emissions, individuals may even perceive that their contribution to the externality is so small relative to the total, that the total level of the externality and therefore the damage suffered, are unaffected by the individual's behaviour. In such a case, the rational, and purely self-interested individual will ignore the effect of the consumption choice on total emissions.

The classic prescription to deal with such a problem is the introduction of a Pigovian tax (equal social marginal damage), which makes individuals face the full economic cost of their consumption. However, because it is global greenhouse gas emissions that cause climate change, emissions in one country do not just cause a negative externality in that country but also in every other country. At the same time, while it is aggregate global emissions that cause climate change, the damage experienced from climate change may be much more significant in some countries compared to others.

Governments may implement a tax that maximises the welfare of that country, taking as given the emissions in other countries and the resulting damage for that country, but ignoring the effect of the country's emissions on other countries. While this corrects for individuals' free-riding behaviour within a given country, it is now governments that are free-riding by ignoring the effect the country's emissions have on other countries. And if every country acts that way, such a non-cooperative equilibrium will lead to higher emissions and lower welfare compared to the global cooperative solution where global aggregate welfare is maximised and global damage is fully internalised. For example, van der Ploeg and de Zeeuw (1992) model the Pigovian taxes for a global externality when each country sets their own tax in a non-cooperative way and compare it to the global cooperative solution. They start with a simple static model with flow pollution but the bulk of the analysis focusses on a dynamic approach with stock pollution. Furthermore, Aronsson and Löfgren (2001) develop a dynamic two-country model of a global externality and evaluate the non-cooperative and cooperative equilibria with regard to valuation problems implicit in environmental accounting.

While we know that in the absence of a global regulator countries have no incentive to cooperate, there are a number of approaches that aim to show how the cooperative global optimum might be achieved nevertheless. For example, Barrett (1990) explores where cooperative agreements might arise in case of global externalities. He first shows that the biggest discrepancies between the cooperative and non-cooperative equilibria arise for global externalities that carry significant damage but are costly to reduce (such as climate change), and also for low damage externalities that can be reduced at relatively low cost. He further summaries a number of approaches that might lead to a cooperative equilibrium, among which is the aspect of morality where governments may be guided by some moral concerns rather than just maximising their countries' welfare. Similarly, Barrett (1994) looks at self-enforcing international agreements, which may either be modelled as members maximising the collective net-benefits or as an infinitely repeated game. He shows that neither approach can sustain full cooperative behaviour when the differences between the net benefits of the global cooperative solution and the non-cooperative solution are large.

Furthermore, while emission permit based mechanisms, such as used in the Kyoto protocol, are deemed to be of limited potential to establish the required international cooperation, there are other proposals, for example the price influencing climate protection scheme put forward by Nordhaus (2006).[1] Altemeyer-Bartscher et al. (2010) further develop this approach by proposing a scheme that also includes side-payments. The literature also points to the importance to measure and take into account ancillary benefits of climate protection because regional secondary benefits of climate protection efforts may overcome free-riding behaviour (e.g. Markandya and Rübbelke 2004), and in particular in conjunction with side-payments or tax transfers (e.g. Altemeyer-Bartscher et al. 2011, Markandya and Rübbelke 2012 and Altemeyer-Bartscher et al. 2014). Other contributions in the area of global externalities or global public goods evaluate the effects of labour mobility (e.g. Aronsson and Blomquist 2003)[2], non-competitive markets (Tahvonen 1995), Veblen effects (e.g. Aronsson and Johansson-Stenman 2014)[3], and the existence of abatement activities that only mitigate local pollution in addition to those that mitigate global pollution (e.g. Pittel and Rübbelke 2010).

The potential for free-riding at the country level is of course a result of free-riding at the individual level. It is sometimes thought that if individuals are not completely self-interested, but also have a concern for the welfare of others, this may overcome free-riding behaviour and reduce the required tax on the externality. And although standard economic theory predicts that in a non-cooperative setting individuals will make only negligible contributions to public goods (see for example Andreoni (1988) and Bergstrom et al. (1986)), there a number of theories that show that when contributing yields some sort of utility benefits, voluntary contributions can be consistent with standard economic models. One example of this is Impure Altruism (also called warm-glow effect) which captures that individuals may have a utility benefit from the contribution itself (Andreoni 1989, 1990). Based on this approach, other approaches have made more specific assertions about the underlying psychological motivation for such a utility benefit. This is commonly based on a type of self-image concern or social norm. Examples of this are, among others, Bénabou and Tirole (2006), Ellingsen and Johannesson (2008), Nyborg et al. (2006) and Brekke et al. (2003).

While these models are concerned with a utility benefit to individuals from the contribution itself, other forms of altruism are dealing with a concern for the welfare of others.

[1] This is "essentially a dynamic Pigovian pollution tax" (p.32).

[2] Aronsson and Blomquist (2003) look at optimal tax policy for trans-boundary environmental problems with labour mobility. Using a two-type approach they show that ability has an impact on the optimal tax, but even with labour mobility part of the externality remains uninternalised in a non-cooperative equilibrium. Only a cooperative approach can fully internalise the externality.

[3] Aronsson and Johansson-Stenman (2014) look at optimal provision of both national and global public goods when individuals care about relative consumption levels, both relative to others in their country as well as relative to individuals in other countries.

In the economics literature, there two main types of altruism that deal with individuals' concern for others, namely Pure Altruism and Paternalistic Altruism. Pure Altruism means that an individual's utility may to some degree be a function of others' utility, but not just a specific component of it. Applications of this type of altruism are often used in smaller environments such as the family where one might care about the welfare of specific individuals (for example Becker 1974, 1981). In large-scale contexts it can also be assumed that an individual cares for the total or average welfare of all other individuals in the population (see Hammond 1987; Johansson 1997 for examples). On the other hand, Paternalistic Altruism means that an individual's utility is not a function of others' utility as such, but a specific component of that utility (see Archibald and Donaldson 1976). In an environmental context, this component may be the damage others experience from the dirty good.[4] In both of these cases, the individual's welfare is affected by the damage others experience and therefore it is intuitive that this would induce people to cut back on the harmful activity voluntarily. However, in the context of the global climate change problem Daube and Ulph (2016) demonstrate that as long as individuals continue to believe that total emissions are unaffected by their consumption choice, then no matter how much they care about the welfare of others, their behaviour will not change.[5] Furthermore, it was shown in a single country context that the optimal tax is still the standard Pigovian tax equal to social marginal damage.

This chapter is closely linked to the model setup used in Daube and Ulph (2016) and extends their static model of Pure Altruism to a setting of multiple countries. The model will first be developed under standard theory with self-interested individuals only, and then add individuals' altruistic concern for the utility of others. However, individuals in any given country may exhibit a different level of altruistic concern for the total utility in their own country compared to the concern for total utility in other countries. Indeed, individuals can have different levels of altruistic concern for every country. Intuitively, it is plausible to assume that an individual may care more for the well-being of the people in their own country, or people closer to the individual, compared to individuals who may live half way around the world. The analysis covers individual behaviour but focusses mostly on socially optimal levels of consumption under the non-cooperative solution where governments individually aim to maximise their country's welfare and the global cooperative solution where global welfare is maximised. This chapter makes no efforts to show how such a global equilibrium may be achieved but simply compares the

[4] While Impure Altruism only takes into account the individual's contribution to the externality, Paternalistic Altruism means that the individual is affected by others' experience of the externality, regardless of the individual's contribution.

[5] This result is also consistent with the analysis of Johansson (1997) for large populations.

consumption levels under those different equilibria. The main results are:

- In a non-cooperative equilibrium of two countries a population increase in one country may either increase or decrease equilibrium consumption in another country depending on the relative size of the countries.
- Under standard theory the cooperative equilibrium achieved by a global social planner will internalise global damage fully and a single global tax rate equal to global marginal damage can induce this global optimum.
- Individuals' altruistic concern for the welfare of others (Pure Altruism) will alter the non-cooperative equilibrium resulting from domestic social planners and lead countries to internalise some, but not all, of the damage caused by emissions in their country, depending on the level of altruistic concern relative to the concern for themselves and others in their country.
- However, altruistic concern will also alter the welfare-maximising global optimum, requiring a lower level of emissions as individuals effectively experience damage from the externality through

  1. the direct effect of their country's damage function on personal welfare, and
  2. the effect of altruistic concern and thus the damage individuals in other countries experience from the externality.

  Since socially optimal consumption in the global optimum then depends on the relative levels of altruistic concern, the degree to which damage is internalised in each country will now differ across countries and thus there is no single global tax rate anymore to achieve the optimum.

The chapter will proceed as follows. Section 2.2 describes the basic model setup. Section 2.3 will then develop the general model and evaluate the results under standard theory. This serves as counterfactual to the analysis with Pure Altruism in Section 2.4. Finally, Section 2.5 will provide some concluding remarks and discuss ideas for potential further research.

2.2 Model Setup

We start off with a continuum of individuals living within n discrete countries. Individuals within a given country have the same initial endowment of income $y_i > 0$, where $1 \leq i \leq$

n. An individual chooses consumption levels of a clean good x and a dirty good z, where the clean good is a numeraire good with a price of 1 and therefore represents expenditure on all other goods but the dirty good. Consumption of the clean good generates no externalities. The dirty good, however, generates one unit of emissions, which is a negative externality to all individuals across all countries. The average consumption of the dirty good within country i is denoted by $\bar{z}_i$. The size of the population of each country i is measured by M_i. Therefore total emissions in country i are captured by

$$E_i = M_i, \bar{z}_i.$$

Furthermore, aggregate global emissions are simply the sum of the total emissions in each country, or formally

$$E_T = \sum_{i=1}^{n} E_i = M_T \bar{z}_T.$$

where M_T measures the global population size and $\bar{z}_T$ captures global average consumption of the dirty good. Individuals derive private utility from consumption of the two goods. For simplicity let us assume that individuals within a given country have the same preferences over the two goods and therefore individuals within each country will have identical utility functions. However, preferences over the two goods may differ across countries. Furthermore, utility is linear in consumption of the clean good. This ensures there are no issues of income distribution in the analysis.

The damage individuals in country i experience from the negative emissions externality of the dirty good is captured by the damage functions $D_i(E_T)$. A key component of this multi-country model is that each country can experience different damage as a result of the emissions externality and therefore we have n different damage functions. Therefore this setup also captures any ancillary costs of the emissions externality, which are local to a particular country. The damage function can vary for each country and includes all of the damage associated with the externality, including any ancillary cots. In addition, total global damage is captured by the damage function $D_T(E_T)$. However, note that the damage function is a function of E_T, and therefore it is always total global emissions that determine the damage experienced. This means that emissions in every country contribute equally to global emissions. In a different setting one could model that emissions generated in a particular country cause more damage in that country and only a fraction of the emissions generated in other countries enter the damage function for that country. For example, damage in country i could be given by the function $D_i(E_i + \theta_i \sum_{j \neq i} E_j)$.

However, in the model developed in this chapter it is always total global emissions which matter for the damage function. Therefore every country's experienced damage is a consequence of the same total emissions. Different countries may suffer differently from those total emissions (i.e. different countries may have different damage functions), but the underlying global emissions are always the same for each country. As such we are modelling a truly global pollutant where the aggregate global emissions are what matters. This is in line with the aim of capturing some of the distinguishing features of climate change. Formally, damage experienced by individuals in country i is given by

$$D_i(M_T\bar{z}_T) = D_i(\sum_{i=1}^{n} M_i\bar{z}_i) \qquad \forall \quad 1 \leq i \leq n,$$

where $M_T = \sum_{i=1}^{n} M_i$ and $\bar{z}_T = \frac{1}{M_T}\sum_{i=1}^{n} M_i\bar{z}_i$. M_i captures the population size of each country while $\bar{z}_i$ measures average consumption of the dirty good in each country. Furthermore we have

$$D_T(E_T) = \frac{1}{M_T}\sum_{i=1}^{n} M_iD_i(M_T\bar{z}_T), \quad \text{where} \quad D_T'(E_T) > 0, D_T''(E_T) > 0, \quad \forall \quad E_T > 0.$$

From this it is also straightforward to derive that

$$\frac{\partial D_T(M_T\bar{z}_T)}{\partial z_i} = \frac{M_i}{M_T}\sum_{i=1}^{n} M_iD_i'(M_T\bar{z}_T) \qquad \forall \quad 1 \leq i \leq n,$$

and

$$M_TD_T'(M_T\bar{z}_T) = \sum_{i=1}^{n} M_iD_i'(M_T\bar{z}_T).$$

Therefore, utility derived from consumption of the two goods is given by

$$u_i(x_i, z_i; E_T) = x_i + \phi_i(z_i) - D_i(E_T) \qquad \forall \quad 1 \leq i \leq n, \tag{2.1}$$

where

$$\phi_i'(z_i) > 0, \phi_i''(z_i) < 0; \quad \text{and} \quad D_i'(E_T) > 0, D_i''(E_T) > 0, \quad \forall \quad E_T > 0, \quad 1 \leq i \leq n.$$

Damage experienced from the emissions externality is a strictly increasing and and strictly convex function of total global emissions for all positive levels of emissions. The private gross benefit derived by individuals in country i from consumption of the dirty good is captured by $\phi_i(z_i)$ and is a strictly increasing and strictly concave function of consumption of the dirty good z_i. The dirty good is produced with constant unit cost $c_i > 0$, which again is the same for individuals within country i but may differ across countries. In addition, governments of the different countries impose an emissions tax $t_i \geq 0$ on consumption of z_i. However, in each country this tax revenue is redistributed to the individuals through a lump-sum transfer σ_i which is identical for all individuals in country i. The government budget constraint for country i is therefore defined by

$$\sigma_i = t_i \bar{z}_i, \qquad \forall \quad 1 \leq i \leq n. \tag{2.2}$$

2.3 Standard Theory

We can now develop the model under standard theory, where individuals simply maximise their own private utility of consumption without any altruistic concern for the utility of others.

2.3.1 Individual Behaviour

Given the model setup described in Section 2.2, and in particular the definition of global emissions and the government budget constraint, the individual's private utility can be expressed as

$$u_i(z_i; \bar{z}_i, \bar{z}_T, t_i) = (y_i + t_i \bar{z}_i) - (c_i + t_i) z_i + \phi_i(z_i) - D_i(M_T \bar{z}_T) \qquad \forall \quad 1 \leq i \leq n. \tag{2.3}$$

A key feature of this model is the atomistic nature of the consumption choice, formally captured by the continuum of individuals. This means that the individual's consumption choice has no impact on damage experienced as well as no impact on the government budget constraint. At the same time, the standard Nash assumption that individuals take the consumption of all other individuals as given, also applies. The consequence of those two fundamental assumptions is that individuals take total global emissions, and therefore the damage experienced from the emissions externality, as given. An individual in country i will choose their consumption of the dirty good by maximising their utility shown in (2.3). Using the first order condition it is then straightforward to show that the consumption choice for an individual in country i is characterised by[6]

$$\phi_i'[\tilde{z}_i(t)] = c_i + t_i \qquad \forall \quad 1 \leq i \leq n. \tag{2.4}$$

The left hand side of (2.4) describes the marginal private gross benefit from consumption of the dirty good while the right hand side describes the private marginal cost. Indeed this is the same result as in Daube and Ulph (2016), the only difference being that individuals in different countries may have different private preferences over consumption of the dirty good, may be subject to differences in the production cost of the dirty good, and have their government impose a different tax on the dirty good. Note that since the individual's choice has no impact on total emissions the fact that that utility is a function of the damage experienced by the individual in country i is irrelevant to the choice.

2.3.2 Domestic Social Planner

For the social welfare functions we now have two possible approaches. The first approach is the non-cooperative approach where governments aim to maximise total welfare for their country only (taking consumption in all other countries as given). This will be called the 'Domestic Planner' approach. The second approach is the cooperative global solution where global welfare (i.e. the sum of all countries' welfare functions) is maximised. This is labelled 'Global Planner'. It is obvious from a theoretical viewpoint that the Global Planner's solution yields the globally optimal levels of consumption and emissions.

We start by looking at the optimal level of consumption from the Domestic Planner's perspective. For this let us first define $\bar{z}_{-i} = \sum_{j \neq i} \bar{z}_j$ and $M_{-i}\bar{z}_{-i} = \sum_{j \neq i} M_j \bar{z}_j$, which captures the total consumption of the dirty good across all countries other than country i. Because individuals in country i have identical and strictly concave utility functions, everybody in country i consumes the same amount of the dirty good in the domestic optimum. Therefore country i's social utility function is given by

$$S_i(z_i; \bar{z}_{-i}) = y_i - c_i z_i + \phi_i(z_i) - D_i(M_i z_i + M_{-i}\bar{z}_{-i}) \qquad \forall \quad 1 \leq i \leq n. \tag{2.5}$$

The Domestic Planner takes into account the link between the taxes paid on the dirty good and the lump-sum transfer individuals receive through the government budget con-

[6] All the necessary second-order condition are assumed to hold.

straint. This means the socially optimal level of consumption is independent of the tax rate. Furthermore, the Domestic Planner also takes account of the connection between consumption of the dirty good in country i and global emissions E_T. However, in the multi-country context the Domestic Planner also has to take into account consumption of the dirty good in all other countries because the damage experienced in country i is a function of global emissions which is in turn determined by the consumption of the dirty good in each of the n countries. Since we are dealing with a non-cooperative situation, the standard Nash assumption where the planner takes consumption in all other countries as given applies. The optimal level of consumption in country i from the Domestic Planner's perspective is therefore defined by[7]

$$\phi_i'(\hat{z}_i^D) = c_i + M_i D_i'(M_i \hat{z}_i^D + M_{-i}\bar{z}_{-i}) \qquad \forall \quad 1 \leq i \leq n. \tag{2.6}$$

The above shows that the socially optimal level is achieved when the private marginal gross benefit of consumption is equal to the social marginal cost of consumption in country i. This characterisation of the socially optimal level of consumption is similar to that of the single country model developed by Daube and Ulph (2016). The characterisation fully internalises the damage experienced in country i and takes full account of the government budget constraint and the redistribution of the tax revenues. However, since the Domestic Planner optimises only domestic welfare, this socially optimal level ignores the effect of consumption in country i on the damage experienced in other countries. This is of course an illustration of the free-riding behaviour countries may engage in by ignoring the effects of their actions on other countries.[8]

Result 1 *In a multi-country setting a domestic social planner will free-ride on the damage caused by consumption in their country but experienced in the other countries.*

This result is similar to van der Ploeg and de Zeeuw (1992) although their model has identical damage functions for all countries. Now, by comparing (2.4) and (2.6) it is straightforward to see that the optimal tax on the dirty good in country i inducing everyone in country i to consume the socially optimal level is

$$\hat{t}_i^D = M_i D_i'(M_i \hat{z}_i^D + M_{-i}\bar{z}_{-i}) \qquad \forall \quad 1 \leq i \leq n. \tag{2.7}$$

[7] All the necessary second-order condition are assumed to hold.

[8] Note that if the damage function in country i were linear in total global emissions E, then marginal damage would be constant and therefore marginal damage would be independent of consumption in other countries. In that case both the socially optimal consumption level and the corresponding optimal tax on the dirty good would be independent of consumption elsewhere.

Since it is assumed that the damage function is non-linear in total emissions E, we have already noted that marginal damage is a function of the average consumption level in all other countries. Therefore (2.6) describes the reaction function leading to the Nash equilibrium. In order to explore the equilibrium level of consumption, let us now assume there are only two countries and each sets its socially optimal tax. Then the reaction function for country 1 becomes

$$\phi_1'(\hat{z}_1^D) = c_1 + M_1 D_1'(M_1\hat{z}_1^D + M_2\hat{z}_2^D). \tag{2.8}$$

We already know that consumption in country 1 depends on consumption in country 2 and vice versa. Specifically, we can determine that the degree to which optimal consumption in one country changes as the result of consumption change in the other country is given by

$$\frac{\partial \hat{z}_1^D}{\partial \hat{z}_2^D} = \frac{M_1 M_2 D_1''(.)}{\phi_1''(.) - M_1^2 D_1''(.)} < 0, \tag{2.9}$$

and

$$\frac{\partial \hat{z}_2^D}{\partial \hat{z}_1^D} = \frac{M_1 M_2 D_2''(.)}{\phi_2''(.) - M_2^2 D_2''(.)} < 0. \tag{2.10}$$

It makes intuitive sense that consumption in one country will decrease as consumption in the other country increases. This is because if consumption in one country increases this also increases marginal damage in the other country and therefore the social optimum under the Domestic Planner has to decrease in the other country. From this we can also determine the conditions in the two-country case that need to hold in order to have a unique and stable Nash Equilibrium where both countries consume positive amounts of the dirty good. For this we require $|\frac{\partial \hat{z}_1^D}{\partial \hat{z}_2^D}| < 1$ and $|\frac{\partial \hat{z}_2^D}{\partial \hat{z}_1^D}| < 1$.[9] These conditions are

$$M_1(M_2 - M_1) < \frac{-\phi_1''(.)}{D_1''(.)} \qquad \text{and} \quad M_2(M_1 - M_2) < \frac{-\phi_2''(.)}{D_2''(.)}. \tag{2.11}$$

If one country is bigger than the other, then one of the two conditions will always hold. However, the other is more restrictive and essentially guarantees that one country is not

[9] These are the standard conditions for a unique and stable Nash equilibrium in a Cournot duopoly with linear reaction functions.

too large compared to the other. We will explore this condition further later in this section when we use an example of a specific functional form for $\phi(.)$ and the damage function $D(.)$.

Returning to the analysis of comparative statics, it is straightforward to see that the equilibrium level of consumption in both countries depends on the population size of each country. Looking at the comparative statics of the equilibrium levels of consumption, we can derive from (2.8) and the equivalent for country 2 that

$$\frac{\partial \hat{z}_1^D}{\partial M_1} = \frac{\phi_2''(.)[D_1'(.) + M_1 \hat{z}_1^D D_1''(.)] - M_2^2 D_1'(.) D_2''(.)}{\phi_1''(.)\phi_2''(.) - \phi_2''(.) M_1^2 D_1''(.) - \phi_1''(.) M_2^2 D_2''(.)} < 0. \tag{2.12}$$

Therefore an increase in the population size of country 1 leads to lower consumption of the dirty good in country 1 under the Domestic Planner. This makes intuitive sense as an increase in population leads to an increase in marginal damage which has to be offset by lower consumption levels. However, since the Domestic Planner has to take account of the emissions generated in other countries it is also important to explore how consumption in other countries might change with an increase in the population of country 1. Consumption in country 2 will be affected by the total emissions in country 1, and these are of course a function of the size of country 1 as well as the level of consumption of the dirty good in country 1. However, it is not definitive whether consumption in country 2 will increase or decrease. Specifically we find that

$$\frac{\partial \hat{z}_2^D}{\partial M_1} = \frac{M_2 D_2''(.)[\phi_1''(.)\hat{z}_1^D + M_1 D_1'(.)]}{\phi_1''(.)\phi_2''(.) - \phi_2''(.) M_1^2 D_1''(.) - \phi_1''(.) M_2^2 D_2''(.)} > 0 \quad \text{if} \quad \hat{z}_1^D < \frac{M_1 D_1'(.)}{-\phi_1''(.)}. \tag{2.13}$$

The above shows that consumption in country 2 will increase as a result of a population increase in country 1 if consumption in country 1 is less than the ratio between the marginal damage experienced in country 1 and the rate at which the marginal private gross benefit derived from consumption of the dirty good changes for individuals in country 1. There are two effects driving this. First, an increase in M_1 increases the marginal damage for country 2 and, ceteris paribus, works towards reducing consumption of the dirty good in country 2. However, because the increase in M_1 will reduce consumption of the dirty good in country 1, this reduction may offset the effect on $D_2'(.)$ sufficiently to allow country 2 to actually increase its consumption of the dirty good. From the condition given in (2.13) we see that this is the case if consumption in country 1 is sufficiently low. The lower consumption in country 1, the less an increase in their population will

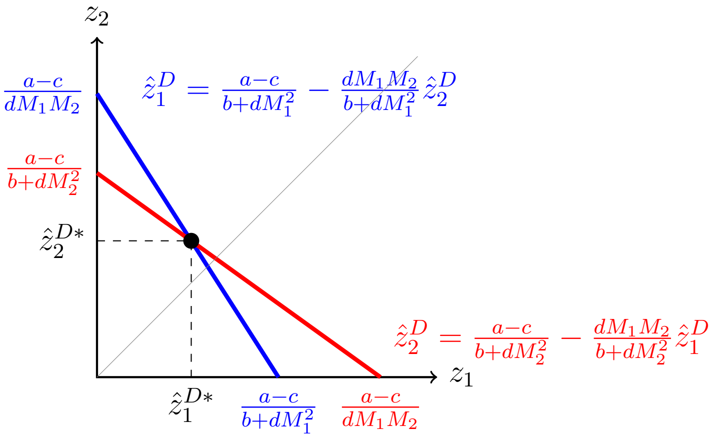

Figure 2.1: Non-Cooperative Nash Equilibrium wtih Domestic Planners

affect marginal damage of country 2 and therefore the more likely that it is that the resulting decrease in consumption in country 1 is sufficient to allow country 2 to increase its consumption of the dirty good.

To illustrate the Nash Equilibrium further it is useful to look at an example of a simple functional form for the private gross benefit from consumption of the dirty good $\phi(z)$ and the damage function $D(E_T)$. In order to simplify the problem further and isolate the importance of relative population sizes, let us also assume that both countries are identical in everything but size. This means they have the same preferences over consumption of the dirty good and experience the same damage from global emissions. Specifically let us assume

$$\phi(z_i) = az_i - \frac{b}{2}z_i^2 \qquad \forall \quad i = 1, 2,$$

and

$$D_i(E_T) = \frac{d}{2}(M_1z_1 + M_2z_2)^2 \qquad \forall \quad i = 1, 2,$$

where $a > c$.

Plugging this into the reaction function for country 1 described in (2.8) we find that the reaction function becomes

$$\hat{z}_1^D = \frac{a-c}{b+dM_1^2} - \frac{dM_1M_2}{b+dM_1^2}\hat{z}_2^D. \tag{2.14}$$

Of course there is an equivalent reaction function for country 2. Solving this system of equations it is then straightforward to derive that the equilibrium level of consumption in country 1 is given by

$$\hat{z}_1^D = \frac{[b + d(M_2 - M_1)M_2]}{b[b + d(M_1^2 + M_2^2)]}(a - c). \tag{2.15}$$

Figure 2.1 further illustrates the reaction functions and the resulting equilibrium level of consumption. Note that this depicts a case were $M_1 > M_2$ since the equilibrium is above the 45 degree line. If both countries had the same population size then of course the equilibrium would be completely symmetric and consumption in both countries would be equal.

We already know from (2.11) the conditions to guarantee the existence and stability of a unique Nash Equilibrium. Given the functional forms we have used, these conditions are now

$$(M_1 - M_2)M_2 < \frac{b}{d} \qquad \text{and} \quad (M_2 - M_1)M_1 < \frac{b}{d}. \tag{2.16}$$

To explore this further let us now say that $M_1 = sM_T$ and $M_2 = (1 - s)M_T$, where the parameter $0 \leq s \leq 1$ captures the size of country 1 relative to the total population mass. Then the two conditions for the Nash Equilibrium become

$$2s^2 - 3s + 1 > \frac{-b}{M_T^2 d} \qquad \text{and} \quad 2s^2 - s > \frac{-b}{M_T^2 d}. \tag{2.17}$$

Figure 2.2 shows that both conditions in (2.17) will hold for any relative population sizes as long as $M_T^2 d < 8b$. The parameter d is the damage parameter and captures the degree of damage caused by the emissions externality. The term $M_T^2 d$ is the second derivative of the damage function with respect to global average consumption (i.e. $\frac{\partial^2 D(.)}{\partial \bar{z}^2}$). Furthermore, $-b$ is the second derivative of the private gross benefit with respect to consumption of the dirty good and therefore b captures the rate at which the marginal benefit of consumption decreases as consumption increases.[10] As such we know that if that rate at which marginal damage increases is less than eight times the rate at which the marginal benefit decreases we will always have a stable and unique Nash Equilibrium under the Domestic Planner. However, if $M_T^2 d > 8b$, then we require either both countries to have similar size or for one country to be much larger than the other. For example, for the

[10] Note that b is the same for both countries in our simplified example.

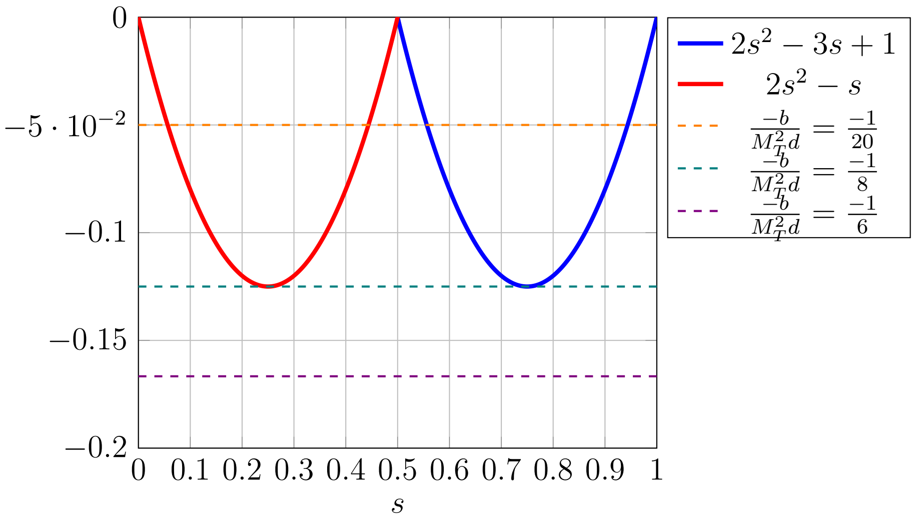

Figure 2.2: Conditions for Existence of Nash Equilibrium

case shown in the figure above where $\frac{-b}{M_T^2 d} = -0.05$ (i.e. $M_T^2 d = 20b$), we require that $s \lessapprox 0.056$, $0.444 \lessapprox s \lessapprox 0.556$ or $s \gtrapprox 0.944$.[11] We can illustrate this example further by determining the consumption levels in each country when we have $M_T^2 d = 20b$. Plugging $M_1 = sM_T$, $M_2 = (1-s)M_T$ and $M_T^2 d = 20b$ into (2.15) we find that the domestic optimum for country 1 is

$$\hat{z}_1^D = \frac{1 + 20(2s^2 - 3s + 1)}{1 + 20(2s^2 - 2s + 1)} \left[\frac{(a-c)}{b} \right]. \tag{2.18}$$

Doing the equivalent for country 2, the domestic optimum for country 2 is

$$\hat{z}_2^D = \frac{1 + 20(2s^2 - s)}{1 + 20(2s^2 - 2s + 1)} \left[\frac{(a-c)}{b} \right]. \tag{2.19}$$

Figure 2.3 plots (2.18) and (2.19) for the whole range of possible values of s. This nicely illustrates the relative population ranges where we have a stable Nash Equilibrium, which are marked by the thick blue and red lines and is consistent with the range determined earlier. If $s < 0.056$ we have positive consumption levels of the dirty good for both coun-

[11] These intervals are of course symmetric since the countries are identical in every aspect but populations size.

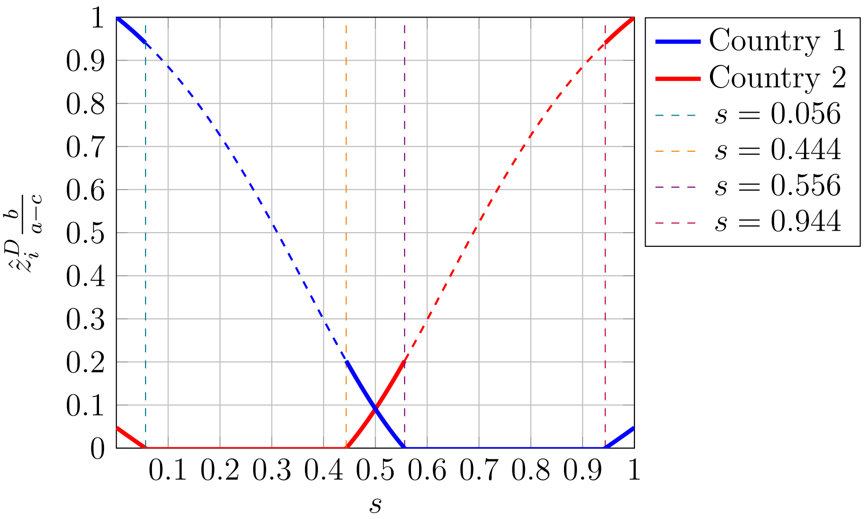

Figure 2.3: Equilibrium Consumption Levels at $M_T^2 d = 20b$

tries and the equilibrium is stable. However, if $0.056 < s < 0.444$, optimal consumption of the larger country 2 would be zero given the consumption of country 1. But this cannot be a stable solution because if $\hat{z}_2^D = 0$ we would have $\hat{z}_1^D = \frac{a-c}{21b}$ which in turn would mean that optimal consumption in country 2 should be positive as well. It makes intuitive sense that the middle section where both countries are of similar size leads to stability as no one country dominates and induces the other country to consume none of the dirty good. Indeed, we already know that if the countries are of exactly the same population size (and therefore completely identical) then we always have a unique and stable Nash Equilibrium. The interesting result, however, is the case where one country is much larger than the other country and this still leads to a stable equilibrium.[12]

Suppose we start with $s = 0.5$ and then slowly decrease s, so that country 1 becomes relatively smaller and country 2 relatively larger, while we hold the total size of the population M_T constant. As country 1 becomes smaller it can increase its domestically optimal consumption level and in turn country 2 decreases theirs as it increases in population size.[13] This effect is true for any relative size between $M_T^2 d$ and b as we can see from Figure 2.4 where $M_T^2 d = 6b$. This continues as s decreases further. In the case of $M_T^2 d = 20b$, the increase in optimal consumption for country 1 eventually has such

[12] The model in this chapter does not consider a production function. In a more general model one should of course reflect that an increase in the population of a country will also have an impact on output.

[13] As the share of the total population in a country approaches zero each individual can consume close to the level that would be consumed if no account of the damage were taken into account ($\tilde{z}(0) = \frac{a-c}{b}$) as the contribution to total emissions becomes negligble.

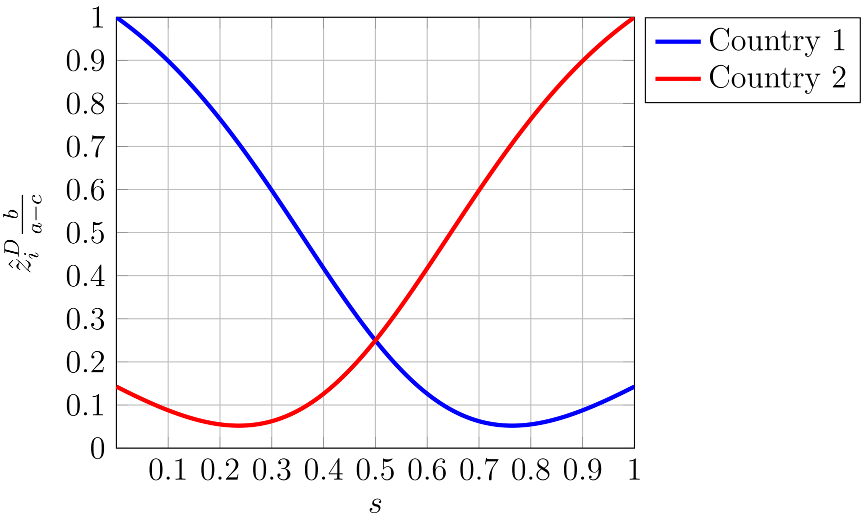

Figure 2.4: Equilibrium Consumption Levels at $M_T^2 d = 6b$

a significant effect on the marginal damage relative to the marginal benefit from consumption in country 2 that it can no longer consume a positive amount of the dirty good given the optimal level in country 1 and the larger population size of country 2. In the equivalent case where $M_T^2 d = 6b$ the only difference is that the marginal damage does not increase at as high a rate relative to the decrease in the marginal benefit so that it is still optimal for country 2 to consume a positive level. However, as is evident from Figure 2.4, as s decreases further eventually a minimum level of consumption for country 2 is achieved and as s decreases the optimal level of country 2 starts to increase again. While consumption levels in country 1 continue to increase with the falling population size, at some point the decrease in population outweighs the increase in consumption in country 1 and the population increase in country 2 with regard to the damage experienced in country 2. At that point country 2 can increase its consumption of the dirty good again. This turning point is consistent with the condition shown in (2.13) where at some point a population decrease of the country 1 may increase consumption in country 2 although that specific condition is based on the case where only M_1 changing, while the example here has M_1 decrease and M_2 increase at the same rate so that M_T stays constant. When we don't have stability over the entire range of population sizes we may have an interval of population shares where there is no stable equilibrium but as s approaches either 0 or 1 a stability will emerge again when the optimal level of consumption of the larger country can increase sufficiently.

Note that these specific results are due to having used simple quadratic functional forms for both $\phi(.)$ and $D(.)$ and that we assumed that the countries are identical in everything

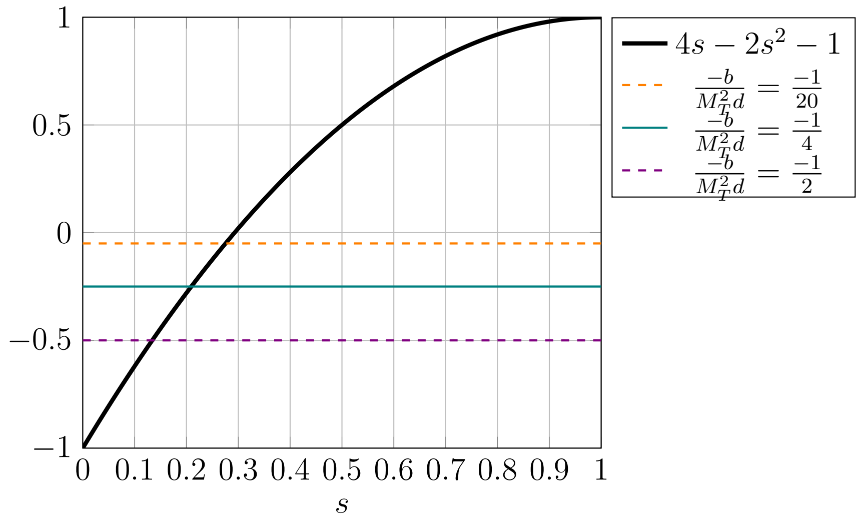

Figure 2.5: Condition for Comparative Statics Result

but size. For example, the number 8 in the threshold of $M_T^2 d < 8b$ has no particular meaning but is a result of those quadratic functional forms. However, the intuition described for the results above holds in general.

Next let us look at the comparative statics for this simplified example. It is straightforward to show that consistent with (2.12) we have $\frac{\partial z_1}{\partial s} < 0$.[14] Furthermore we find that, equivalent to (2.13), the condition required for $\frac{\partial z_2}{\partial s} < 0$ is

$$4s - 2s^2 - 1 < \frac{-b}{M_T^2 d}. \tag{2.20}$$

This is illustrated in Figure 2.5 and shows that since $\frac{-b}{M_T^2 d} < 0$ we know that we never have $\frac{\partial z_2}{\partial s} < 0$ for any $s > \frac{2-\sqrt{2}}{2}$ and that if $d < b$ (i.e. $\frac{-b}{M_T^2 d} < -1$), we never have $\frac{\partial z_2}{\partial s} < 0$ regardless of the relative population sizes.

While it is not possible to draw definitive conclusions from these specific results, they do underline the importance of the relative size of the countries in determining the socially optimal level of consumption in equilibrium even though the Domestic Planner does not take into account the damage that country imposes on other countries. The key to these results is that it is total global emissions which drive the damage experienced from the externality in each country, creating dependencies even when a country tries to only maximise its own welfare with no regard for other countries, and individuals also have

[14] This is effectively the same as $\frac{\partial z_1}{\partial M_1}$ since s measures the relative size of country 1.

no altruistic concern for others. The next step is to see how these results may change if instead of the Domestic Planner there is a Global Social Planner.

2.3.3 Global Social Planner

In contrast to the Domestic Planner a Global Social Planner aims to maximise global welfare across all n countries. This global welfare function consists of the sum of all countries' social welfare functions and is therefore given by

$$S_T(z_1, \ldots, z_n) = \sum_{i=1}^{n} \Big\{ M_i \big[y_i - c_i z_i + \phi_i(z_i) - D_i(M_1 z_1 + \ldots + M_n z_n) \big] \Big\}. \tag{2.21}$$

In the case of the Domestic Planner each country took consumption in every other country as given. However, the Global Planner maximises aggregate welfare and therefore simultaneously chooses the globally optimal consumption level of the dirty good for each of the n countries. This cooperative global optimum for country i is characterised by[15]

$$\begin{aligned} \phi_i'(\hat{z}_i^G) =& c_i + \sum_{j=1}^{n} M_j D_j'(M_1 \hat{z}_1^G + \ldots + M_n \hat{z}_n^G) \\ =& c_i + M_T D_T'(M_1 \hat{z}_1^G + \ldots + M_n \hat{z}_n^G) \qquad \forall \quad 1 \leq i \leq n. \end{aligned} \tag{2.22}$$

Contrary to the case of the Domestic Planner, (2.22) shows that in a global equilibrium global damage of consumption of the dirty good is fully internalised in each country. This means that consumption in country i does not just internalise the damage experienced in country i, but also the damage in all other countries. It also means that if preferences over the dirty good and the cost of production were the same for each country ($\phi(z) = \phi_i(z)$ and $c = c_i$, $\forall\ 1 \leq i \leq n$), then we would have the same level of optimal consumption in each of the n countries. It is important to note that it would be the same level of consumption for each country regardless of their size or damage function. This is because global damage is now equally distributed across all individuals globally and each individual carries an equal share of global damage. The global optimum is driven by global marginal damage rather than the damage experienced in each country. A key factor for this result is the assumption that it is total global emissions that cause the damage experienced in each country. Note that if only a fraction of the emissions caused in other countries would contribute to the damage experienced in country i (as discussed

[15] All the necessary second-order condition are assumed to hold.

in Section 2.2), then the global optimum would not internalise global emissions equally, but country i would only internalise other countries' emissions to the extent that they spill over to other countries.

Furthermore, it is noteworthy that in the global social optimum consumption in country i internalises the damage experienced in country i as a result of a consumption change in country i as well as the damage experienced in country j as a result of a consumption change in country j, which is why it equates to internalising global damage. To illustrate this point, let us look at the simplified version of a 2-country case. Then the global social optimum for country 1 is characterised by

$$\phi_1'(\hat{z}_1^G) = c_1 + M_1 D_1'(M_1 \hat{z}_1^G + M_2 \hat{z}_2^G) + M_2 D_2'(M_1 \hat{z}_1^G + M_2 \hat{z}_2^G). \tag{2.23}$$

The last part of this is the marginal damage experienced in country 2 as a result of a marginal increase of average consumption in country 2. However, this is not the same as the marginal damage experienced in country 2 as a result of a marginal increase in consumption in country 1, which would capture the damage inflicted by country 1 on country 2. Yet, we can also rewrite (2.23) as

$$\phi_1'(\hat{z}_1^G) = c_1 + M_1 D_1'(M_1 \hat{z}_1^G + M_2 \hat{z}_2^G) + \frac{M_2}{M_1} M_1 D_2'(M_1 \hat{z}_1^G + M_2 \hat{z}_2^G). \tag{2.24}$$

Then we see that the global optimum internalises the damage inflicted by country 1 on country 2, adjusted for the relative population sizes. Intuitively this is the case because the Global Planner optimises the joint welfare functions across both countries, which is of course weighted by the different population sizes and thus even a global optimum does not necessarily require that the damage inflicted on another country is fully internalised within one country, but the Global Planner is able to distribute the degree of internalisation such that global welfare is maximised by each country internalising global damage caused by consumption of the dirty good across all countries.

Result 2 *A global social planner fully internalises global damage and ensures that each individual carries an equal share of global damage. If private preferences over the dirty good and the cost of production of the dirty good are the same for all individuals globally (i.e.* $\phi(z) = \phi_i(z)$ *and* $c = c_i \; \forall \; 1 \leq i \leq n$*), then all individuals will consume the same amount of the dirty good under the global social optimum regardless of the countries' population sizes and damage experienced from the externality.*

Comparing (2.4) and (2.22), it is then straightforward to see that the global optimum can be achieved by a single tax rate imposed on each country. This globally optimal tax on the dirty good is equal to global marginal damage, or

$$\hat{t}_i = \hat{t} = M_T D'_T(M_1 \hat{z}_1^G + \ldots + M_n \hat{z}_n^G) \qquad \forall \quad 1 \leq i \leq n. \tag{2.25}$$

This is an important result because if a cooperative solution could be achieved this Global Planner could induce the optimum through a single tax rate on the dirty good for all countries. This would not just be simpler to implement in practice, it also is a fair solution since it imposes the same tax on each individual globally but consumption can still differ in line with private preferences over the dirty good.

Result 3 *The global social optimum can be achieved by a single tax on the dirty good applied to each country, equal to global marginal damage of consumption of the dirty good.*

Let us now compare these results to the case of the Domestic Planner. To illustrate the differences, first assume that the Global Planner imposes the optimal tax on each country. Further, we will use the simplified case of two countries identical in every aspect but size, and with the specific functional forms used previously. With this setup it is straightforward to derive that the globally social optimal level of consumption becomes

$$\hat{z}_1^G = \hat{z}_2^G = \frac{a-c}{b+dM_T^2}, \tag{2.26}$$

where $M_T = M_1 + M_2$.

Note that the socially optimal consumption level is the same for both countries because private preferences over the dirty good as well as the cost of production are assumed to be the same. As already described in the more general setting, it is also noteworthy that optimal consumption would still be the same for each country even if the damage functions were different in different countries. In that case the damage parameter d would simply have to be replaced with a d_T parameter capturing global damage. Figure 2.6 illustrates the global optimum in this simplified setup and compares it to the Domestic Planner equilibrium.

First, note that since consumption in equilibrium will be the same in both countries under the Global Planner, the equilibrium level of consumption lies on the 45 degree line. Even

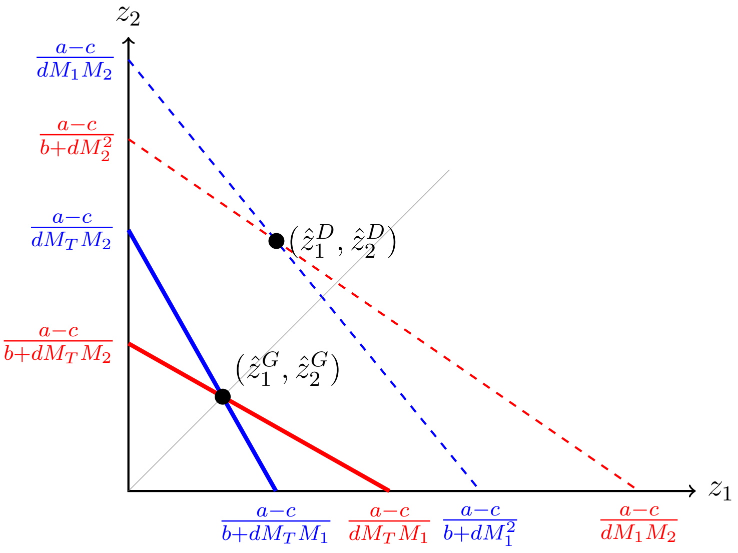

Figure 2.6: Global Planner Equilibrium compared to Domestic Planner Equilibirum

if the size of one of the countries changes or one country has a different damage function, the resulting equilibrium will still lie on the 45 degree line. This is not the case for the Domestic Planner's equilibrium which shifts depending on the relative size of the two countries. In general it is intuitive that average consumption across the two countries will always be lower under the global equilibrium compared to the domestic equilibrium because the Global Planner internalises more damage than each of the Domestic Planners. However, it is theoretically possible that, for example, consumption in country 1 would be higher under the Global Planner than under the Domestic Planner. This would be the case if the size of country 1 is sufficiently large relative to country 2 such that the domestic equilibrium would have a very low level of consumption in country 1 and a comparatively high level in country 2. The Global Planner's equilibrium would then redistribute this imbalance and allow country 1 to consume more, but of course require country 2 to consume less. Formally, in order to have $\hat{z}_1^G < \hat{z}_1^D$ we require that

$$M_1^2 - M_2^2 < \frac{b}{d}. \tag{2.27}$$

This has to hold in addition to the previously stated conditions required for the existence and stability of the Domestic Planner equilibrium.[16] Since $\frac{b}{d} > 0$ we know that the above condition will always hold if $M_2 > M_1$. This means that if country 2 has a larger

[16] Note that under the Global Planner we will always have a unique and stable equiliblrium regardless of the relative population sizes.

population, then the domestic optimum for country 1 will always be larger than the global optimum would be. Intuitively this is because a large country 2 will experience a higher marginal damage for the same average consumption level compared to a smaller (but otherwise equal) country 1 and therefore the consumption of country 2 would be lower than that of country 1. This in turn enables country 1 to have a higher consumption level in the domestic optimum. If however, country 1 is sufficiently large relative to country 2, then it may consume less under the domestic optimum compared to the global optimum. This is because a large country 1 internalises its own experienced damage fully (although it ignores its effect on the other countries) while a global optimum equally shares the burden of global emissions (which are disproportionately caused by country 1 in this example). Of course this may be different if country 1 has a 'lower' damage function compared to country 2.

In the example case of the simple functional forms used above individuals have identical private preferences over the dirty good. For this case it is easy to derive that $\frac{\partial \hat{z}_1^G}{\partial M_1} = \frac{\partial \hat{z}_1^G}{\partial M_2} = \frac{\partial \hat{z}_2^G}{\partial M_1} = \frac{\partial \hat{z}_2^G}{\partial M_2} < 0$. This makes intuitive sense because both countries have to consume the same level of the dirty good after an increase in any of the countries' population size and therefore the change in consumption has to be the same. Of course an increase in population size anywhere leads to an increase in global marginal damage and therefore the consumption of the dirty good by each individual has to decrease under a global social optimum. In more general terms, under a global optimum in the two-country case, the change in consumption in country 1 as a result of the population increase in country 1 is given by

$$\frac{\partial \hat{z}_1^G}{\partial M_1} = \frac{\phi_2''(.)[D_1'(.) + M_1 \hat{z}_1^G D_1''(.) + M_2 \hat{z}_1^G D_2''(.)]}{\phi_1''(.)[\phi_2''(.) - M_1 M_2 D_1''(.) - M_2^2 D_2''(.)] - \phi_2''(.)[M_1^2 D_1''(.) + M_1 M_2 D_2''(.)]} < 0, \tag{2.28}$$

and the change in consumption in country 2 as a result of the population increase in country 1 is given by

$$\frac{\partial \hat{z}_2^G}{\partial M_1} = \frac{\phi_1''(.)[D_1'(.) + M_1 \hat{z}_1^G D_1''(.) + M_2 \hat{z}_1^G D_2''(.)]}{\phi_1''(.)[\phi_2''(.) - M_1 M_2 D_1''(.) - M_2^2 D_2''(.)] - \phi_2''(.)[M_1^2 D_1''(.) + M_1 M_2 D_2''(.)]} < 0. \tag{2.29}$$

This demonstrates that any differences in the amount of change between the two countries as a result of the population increase in country 1 is only driven by the private preferences over the dirty good. This is consistent with the earlier finding that in a global equilibrium differences in consumption are only caused by differences in the private benefit from consumption of the dirty good as well as differences in the cost of production. We established from (2.22) that differences in globally optimal consumption between countries are only driven by differences in private preferences and the cost production. Since the cost of production is assumed to be a constant, it has no impact on the degree to which optimal consumption changes as the result of a population increase. Therefore differences in the degree to which consumption changes are only determined by the different private preferences over the dirty good.

After having developed the model under standard theory we can now begin to explore how these results change when individuals may exhibit altruistic concern for the utility of others.

2.4 Pure Altruism

This chapter uses Pure Altruism in the sense that individuals' utility is a function of the direct utility of all other individuals in population rather than just specific individuals.[17] However, given the multi-country setup, the individual can have a different level of altruistic concern for others' utility in their country compared to others' utility in other countries. Indeed, the individual may have a different level of altruistic concern for each of the n countries. The degree of altruistic concern of individuals in country i for an individual in country i is captured by the parameter $\alpha_{ii} \geq 0$, while altruistic concern of individuals in country i for an individual in country j is captured by $\alpha_{ij} \geq 0$.[18] Therefore the total level of altruistic concern an individual in country i has for the whole population of country i is given by $\alpha_{ii}M_i$, and similarly the total level of altruistic concern an individual in country i has for the whole population of country j is given by $\alpha_{ij}M_j$. Since the degree of altruism is defined as the degree of care for an individual, and the degree of altruism is assumed to be equal for all individuals in that country, the total weight of altruistic concern is greater the larger the population given a fixed level of α. This reflects the idea that while an individual could have a very high level of altruistic concern for individuals in a very small country, this will be offset to some degree by the fact that a large amount of people are affected by the emissions externality in another country, even

[17] See for example Johansson (1997) and Hammond (1987).

[18] Note that the altruism parameter is assumed to be non-negative since a negative α would not capture a degree of altruism but rather some type of 'Schadenfreude'.

though the concern for each individual may be lower. Furthermore, in order to maintain the reasonable assumption that the individual cares at least as much for their private utility from consumption as for the utility of all others, we further require that

$$0 \leq \sum_{j=1}^{n} M_j \alpha_{ij} \leq 1 \qquad \forall \quad 1 \leq i \leq n. \tag{2.30}$$

As will become evident, this assumption is crucial to the interpretation of the results. The condition above effectively ensures that the welfare derived directly from consumption of the dirty good does not become negligible compared to the welfare derived from altruistic concern for others.[19] Also note that for simplicity this chapter assumes that all individuals in any particular country have the same levels of altruistic concern, but altruistic concern between countries is of course allowed to differ.

2.4.1 Individual Behaviour

This section will refer to an individual's utility function that includes utility derived from altruistic concern as the individuals 'welfare'.[20] An individual's personal welfare is given by the sum of their direct utility of consumption and the total utility of all other individuals weighted by the corresponding degrees of altruistic concern, α_{ij}. Therefore we have

$$\begin{aligned} w_i(z_i; \bar{z}_i, t_i, \alpha_{ij}) =& (y_i + t_i \bar{z}_i) - (c_i + t_i) z_i + \phi_i(z_i) \\ & - D_i(M_T \bar{z}_T) + \sum_{j=1}^{n} \alpha_{ij} M_j \bar{u}_j(.) \qquad \forall \quad 1 \leq i \leq n, \end{aligned} \tag{2.31}$$

where $\bar{u}_j$ captures the average personal utility of individuals in country j. Note that individuals have altruistic concern for the direct utility derived from their consumption choice, but no altruistic concern for the indirect utility others derive from their altruistic concern. This is done to avoid a loop effect where one individual would be affected by another's altruistic concern for that individual. From (2.31) it is straightforward to determine that due to the atomistic nature of consumption, the individual's consumption choice has no impact on the utility of others, neither in their own country nor in any other country, and therefore the individual's consumption choice is still characterised by

[19] This approach is consistent with the same argument made in Johansson (1997).

[20] This is of course a personal welfare function rather than a social welfare function.

(2.4). Thus altruism has no effect on an individual's consumption choice. However, it may affect the socially optimal levels of consumption as we will determine in the next section. Just as we did for standard theory we will look at two cases of social welfare optimisation, under a Domestic Planner and under a Global Planner.

2.4.2 Domestic Social Planner

The Domestic Planner maximises total welfare in their country. We assume that total welfare is simply the sum of all individuals' welfare including the effects of altruistic concern for others. Therefore the social welfare function for a Domestic Planner is given by

$$\begin{aligned}
W_i(z_i, \alpha_{ij}) =& S_i(.) + \sum_{j=1}^{n} \Big\{ \alpha_{ij} M_j S_j(.) \Big\} \\
=& (1 + \alpha_{ii} M_i) S_i(.) + \sum_{j \neq i} \Big\{ \alpha_{ij} M_j S_j(.) \Big\} \\
=& (1 + \alpha_{ii} M_i) \Big[y_i - c z_i + \phi(z_i) - D_i(M_i z_i + M_{-i} \bar{z}_{-i}) \Big] \\
&+ \sum_{j \neq i} \Big\{ \alpha_{ij} M_j \Big[(y + t_j \bar{z}_j) - (c + t_j) \bar{z}_j + \phi(\bar{z}_j) - D_j(M_i z_i + M_{-i} \bar{z}_{-i}) \Big] \Big\} \\
&\quad \forall \quad 1 \leq i \leq n.
\end{aligned} \tag{2.32}$$

We still have identical personal welfare functions for all individuals in country i and therefore in the domestic optimum, each individual in country i will consume the same amount of the dirty good. Maximising the social welfare function we find that the domestically optimal level of consumption of the dirty good is characterised by[21]

$$\begin{aligned}
\phi_i'(\hat{z}_i^D) =& c_i + M_i D_i'(M_i \hat{z}_i + M_{-i} \bar{z}_{-i}) \\
&+ \sum_{j \neq i} \Big\{ \frac{\alpha_{ij} M_j}{1 + \alpha_{ii} M_i} M_i D_j'(M_i \hat{z}_i + M_{-i} \bar{z}_{-i}) \Big\} \qquad \forall \quad 1 \leq i \leq n.
\end{aligned} \tag{2.33}$$

The above characterises the reaction function for each country analogous to those in Section 2.3.2 leading to the non-cooperative Nash Equilibrium. Daube and Ulph (2016) found that in the single country setting the level of altruistic concern for the utility of

[21] All the necessary second-order condition are assumed to hold.

others has no impact on the socially optimal level of consumption. However, from (2.33) we see that altruism does affect the domestic optimum in a multi-country setting. Taking a closer look we can see that the degree to which the damage experienced in country i is internalised is independent of the level of altruistic concern individuals in country i may have for others in their country. This is consistent with the results of the model in Daube and Ulph (2016). However, we also see that the degree to which damage experienced in the other countries as the result of a consumption in country i is internalised by the Domestic Planner is driven by the relative levels of altruistic concern. The key factor in this multi-country setting is that the chosen level of consumption by the Domestic Planner affects total emissions and thus also has an impact on the damage experienced in the other country, which in turn has an impact on domestic welfare driven by their level of altruistic concern for other countries welfare.

To clarify this, note that if individuals in country i did not have any altruistic concern for the individuals in any other country (i.e. if $\alpha_{ij} = 0 \ \forall \ j \neq i$), then this would be the same as under standard theory shown in (2.6). However, since individuals may have altruistic concern for individuals in the other countries (potentially a different level for each country), the damage experienced in country $j \neq i$ but caused as a result of consumption in country i now also enters the socially optimally level of consumption in country i. Furthermore, note that the degree to which this influences the domestic optimum is not simply a function of the altruistic concern for the other country, but the level of altruistic concern for the other country relative to the concern for their own private utility and altruistic concern for their own country. This means that the lower the altruistic concern for the other country compared to the concern for their own country, the less of the damage caused to other countries will be internalised.

To get a clearer picture of how exactly the level of altruistic concern influences the domestic optimum it is helpful to look at the simpler case of just two countries. Then the characterisation shown in (2.33) becomes

$$\phi_1'(\hat{z}_1^D) = c_1 + M_1 D_i'(M_1\hat{z}_1 + M_2\bar{z}_2) + \frac{\alpha_{12}M_2}{1+\alpha_{11}M_1} M_1 D_2'(M_1\hat{z}_1 + M_2\bar{z}_2). \tag{2.34}$$

The coefficient $\frac{\alpha_{12}M_2}{1+\alpha_{11}M_1}$ determines to what degree the damage inflicted on country 2 is internalised, and captures the level of altruistic concern for country 2 relative to the total weight given to the individual's private utility of consumption (equal one) and their altruistic concern for all other individuals in country 1.

Due to the constraint on the total level of altruistic concern relative to the private utility of consumption as defined in (2.30), which is $\alpha_{11}M_1 + \alpha_{12}M_2 \leq 1$ in the two country case, and assuming that there is at least some degree of altruistic concern for domestic welfare if there is concern for the other country (i.e. $\alpha_{11} > 0$ if $\alpha_{12} > 0$), then we know that $\frac{\alpha_{12}M_2}{1+\alpha_{11}M_1} < 1$. Therefore the Domestic Planner will never fully internalise the damage caused by consumption in country 1 but experienced in country 2. The only way for the damage to be fully internalised is if there were no domestic altruism ($\alpha_{11} = 0$) while at the same time having the maximum level of altruistic concern for the other country ($\alpha_{12}M_2 = 1$). However, this special case could only occur in the case of two countries. When there are more than two countries, none of coefficients could be equal to one. We can therefore generalise this result to the n-country case and say that for $n > 2$ the Domestic Planner will never fully internalise the damage imposed on any other country caused by consumption in their country, regardless of the level of altruistic concern (given the conditions imposed on the size of altruistic concern). While, unlike in the single country model, the level of altruism does matter to the Domestic Planner, and the impact of consumption on damage experienced in other countries is internalised to some degree, it is not fully internalised.

Result 4 *In a multi-country setting altruistic concern for the welfare of individuals in another country is necessary for a domestic social planner to internalise to some degree the damage inflicted on another country as a result of domestic consumption. The degree to which damage in other countries is internalised depends on the altruistic concern for the other country relative to the concern for the individuals direct utility and the welfare of others in that country. However, for any $n > 2$ the Domestic Planner will never fully internalise the damage inflicted on another country, regardless of the levels of altruistic concern, given the constraints on the levels of altruistic concern as defined in (2.30).*

From (2.33) we can also derive that $\frac{\partial \hat{z}_i^D}{\partial \alpha_{ij}} < 0 \; \forall \; 1 \leq i, j \leq n$, where $i \neq j$. This makes intuitive sense. As the degree of altruistic concern increases relative to the concern for their own country, individuals will internalise more of the other countries' damage and the resulting equilibrium level of consumption will be lower. Similarly, we find that $\frac{\partial \hat{z}_i^D}{\partial \alpha_{ii}} > 0 \; \forall$ $1 \leq i \leq n$. An increase in the altruistic concern for the utility of their own country lowers the relative importance of the damage experienced in other countries and therefore country i can consume more under the domestic equilibrium. Here it is of course important to keep in mind that a decrease in the country i's consumption will also have an effect on the equilibrium consumption in all other countries. Furthermore, as we have established in detail in Section 2.3.2, the Nash Equilibrium level of consumption under the Domestic Planner is a function of the relative population sizes and damage experienced in the other countries, regardless whether there is altruistic concern or not.[22]

So far we have looked at this social optimum as internalising the damage inflicted on another country. However, we can also evaluate whether the levels of altruistic concern could be sufficient to induce a Domestic Planner to internalise global damage. To investigate this we start by rewriting (2.34) as

$$\phi_1'(\hat{z}_1^D) = c_1 + M_1 D_i'(M_1\hat{z}_1 + M_2\bar{z}_2) + \frac{\alpha_{12}M_1}{1+\alpha_{11}M_1} M_2 D_2'(M_1\hat{z}_1 + M_2\bar{z}_2). \tag{2.35}$$

This is exactly the same as before only rewritten such that the last part of the equation captures the marginal damage experienced in country 2 as the result of a consumption change in country 2. It shows that the damage in country 2 is internalised to the degree defined by the coefficient $\frac{\alpha_{12}M_1}{1+\alpha_{11}M_1}$. If this coefficient is equal to one, the Domestic Planner will fully internalise global damage, the same as the Global Planner would determine under standard theory without altruism. We know that this is the case if

$$(\alpha_{12} - \alpha_{11})M_1 = 1.$$

From this we see immediately that the condition requires that $\alpha_{12} > \alpha_{11}$, which means it requires that the level of altruistic concern for the other country is larger than it is for their own country. Furthermore, we can also rewrite this condition as

$$\frac{1+\alpha_{11}M_1}{\alpha_{12}M_2} = \frac{M_1}{M_2}. \tag{2.36}$$

We know that the left hand side of (2.36) is greater than one due to the constraint imposed earlier on the total level of altruistic concern relative to the private utility of consumption ($\alpha_{11}M_1 + \alpha_{12}M_2 \leq 1$). Therefore for the condition to hold, we require M_1 to be larger than M_2 and in such a way that the ratio between the sum of the concern for the individual's private utility and the welfare of all others in the same country and the concern for individuals in the other country is the same as the ratio between the population sizes of the two countries. Intuitively this means that the size of country 2

[22] Assuming non-linear damage functions.

has to be sufficiently small and the level of altruistic concern for the other country be sufficiently larger than the concern for the domestic welfare in order for the total weight of the altruistic concern to have enough weight in the domestic social welfare function that it fully internalises the damage of country 2. In the n-country case this would have to hold for each of the other countries and for this to be consistent with the restriction in (2.30) country 1 has to become larger and larger relative to all other countries the more countries there are.[23] It is noteworthy that it is counter-intuitive that this condition would actually hold, since it is reasonable to expect that the level of altruistic concern for the welfare in another country is not greater that the level of concern for welfare in their own country.

Next we want to establish how the domestic optimum with Pure Altruism is affected by a change in population size in country i. Again turning to the case of only two countries, we find that the change in the domestic optimum in country 1 as a result of the population increase in country 1 is given by

$$\begin{aligned}\frac{\partial \hat{z}_1^D}{\partial M_1} = \frac{1}{|\mathbf{H}^D|}\Bigg\{ & \phi_2''(.)\Big[D_1'(.) + M_1 z_1 D_1''(.) + A_1^D M_2 z_1 D_2''(.)\Big] \\ & - (1 - A_1^D A_2^D)\Big[M_2^2 D_1'(.) D_2''(.)\Big]\Bigg\} < 0,\end{aligned} \tag{2.37}$$

$$\begin{aligned}\text{where} \qquad A_1^D &= \frac{\alpha_{12} M_1}{1 + \alpha_{11} M_1}, \qquad A_2^D = \frac{\alpha_{21} M_2}{1 + \alpha_{22} M_2}, \\ \text{and} \qquad |\mathbf{H}^D| &= \phi_1''(.)\Big[\phi_2''(.) - A_2^D M_1 M_2 D_1''(.) - M_2^2 D_2''(.)\Big] \\ &\quad - \phi_2''(.)\Big[M_1^2 D_1''(.) + A_1^D M_1 M_2 D_2''(.)\Big] > 0.\end{aligned}$$

Since we know that $A_1^D A_2^D < 1$ from the restriction put on the total level of altruistic concern we also know that, as in the case without altruism, an increase in consumption in country 1 leads to a decrease in domestically optimal consumption of the dirty good in country 1. Furthermore, the resulting change in consumption in country 2 is given by

$$\begin{aligned}\frac{\partial \hat{z}_2^D}{\partial M_1} = \frac{1}{|\mathbf{H}^D|}\Bigg\{ & \phi_1''(.)\Big[A_2^D D_1'(.) + A_2^D M_1 z_1 D_1''(.) + M_2 z_1 D_2''(.)\Big] \\ & - (1 - A_1^D A_2^D)\Big[M_1 M_2 D_1'(.) D_2''(.)\Big]\Bigg\}.\end{aligned} \tag{2.38}$$

[23] For example in the case of three countries we would require that $M_1 \geq 2M_2$ and $M_1 \geq 2M_3$.

Similar to the findings under standard theory, it is not clear whether consumption in country 2 will increase or decrease as a result of the population increase in country 1. The condition for $\frac{\partial z_2}{\partial M_1} > 0$ is given by

$$-\phi_1''(.) < \frac{(1 - A_1^D A_2^D)[M_1 M_2 D_1'(.) D_2''(.)]}{A_2^D D_1'(.) + A_2^D M_1 z_1 D_1''(.) + M_2 z_1 D_2''(.)}, \tag{2.39}$$

where of course the left hand side is a positive value. While this condition is not straightforward to interpret, the intuition is the same as for the case without altruism. If country 1 is sufficiently large or the marginal damage in country 1 is sufficiently high relative to the private benefit of consumption in country 1 and the size and marginal damage in country 2, the decrease in consumption in country 1 can be such that country 2 could actually increase consumption of the dirty good under a Domestic Planner. Of course the levels of altruistic concern the two countries have for themselves and the other country influences how much a country changes its consumption of the dirty good.[24]

We have already determined earlier that equilibrium consumption in country 1 will decrease with an increase in their level of altruistic concern for country 2 (i.e. $\frac{\partial \hat{z}_1^D}{\partial \alpha_{12}} < 0$), and that equilibrium consumption in country 1 will increase with an increase in the level of altruistic concern for their own country (i.e. $\frac{\partial \hat{z}_1^D}{\partial \alpha_{11}} > 0$). In the result shown in (2.37) we defined the parameters A_1^D and A_2^D. These represent the level of altruistic concern by individuals in country 1 and country 2 for the other country relative to the concern for their country and the direct utility from consumption respectively. As already established it is the relative levels of altruism that matter, and therefore it is these parameters that dictate to what degree the damage in another country is internalised. Using these, we can derive that

$$\frac{\partial \hat{z}_1^D}{\partial A_1^D} = \frac{1}{|\mathbf{H}^D|} \left\{ M_2 D_2' \Big[\phi_2''(.) - A_2^D M_1 M_2 D_1''(.) - M_2^2 D_2''(.) \Big] \right\} < 0. \tag{2.40}$$

Therefore, as we would expect, consumption in country 1 decreases as the relative level of altruistic concern for country 2 increases. However, equilibrium consumption in country 1 also depends on consumption in country 2. Since we know that an increase in A_2^D will decrease consumption in country 2, we can deduce that this means that consumption in

[24] Also note that if both $A_1^D = 0$ and $A_2^D = 0$, then the condition described in (2.39) reduces to that shown in (2.13).

country 1 will increase. Formally this is shown by

$$\frac{\partial \hat{z}_1^D}{\partial A_2^D} = \frac{1}{|\mathbf{H}^D|}\left\{ M_1 D_1' \Big[M_1 M_2 D_1''(.) + A_1^D M_2^2 D_2''(.) \Big] \right\} > 0. \tag{2.41}$$

This shows that consumption in country 1 increases as the relative level of altruistic concern by individuals in country 2 for country 1 increases. This is consistent with our earlier findings. Note that A_2^D does not affect the consumption choice of country 1 directly but only through the change in the consumption level for country 2. Therefore we can also express (2.41) as

$$\frac{\partial \hat{z}_1^D}{\partial A_2^D} = \frac{\partial \hat{z}_2^D}{\partial A_2^D}\frac{\partial \hat{z}_1^D}{\partial \hat{z}_2^D} > 0.$$

This illustrates that as A_2^D increases, consumption in country 2 decreases in line with the equivalent of (2.40) for country 2. And since consumption in country 2 decreases, consumption in country 1 increases.

We can now turn to see how the socially optimum level from the perspective of a Global Planner is affected by the presence of Pure Altruism.

2.4.3 Global Social Planner

As before, the global social planner simultaneously maximises the joint welfare functions of all n countries. Therefore total global welfare is given by

$$W_T(z_1, \ldots, z_n) = \sum_{i=1}^{n} \left\{ M_i \Big[(1 + \alpha_{ii} M_i) S_i(.) + \sum_{j \neq i} [\alpha_{ij} M_j S_j(.)] \Big] \right\}. \tag{2.42}$$

Taking the first order condition with respect to z_i, it is then straightforward to derive that the globally socially optimal consumption of the dirty good in country i is characterised by[25]

$$\begin{aligned}\phi_i'(\hat{z}_i^G) =& c_i + M_i D_i'(M_1\hat{z}_1^G + \ldots + M_n\hat{z}_n^G) \\ &+ \sum_{j\neq i}\left\{\frac{1+\sum_{h=1}^n[\alpha_{hj}M_h]}{1+\sum_{h=1}^n[\alpha_{hi}M_h]}M_jD_j'(M_1\hat{z}_1^G + \ldots + M_n\hat{z}_n^G)\right\} \qquad \forall \quad 1\leq i\leq n.\end{aligned} \tag{2.43}$$

The above shows that - just as in the case of the Domestic Planner - the damage in country i is fully internalised, but the degree to which damage in the other countries is internalised depends on the relative levels of altruistic concern. Although this relative level of altruism is different from the Domestic Planner case, it shows that it is by no means given that the Global Planner fully internalises global damage equally in each country. One may intuitively expect that for a Global Planner the degrees of altruistic concern would cancel out somehow and we would still arrive at the same global optimum as described under standard theory in (2.22). However, this is not the case and altruism clearly does matter. In order to get a better understanding of what exactly is going on, let us look at the simplified case of only two countries. Then the global optimum for country 1 is characterised by

$$\phi_1'(\hat{z}_1^G) = c_1 + M_1D_1'(M_1\hat{z}_1^G + M_2\hat{z}_2^G) + \frac{1+\alpha_{12}M_1+\alpha_{22}M_2}{1+\alpha_{11}M_1+\alpha_{21}M_2}M_2D_2'(M_1\hat{z}_1^G + M_2\hat{z}_2^G). \tag{2.44}$$

The first point that is evident from (2.44) is that the degree to which damage in country 2 is internalised does not just depend on the relative level of altruistic concern in country 1 for their own country and country 2 (as is the case for the Domestic Planner), but now the combined levels of altruistic concern in both countries for country 2 relative to the combined levels of altruistic concern in both countries for country 1 drive the degree to which damage in country 2 is internalised. This is because the Global Planner maximises the joint global welfare function that also includes the effect of country 2 on country 1 and their levels of altruistic concern. Indeed, it is straightforward to see that if in both countries the level of altruistic concern is at the same level for their own country as it is for the other country (or in other words $\alpha_{11} = \alpha_{12}$ and $\alpha_{22} = \alpha_{21}$), then the coefficient is equal to one and the socially optimal level fully internalises global damage (i.e. $\phi_1'(\hat{z}_1^G) = c_1 + M_TD_T'(M_1\hat{z}_1^G + M_2\hat{z}_2^G)$). And of course it is also evident that in this case the socially optimal level in country 2 fully internalises damage in country 1 as well, and

[25] All the necessary second-order condition are assumed to hold.

we have the same optimum as under standard theory without altruism. To generalise this to the n-country case we would require $\alpha_{ii} = \alpha_{ij} \; \forall \; 1 \leq i, j \leq n$. But because the levels of altruistic concern may be different for each country, and each of those parameters enters the global welfare function, the global optimum with Pure Altruism may very well be different from the standard theory result where each country internalises global damage to the same degree. Also note that even if there is only altruistic concern by individuals for their own countries ($\alpha_{ij} = 0 \; \forall \; 1 \leq i, j \leq n, \; i \neq j$) then the global optimum is still not the same as under standard theory unless the degree of altruistic concern within each country is also the same.

However, it is not strictly necessary to have exactly the same levels of altruistic concern for the other countries as for their own country in order for the coefficient for the marginal damage of the other country to be 1. More generally we can see that the marginal damage in the other country will be fully internalised if

$$\frac{\alpha_{11} - \alpha_{12}}{\alpha_{22} - \alpha_{21}} = \frac{M_2}{M_1}.$$

Let us refer to the difference between the altruistic concern for their own country and the concern for the other country as the 'additional domestic altruism' (e.g. $\alpha_{11} - \alpha_{12}$), assuming for simplicity that the altruistic concern for their own country is larger than for the other country. Then we can see from the above that if the relative level of additional domestic altruism in both countries is equal to the inverse of the relative size of the two countries the global optimum will internalise global damage equally for each country. In the n-country case, this requirement becomes a little more complicated. In order for the damage of country j to be fully internalised in the socially optimum consumption of country i we would require the population weighted additional domestic altruism in country i to be equal to the sum of the population weighted additional altruism for country j relative to country i by all other countries. Specifically we would require

$$(\alpha_{ii} - \alpha_{ij})M_i = \sum_{h \neq i} \left\{ (\alpha_{hj} - \alpha_{hi})M_h \right\} \qquad \forall \quad 1 \leq i, j \leq n, \quad i \neq j. \tag{2.45}$$

The above condition is necessary for the global optimum to equally internalise global damage in each country. However, the key intuition behind a global social optimum including degrees of Pure Altruism in the welfare function is that one country may internalise more of the damage due to the relative levels of altruism compared to another country. Indeed, if the population weighted sum of the altruistic concern for country j across all countries is greater than the sum of population weighted altruistic concern for country i, then the social optimum for country i would imply internalising more than global marginal damage. Of course in turn country j's optimal consumption level would internalise less than global marginal damage. If, in a two-country example, country 1 has a large amount of altruistic concern for the welfare of country 2 but country 2 has a low level of altruistic concern for the welfare of individuals in country 1 but a high level of concern for the welfare of their own country, then the global welfare function will put more weight on the damage experienced in country 2 than a Global Planner would do under standard theory. Altruistic concern of individuals in country 1 for country 2 means that the individuals in country 1 do not just experience the damage directly, but they also suffer to some degree the damage the individuals in country 2 experience. Therefore they are affected more by the externality than they would be without altruism.

Result 5 *In a multi-country setting the socially optimal level of consumption from a global social planner's perspective is driven by the relative levels of altruistic concern. If, in each country, the levels of altruistic concern for the domestic welfare is the same as the level of altruistic concern for welfare in all other countries, then the social optimum internalises the global damage equally across each country. However, if the relative levels of altruistic concern for a particular country by other countries are large, the welfare of that country - and therefore the damage function of that country - will carry more weight in determining the global optimum.*

Comparing both the optimal consumption levels under the Domestic Planner with that of the Global Planner, we can further determine that as long as the level of altruistic concern for domestic welfare is larger than the level of concern for the welfare in other countries (i.e. $\alpha_{ii} > \alpha_{ij}\ \forall\ 1 \leq i, j \leq n,\ i \neq j$), then the global social optimum will lead to a lower consumption level of the dirty good compared to the domestic social optimum. This makes intuitive sense since the Global Planner would, on average, internalise more of the global effect of the externality than a Domestic Planner could achieve through the effect of altruism alone.

Next, we again want to show how the optimal consumption levels in each country are affected by a population increase in one country. For this, let us turn back to the case of only two countries. Then we can show that

$$\begin{aligned}\frac{\partial \hat{z}_1^G}{\partial M_1} = \frac{1}{|\mathbf{H}^G|}\Bigg\{ & \phi_2''(.)\Big[D_1'(.) + M_1 z_1 D_1''(.) + A_1^G M_2 z_1 D_2''(.)\Big] \\ & - (1 - A_1^G A_2^G)\Big[M_2^2 D_1'(.) D_2''(.)\Big]\Bigg\} < 0,\end{aligned} \tag{2.46}$$

$$\begin{aligned}\text{where} \qquad & A_1^G = \frac{1+\alpha_{12}M_1+\alpha_{22}M_2}{1+\alpha_{11}M_1+\alpha_{21}M_2}, \qquad A_2^G = \frac{1+\alpha_{11}M_1+\alpha_{21}M_2}{1+\alpha_{12}M_1+\alpha_{22}M_2}, \\ \text{and} \qquad & |\mathbf{H}^G| = \phi_1''(.)\Big[\phi_2''(.) - A_2^G M_1 M_2 D_1''(.) - M_2^2 D_2''(.)\Big] \\ & \qquad - \phi_2''(.)\Big[M_1^2 D_1''(.) + A_1^G M_1 M_2 D_2''(.)\Big] > 0.\end{aligned}$$

Since $A_1^G A_2^G = 1$ we can see that the second part of the numerator falls away. Note that the only differences between (2.46) and (2.37) are the altruism parameters A^D and A^G. With regard to the impact of a change in M_1 on the global optimum for country 2 we find that

$$\frac{\partial \hat{z}_2^G}{\partial M_1} = \frac{1}{|\mathbf{H}^G|}\Bigg\{\phi_1''(.)\Big[A_2^G D_1'(.) + A_2^G M_1 z_1 D_1''(.) + M_2 z_1 D_2''(.)\Big]\Bigg\} < 0. \tag{2.47}$$

Therefore optimal consumption falls in both countries as a result of a population increase in either country. This is consistent with Global Planner results under standard theory shown in Section 2.3.3. Finally, similar to the Domestic Planner case, it is intuitive and easy to show that

$$\frac{\partial \hat{z}_1^G}{\partial A_1^G} < 0 \qquad \text{and} \qquad \frac{\partial \hat{z}_1^G}{\partial A_2^G} > 0.$$

As we would expect, if the level of altruistic concern for country 2 (by either country) increases relative to the altruistic concern for country 1, the optimal consumption level in country 1 decreases as the welfare of country 2 then has more weight in the global welfare function compared to country 1. Similarly, if the altruistic concern for country 1 increases relative to the altruistic concern for country 2, the optimal consumption level for country 1 increases. We have now completed the analysis of the non-cooperative and cooperative equilibria in the presence of Pure Altruism. To complete the analysis let us now see how the optimal tax to induce this optimal behaviour under both the Domestic Planner and the Global Planner has changed compared to standard theory.

2.4.4 Optimal Tax

We have seen in Section 2.4.1 that adding Pure Altruism to the individual's utility function does not alter individual behaviour due to the atomistic nature of the consumption decision in this model. However, we have also determined that altruistic concern for other countries does influence the socially optimal level of consumption, whether this is the level determined through the non-cooperative equilibrium as a result of Domestic Planners or the global optimum determined by the cooperative solution of the Global Planner. This is a key difference compared to the results under standard theory as well as compared to a single country model with Pure Altruism.

Yet, because individual behaviour is unaffected by their altruism, under both solutions discussed every individual in country i still consumes the same amount of the dirty good and we can induce this solution through a tax on the dirty good. For the Domestic Planner, it is straightforward to derive that the required tax on the dirty good is

$$\begin{aligned}\hat{t}_i^D =& M_i D_i'(M_i \hat{z}_i^D + M_{-i}\bar{z}_{-i}) \\ &+ \frac{M_i}{1+\alpha_{ii}M_i}\sum_{j\neq i}\alpha_{ij}M_j D_j'(M_i\hat{z}_i^D + M_{-i}\bar{z}_{-i}) \qquad \forall \quad 1 \leq i \leq n.\end{aligned} \tag{2.48}$$

As under standard theory, this tax may be different for each country depending on their damage function. In addition, it is now also a function of the altruistic concern individuals in this country may have. However, we also see now that, unlike under standard theory, to achieve the global optimum we cannot impose the same tax in every country and need to have a different tax for each country depending on the relative levels of altruistic concern. As such the required tax for country i to induce the global optimum is

$$\begin{aligned}\hat{t}_i^G =& M_i D_i'(M_i \hat{z}_i^G + M_{-i}\hat{z}_{-i}^G) \\ &+ \sum_{j\neq i}\left\{\frac{1+\sum_{h=1}^n[\alpha_{hj}M_h]}{1+\sum_{h=1}^n[\alpha_{hi}M_h]}M_j D_j'(M_i\hat{z}_i^G + M_{-i}\hat{z}_{-i}^G)\right\} \qquad \forall \quad 1 \leq i \leq n.\end{aligned} \tag{2.49}$$

As discussed in the previous section already, the required tax on the dirty good would only be the same for each country if for each country the degree of altruistic concern for their own country is the same as that for every other country. Even if that were the case, it means that inducing the global optimum requires significantly more information about individuals and the damage caused in individual countries; knowledge about global damage is not sufficient anymore. Not only does the Global Planner require information about the various levels of altruistic concern, but also about the different damage functions for each country. This is important because with altruistic concern, even if a cooperative solution can be achieved, it is not straightforward for policy makers to determine the Pigovian tax for each country.

Result 6 *With Pure Altruism in a multi-country setting the cooperative global (first-best) solution can no longer be induced by a single global tax on the dirty good unless in each country the levels of altruistic concern for the domestic welfare are the same as the concern for welfare in all other countries.*

2.5 Concluding Remarks

This chapter has shown that while individual behaviour is unaffected by altruism in the multi-country setting, meaning individuals will free-ride in the absence of a tax on the dirty good regardless of their level of altruistic concern for others, determining the optimal level of consumption of the dirty good for both a non-cooperative and cooperative welfare maximising solution, as well as determining the right tax to induce that level, becomes more complex. Individual behaviour is unaffected due to the atomistic nature of the consumption decision. Therefore no matter how much they may care about the welfare of others, in their country or another, their consumption choice has no impact on global emissions. The chosen consumption level is only determined by the private marginal benefit and cost of consumption.

Given the global nature of climate change, non-cooperating governments might maximise domestic welfare, taking consumption in other countries as given. However the social optimum is determined by the global cooperative solution where global welfare is maximised for each country simultaneously. With or without altruism, the global optimum implies lower average consumption compared to non-cooperative solution. While individuals will free-ride on the damage their consumption causes, governments maximising domestic welfare only will also free-ride on the damage caused to other countries. Under standard theory the Domestic Planner will only internalise domestic damage, while a Global Planner will internalise global damage completely and differences in optimal consumption are only caused by differences in private preferences and the cost of production of the dirty good. Therefore the global optimum can be induced by a single tax rate on the dirty good for everyone globally, equal to global marginal damage from the dirty good. This is a common result in the literature.

However, this chapter has shown that if individuals exhibit altruistic concern for the utility of others, both the non-cooperative solution as well as the global optimum will change. If individuals exhibit altruistic concern for the welfare of other countries, then the Domestic planners' non-cooperative solution will internalise the damage caused to other countries depending on the level of altruistic concern for that country relative to the altruistic concern for their own country and their own private utility. However, the Domestic Planner will never fully internalise global damage if there are more than two countries. Altruistic concern for other countries effectively increases the damage individuals experience as they will not just be affected by the damage they experience directly, but also the damage others experience. Consequently, even the Domestic Planner has to take account of damage caused to other countries. The key result, however, is that the global optimum is affected by the existence of altruistic concern for the same reasons. Effectively altruism means that global damage is now not the simple sum of all the individual damage functions but the weighting has to be adjusted for the damage experienced via altruistic concern for others. This depends on the relative levels of altruistic concern each country has for their own country and for other countries, and means that unless for each country the altruistic concern for their own country is the same as for all other countries, the global optimum will be different from standard theory. From this it also follows that although altruistic concern leads the Domestic Planner to internalise some of the damage, the gap between the non-cooperative solution and the global optimum may not narrow. Furthermore, this chapter has shown that with altruism the global optimum can no longer be achieved by a single tax equal to global marginal damage but has to be adjusted for each country depending on the levels of altruistic concern. This significantly increases the information policy makers require to set the right tax on emissions even if a global cooperative solution could be achieved.

This chapter is inspired by the model developed in Daube and Ulph (2016), whose main contribution is the development of an alternative theory of behaviour where individuals do not necessarily act in a utility-maximising way, but may base their consumption decision on a hypothetical moral benefit determined by asking what would be optimal if they and everybody else were to make the same choice. Therefore a natural extension of this chapter would be to apply this alternative theory of behaviour to the multi-country setting developed here. Furthermore, in order to develop more concrete policy recommendations it may be useful to link this model to empirical analysis of the degrees of altruistic concern individuals may have for welfare in their own country and other countries.

Chapter 3

Moral Behaviour, Altruism & Environmental Policy

3.1 Introduction

In a recent paper assessing the challenges that policy-makers and society face in addressing climate change, Galarraga and Markandya (2009) point out the key ethical and welfare considerations that need to be taken into account - in particular the intra-and inter-generational impact of the damage that climate change may bring. While economists frequently consider these ethical considerations in terms of the formulation of the appropriate welfare objective for policy-makers to pursue, an equally important issue is how far individuals themselves take these factors into account when deciding what actions to take and what policies they are willing to support. This raises the important question of the extent to which policy action might be needed if individuals themselves are willing to alter their behaviour as they recognise the potential harm that their actions might cause.

As both Stern (2007) and Galarraga and Markandya (2009) point out, climate change represents one of the largest externalities that society has had to face, and, like all externalities, the fundamental need for policy intervention arises from free-riding behaviour - in this case not just by individuals and companies but also by governments. In the classical analysis of externalities, free-riding arises because individuals are purely self-interested, and so perceive that, while they bear all the costs of changing their consumption behaviour, they may get only a very small gain in terms of the reduced damage that they themselves will suffer. In the extreme case, individuals may calculate that their emissions are so insignificant relative to the total, that total emissions and hence any damage that they (and others) might suffer will be unaffected by whatever consumption choices they make.[1] In these circumstances individuals make their consumption choices ignoring any effect these choices have on climate change and the damage it will cause.[2] The classic prescription is the introduction of a Pigovian tax (equal to social marginal damage) so that individuals face the full economic cost of their consumption decisions.

It is sometimes thought that if individuals are not self-interested, but instead are altruistic and so take account of the effect of their actions on others, this may overcome free-riding behaviour. However, as we will show, as long as individuals continue to believe that total emissions are completely unaffected by whatever they do, then howevermuch they care about others, their behaviour will not change, and the optimal policy is to impose exactly the same tax as if individuals were self-interested. For individual behaviour to change, individuals need to make their consumption decisions in a different way. While there have been a number of theories of pro-social behaviour, these typically involve individuals obtaining some kind of utility gain from behaving morally. This chapter proposes a new theory of moral behaviour whereby individuals recognise that they will be worse off by not acting in their own self-interest, and balance this cost off against a hypothetical moral value of adopting a Kantian form of behaviour, that is by calculating the consequences of their action by asking what would happen if everyone else acted in the same way as they did. The analysis of the baseline model shows that

1. individuals behaving this way will adjust their behaviour to take account of the impact of their decision on themselves and others.
2. if individuals behave in this way, then altruism can matter and, depending on the choice function, the greater the degree of altruism the more individuals cut back their consumption of a 'dirty' good.
3. nevertheless the optimal environmental tax is exactly the same as that emerging from the classical analysis where individuals are purely self-interested.

While the baseline results are developed under a simple linear choice function, this chap-

[1] This inability to influence the total level of the externality will be referred to as 'atomistic' consumption.

[2] Echoes of such free-riding behaviour can be found in a paper by Longo et al. (2012) reporting on a study conducted in the Basque Country on people's willingness to pay for the ancillary benefits of climate change mitigation. They note that, as in many such studies, while many people reveal a positive willingness to pay, there is a group of protestors who say they do not want to pay for these benefits, either because they do not think the proposed policy actions will be effective, or because they feel that others should contribute.

ter shows that these results also hold with a more generalised choice function.[3] Further extensions to the baseline model show that even if preferences are heterogenous, the optimal tax is still the same as under the baseline model. This also holds when individuals exhibit a desire for conformity in addition to the alternative form of behaviour. Furthermore, if individuals choose consumption of two different dirty goods, an increase in the tax on one good can either increase or decrease consumption of the other good.

A key result of this chapter is that the existence of altruistic and 'moral' behaviour does not change the socially optimal tax on an environmentally harmful good. Johansson (1997) previously modelled the socially optimal tax on an externality for different types altruism and compared it to the standard Pigovian tax level. Using a setting of discrete individuals, he finds that the optimal tax in the presence of altruistic concern for others' utility is the same as the standard tax in large populations. This is in line with the findings of this chapter.[4] However, Johansson (1997) uses a different approach to analysing the case where individuals' behaviour is driven by something other than maximising their utility (i.e. Genuine Altruism), and finds that the optimal tax is lower than under standard theory. Contrary to this chapter, Johansson (1997) is not analysing a Kantian type of behaviour and therefore uses a different type of choice function that includes the individual's and everybody else's utility. He further assumes that all individuals are identical in their degree of altruism or morality.

The chapter is structured as follows. Section 3.2 conducts a brief discussion of the literature on pro-social behaviour and altruism. Section 3.3 then develops the baseline model, starting with standard theory excluding any form of altruism in Section 3.3.1. This serves as counterfactual to the remaining analysis. Section 3.3.2 then proceeds to layer on a type of Pure Altruism that will be defined carefully in what follows. Section 3.3.3 turns to the central element of this chapter by developing the theory of moral behaviour. Sections 3.3.4 and 3.3.5 will then look at two variations of the baseline model to see how the results may change. Following this, Section 3.4 furthers the analysis by looking at a generalisation of the choice function to show that the main results are not dependent on the linear choice function used in the baseline model. Section 3.5 extends the model to the case where individuals choose consumption of a range of different dirty goods rather than just one while Section 3.6 develops the model to capture heterogenous preferences

[3] Note that, as shown in Section 3.4.6, a choice function non-linear in the hypothetical Moral Benefit altruism does not necessarily always lead to a reduction in consumption of the dirty good as the level of altrsuim increases.

[4] However, the model developed in this chapter uses a continuum of individuals in order to capture the atomistic nature of the consumption choice in the context of global environmental externalities such as climate change.

over the dirty good. As a final extension, Section 3.7 combines the theory of alternative behaviour with a model where individuals have a desire for conformity. Finally, Section 3.8 will present a brief discussion of the model and results as well as make suggestions for future research based on the findings in this chapter.

3.2 Review of Literature

3.2.1 Warm-Glow and Concern for Self-Image

As described in the Introduction, pro-social (and pro-environmental) behaviour is usually at odds with standard economic analysis, which predicts that in a non-cooperative setting, individuals only make negligible contributions to public goods. For example, Andreoni (1988) shows that in large populations the share of individuals making contributions to a public good tends to zero as the free-riding effect dominates.[5] However, when contributing to the public good also yields some utility benefit to the individual, voluntary contributions can be consistent with standard economic models. Andreoni (1989, 1990) models the individual's utility not just as a function of the consumption of the private and public goods, but also of the individual's contribution to the public good itself. This is commonly referred to as the 'warm-glow' effect and describes a form of Impure Altruism. Andreoni's development of warm-glow giving is based on the analysis of the provision of impure public goods by Cornes and Sandler (1984). The analysis of impure public goods also has importance in the area of environmental policy. For example, Markandya and Rübbelke (2004) analyse ancillary benefits of climate policy and argue that policymakers should consider these more thoroughly. Further, Markandya and Rübbelke (2012) look at the under-provision of impure public technologies in an environmental context while Altemeyer-Bartscher et al. (2014) take into account that climate policy is an impure public good in their analysis of tax-transfer schemes in relation to international public goods. However, Impure Altruism and impure public goods are not the same concept. Impure public good means that the good itself has some private characteristics and this may impact individuals' contribution to the public good. Whether a public good is a pure or impure public good depends on the characteristics of the good, and is independent of how the good is funded. However, Impure Altruism relates to the decision individuals make regarding the funding for the provision of the public good, and, in particular, to the relative weights they place on the benefits to themselves and to others in making that decision. For example, a public good may of a pure nature (i.e. both non-excludable and non-rivalrous) and confer no private benefits of the type mentioned above. An altruistic

[5] Also see Bergstrom et al. (1986).

individual will value the benefits that the provision of this good brings to others but may, in addition, derive some utility gain (warm-glow) from knowing that they are taking an action (contributing) that gives benefits to others. The label 'impure' therefore refers to the nature of utility derived from an altruistic action, but not to the properties of the public good the individual is contributing to.

'Warm-glow' can also be interpreted as a self-image gain from contributing to the public good.[6] While Andreoni makes no assumptions regarding the psychological cause of this 'warm-glow', various other authors have developed more sophisticated models with regard to the underlying motivation. These models usually work on the premise that individuals derive intrinsic value from a self-image desire or social norms. Bénabou and Tirole (2006), for example, model a "reputational payoff" (p. 1656) from contribution to a public good, which is also a function of the belief others have regarding the type of consumer this individual is. In a model by Ellingsen and Johannesson (2008) the level of social approval depends on whether the individual himself approves of the person who approves him. Nyborg et al. (2006) also construct a model where individuals are motivated by a concern for self-image. However, this self-image is a function of the total benefit a 'green' good yields to the population, as well as their perception of what share of the population is choosing to consume the 'green' option.[7] This means the consumer's intrinsic incentive to be prosocial increases as the share of the population acting that way increases.[8] On the other hand, Brekke et al. (2003) develop a model where individuals are able to make a more sophisticated calculation of the "morally ideal effort" (p. 1971). This is achieved by evaluating the socially optimal contribution to a public good if they and everybody else were to make the same choice. The individual then derives self-image value depending on how close their contribution is to that socially optimal level, while trading off the self-image gain against the utility benefit from consumption. In this setting the individual of course still behaves according to the utility-maximisation principle. However, a key difference to the standard warm-glow approach is that individuals are able to make the quite sophisticated calculation of what the social optimum constitutes. The results of the analysis by

[6] Also note that Andreoni refers to the case of an individual who only cares about the total supply of the public good as Pure Altruism. However, although the total supply of a public good also affects others, this chapter, as well as other literature on altruism, uses the term Pure Altruism to refer to an individual's direct concern for the utility of others. A more detailed description of how Pure Altruism is modelled will be provided later in this section.

[7] To some extent this also captures the idea of a social norm or peer pressure and an individual's belief about what others do can be more important than what they actually do. Their analysis also shows that such norms can lead to herd behaviour. See Section 3.2.2 for more on social norms.

[8] Because what matters in their model is the individual's perception of what others do, Nyborg et al. (2006) further argue that policy makers may be able to influence this perception, for example through advertising. Also, a temporary tax on the environmentally harmful good could move the population to a permanent 'green' equilibrium even when the tax is later removed.

Brekke et al. (2003) still lead to under-provision of a public good and furthermore show that extrinsic incentives can reduce private provision of the public good due to the effect they can have on the individuals' perception of what constitutes the morally best choice.[9]

3.2.2 Social Norms

The concepts on warm-glow and self-image concern link to the notion of social norms where individuals' choices can be affected by what others in the population do. This is a very broad area and there are various ways to approach this. For example, conspicuous consumption, a term introduced by Veblen (1924), describes the consumption of status goods in order to signal status to others. When individuals are motivated this way, the utility an individual obtains is a function of the individual's consumption relative to the consumption of a peer group. This type of competitive consumption can lead to over-consumption of a good. Similarly, an individual may also be willing to pay a higher price for a status good compared to a non-status - but otherwise equal - good. This is called the Veblen effect.[10] Arrow and Dasgupta (2009) use an inter-temporal model to analyse conspicuous consumption behaviour and find that conspicuous consumption might not lead to a market distortion but this is dependent on the number of goods subject to the conspicuous consumption effect as well as the formulation of the utility function. Dasgupta et al. (2015) model conspicuous consumption with multiple goods where consumption creates a negative environmental externality. They also find that conspicuous consumption does not necessarily have to lead to a market distortion and there may be no need for the government to correct for the conspicuous consumption effect. However, they also note that these results are only derived under very specific conditions.

While an individual strives to consume more than others when competitive consumption is present, another approach to social norms is that an individual will benefit from making a consumption choice as close as possible to a certain norm level.[11] Information effects are one way in which the choice of others can influence consumption decisions.

[9] Another area of the literature where individuals may have utility gains from behaving in pro-environmental ways looks at individuals' concern for relative consumption. An example relevant to the analysis of global public goods such as climate protection is the work by Aronsson and Johansson-Stenman (2014), who look at optimal provision of national and global public goods when individuals care about relative consumption levels.

[10] The Veblen effect is used to explain the Easterlin Paradox (Easterlin 1974, 2001) which states that once a certain per capita income is reached, a further increase in income per capita has no impact on individuals' well-being.

[11] See Hargreaves Heap (2013) for an overview of the various ways in which individuals may derive benefit from conforming to a social norm.

For example, Allcott (2011) use a natural experiment where residential electricity customers were sent reports comparing their consumption to that of neighbours, resulting in significant reductions in electricity consumption.[12] Being member of a group can also enhance trust between individuals which is similar to the idea of social capital. However, while group membership may enhance trust within the group it may also reduce trust to individuals outside the group (see for example Putnam (2000), Dasgupta (2000) and Hargreaves Heap 2013).

Another way in which social norms can affect behaviour is when more intrinsic benefits are derived from belonging to a group. Akerlof and Kranton (2000) build a model incorporating self-identity and how this could affect individual interaction. They argue that individuals derive a psychological benefit from being part of a group since this a key aspect of self-identity. This is in line with Adam Smith's concept of "special pleasure of mutual sympathy" (Smith 1759, Chapter 2). Similar to the warm-glow effect from the contribution to a public good described earlier, individuals may also derive a warm-glow from conforming with a social norm. This type of behaviour where individuals are more willing, for example, to contribute to a public good if they know that others in their peer group act similarly, is called conditional cooperation.[13] For example, Azar (2004) models tipping behaviour where the tipping choice is determined both by conformity to a social norm driven by the avoidance of social disapproval, as well as a warm-glow factor such as self-image and impressing others. He finds that for a norm to be sustained over time these other benefits have to be present, otherwise the norm will disappear eventually. Such a warm-glow effect derived from behaviour relative to a social norm can also have an impact on recycling behaviour as investigated by Bruvoll and Nyborg (2004), Brekke et al. (2010) and Abbott et al. (2013), among others. In addition, Buchholz et al. (2014) analyse non-governmental public norm enforcement in a two-stage model where individuals first voluntarily contribute to a non-governmental agency that provides social approval incentives for public good provision in the second stage. That way voluntary public good provision can be maintained in a non-cooperative equilibrium where individuals do not exhibit any altruism.

Bernheim (1994) develops a model where individuals care about the perception others have about their type relative to some ideal type.[14] The analysis shows that if status is

[12] This is consistent with similar studies using smaller samepls such as Nolan et al. (2008) and Schultz et al. (2007).

[13] See Chaudhuri (2011) for a literature survey on sustaining cooperation in laboratory public good experiments.

[14] An individual's type is private information in this model.

important enough then, despite heterogenous preferences, individuals may conform to a single norm. However, individuals with extreme preferences will not conform regardless of the loss of status.[15] Ulph and Ulph (2014) develop a model where individuals have a desire for conformity but choose whether to adhere to a norm before they make their consumption decision. If an individual chooses to adhere to the norm, the individual derives a benefit from conformity but at the same time has disutility depending on how far the chosen consumption level is from the norm level and the individual's strength of adherence to the norm. In addition, the norms are not exogenous but are determined endogenously by the consumption decisions individuals make. This model is the basis for the extension developed in Section 3.7 and will be described in more detail then.

3.2.3 Motivation Crowding

The analysis by Nyborg et al. (2006) described earlier argues that extrinsic incentives such as taxes and subsidies can enhance the level of green consumerism. There are a number of studies aiming to establish how intrinsic incentives of behaviour can be affected by extrinsic incentives such as fines or taxes. The theories of this are usually based on case study observations where people's behaviour changed in response to the imposition of some extrinsic incentive. A key contribution in this field was made by Deci (1971), whose experiment established that people are motivated by intrinsic incentives and found that extrinsic (monetary) rewards dependent on performance reduce the intrinsic motivation for a certain task. Another example is Frey and Oberholzer-Gee (1997) who find that people's willigness to accept a nuclear facility in their community was reduced when monetary compensation was offered. Frey (1997) points out that external incentives can both crowd-out or crowd-in intrinsic motivation depending on the type of incentive applied. In the context of environmental issues, Frey (1999) and Frey and Jegen (2001) argue that policy instruments such as taxes and subsidies can crowd out environmental concern as they reduce the level of self-determination of individuals and violate the idea of reciprocity based on the concept of an implicit contract of acknowledgement of each other's action that is broken through extrinsic incentives. Indeed, the crowding-out effect can even outweigh the relative price effect of an extrinsic incentive. In addition, Frey (1999) develops the proposition that not just does any form of regulation undermine environmental concern, but the complexity of such regulation is of importance. Furthermore, he argues that taxes do crowd out environmental concern but the effect is smaller compared to the case of tradable emission permits as these can be seen as giving the

[15] Such reputational gains are similar to the approach used by Bénabou and Tirole (2006) described earlier and based on earlier work such as Akerlof (1980) and Jones (1984).

owner a legitimate right to pollute and therefore reduce all sense of 'wrongness' associated with the activity. Bénabou and Tirole (2006) also argue that extrinsic incentives can crowd out intrinsic motivation but develop a slightly different reason for this effect. They argue that the existence of extrinsic incentives leads to uncertainty over whether the individual's behaviour is driven by intrinsic motivation or the extrinsic incentives and therefore makes it difficult to assess the 'virtue' of the individual. Bénabou and Tirole (2006) take the view that prosocial behaviour is largely driven by concerns of reputation and extrinsic incentives such as rewards or punishments can crowd out this motivation. In line with this they also argue that the level of intrinsic motivation is a function of memorability and prominence of the prosocial activity as this increases the reputational value of the activity. They note, however, that with heterogeneity in individuals' image concerns, prosocial activities may be suspected of having been undertaken only for image reasons and this would reduce the effectiveness of rewards like public praise. Similarly, the inferences and reputational value that can be drawn from any prosocial activity also depend on what others in the economy choose to do.

While extrinsic incentives can reduce the level of intrinsic motivation, Frey (1999) argues that environmental concern can be enhanced (or crowded-in) through personal relationships, through principals and agents communicating with each other, or when employees are able to participate in the decision making process because these factors increase the level of self-determination and acknowledgement of intrinsically motivated behaviour. In particular he develops the proposition that environmental concern is supported in the short term by appeals and participation, and in the long run by education which increases the self determination aspect. Nyborg and Rege (2003) find that while different models of the motivation for prosocial behaviour such as impure altruism, social norms and fairness lead to mixed conclusions regarding the influence of government provision on private contributions to a public good, in all models subsidies will crowd-in private contributions. Due to the dynamics of the crowding out-effect, Frey (1999) points out that either 'low' or 'high' environmental tax rates are more effective than 'intermediate' tax rates. This is because a 'low' tax would work in support of environmental concern, while 'high' tax rates will completely outweigh moral concerns yet still achieve the desired outcome. Tax rates in between, however, may crowd out moral motivation while the extrinsic incentive is not strong enough to reduce emissions sufficiently. He concludes that a complementary approach of both traditional and behavioural policy instruments is important for effective environmental policy. Traditionally, when a policy instrument such as a tax is seen not to be working as desired, it is usually just applied in a stronger way. However, by doing that it is possible that the crowding-out effect starts dominating and environmental damage may even increase. Instead, the tax should be supported by a policy instrument that supports the crowding-in of environmental concern.

3.2.4 Altruism

Most of the contributions discussed in Section 3.2.1 capture a form of Impure Altruism where, to some extent, the contribution itself matters to an individual's utility. However, altruism is instead often viewed as a concern for the welfare of others.[16] The two main types of altruism that capture an individual's concern for others' welfare are Pure Altruism and Paternalistic Altruism. Pure Altruism uses the idea that an individual's utility may to some degree be a function of others' utility, but not just a specific component of it. Applications of this type of altruism are often used with smaller settings, such as the family where one might care about the welfare of specific individuals (for example Becker 1974, 1981). In large-scale contexts it can also be assumed that an individual cares for the total or average welfare of all other individuals in the population (see Hammond (1987) and Johansson (1997) for example).[17] Paternalistic Altruism differs from Pure Altruism in that it does not assume that an individual's utility is a function of others' utility as such, but a specific component of that utility (see Archibald and Donaldson (1976)). In an environmental context, this component may be the damage experienced by others from the environmentally harmful good. While Impure Altruism only takes into account the individual's contribution to the externality, Paternalistic Altruism means that the individual is affected by others' experience of the externality, regardless of the individual's contribution. All these types of altruism still assume that individuals maximise their utility when acting pro-socially. Genuine Altruism as defined by Kennett (1980), on the other hand, requires that individuals' behaviour is driven by some function other than maximising their utility. For example, the individual may make the consumption choice by maximising a function consisting of their own utility and everybody else's utility, but by doing so will incur a loss in utility compared to standard economic behaviour.[18] Since this implies a deviation from 'rational' behaviour driven by self-interest economists usually assume, it is the most drastic form of altruism.

Johansson (1997) models the socially optimal tax on an externality for all four types of

[16] The literature on altruism is very rich and we will only make a superficial survey as relevant to this chapter. For more detail, Fontaine (2008) provides a summary of the history of the concept of altruism in economics.

[17] Also see Nyborg and Rege (2003) for basic modelling approaches to public good provision including Pure and Impure Altruism.

[18] See Johansson (1997) for an example of this.

altruism described above and compares it to the standard Pigovian tax level. He uses a model of discrete, identical consumers, who choose between a 'clean' good and a 'dirty' good, which in turn causes the externality. The government imposes a tax on the dirty good, but tax revenues are distributed back to the consumer. The analysis finds that under Pure Altruism the optimal level of tax is equal to the standard Pigovian level in large populations. This result is driven by the assumption that the size of altruistic concern has to be small relative to the private consumption utility in order to maintain realistic proportionality of the utility function. However, with a form of Genuine Altruism where individuals maximise a function of the weighted sum of their own utility and everybody else's utility, Johansson (1997) finds that the optimal tax is lower than the Pigovian tax. The socially optimal level of consumption is unchanged from the standard level as this type of altruism does not affect it, but the individual will demand this lower level of consumption due to the function maximised and therefore the requirement on the tax level is reduced. If the weight in the maximisation were equal between the individual's utility and all others' utility, the tax rate would drop to zero. Furthermore, for Paternalistic Altruism Johansson (1997) finds that the tax is higher than the Pigovian tax because, as individuals take into account others' experience of the externality, the socially optimally level of the externality is reduced and therefore the optimal tax level increases. For the case of Impure Altruism, the optimal tax is exactly equal to the Pigovian tax. Although the socially optimal level of consumption is lower than the standard level, this shift is exactly the amount that occurs due to the 'warm glow' from the individual's contribution to the externality. Therefore, the optimal tax level does not need to be higher in order to achieve the socially optimal level of consumption.

3.2.5 Towards an Alternative Theory of Behaviour

In his analysis Johansson (1997) recognises that Genuine Altruism is the most controversial of all types of altruism as it contradicts mainstream economic theory. Yet the traditional model has already for a long time drawn fundamental criticism. Sen (1977) describes why the economist prefers to assume the existence of a selfish being: "It is possible to define a person's interests in such a way that no matter what he does he can be seen to be furthering his own interests in every isolated act of choice." (p. 322). Indeed it means that every action can be explained simply through the concept of revealed preferences. This saves the economist from having to take a closer look at what constitutes preferences and utility, or as Sen (1977) puts it: "a robust piece of evasion." (p. 323). Sugden (1982) further argues that assuming utility-maximising behaviour in the traditional sense may be too constricting and different people might indeed maximise very different functions in order to determine their consumption choice.

The criticism presented here does not seek to suggest that the utility-maximising model is without value, and the critical literature usually acknowledges the usefulness of the traditional model, especially in market-exchange situations. But in the context of this chapter it has to be recognised that environmental issues are often subject to strong moral views.[19] As Frey (1999) summarises, contrary to the utilitarian perspective, a moralist views the environment with unique value and believes that it should be protected purely for ethical reasons and with little or no regard to any trade-offs. Furthermore, Arrow (1986) argues that rationality as the economist defines it often only occurs through market exchange, rather than through the intrinsic motivation of an individual. However, as Shogren and Taylor (2008) point out, environmental resources are frequently not subject to the market exchange required for consistent choices. Furthermore, the approach taken in this chapter puts forward that when individuals realise that the government is not able to set the incentives correctly, they may also recognise that the result will be far from optimal for themselves and everybody else without a change in behaviour. This may, in turn, lead to moral motivation that requires a different type of model to explain. Along the lines of Harsanyi (1955), Sen (1977) develops the idea that people may have two preference relations working at the same time.[20] Utility-maximising behaviour with some consideration for the utility of others he labels as 'sympathy', while actions driven only by their moral value with no consideration of utility he calls 'commitment'.[21] The latter is very much in line with Kant's concept of 'duty'.[22]

Laffont (1975) develops the idea of a Kantian approach to behaviour where individuals consider what would be optimal if they and everybody else were to make the same choice.[23] However, Laffont (1975) assumes that individuals are identical, with all individuals applying this rule equally. In his model this implies that individuals will correctly

[19] Also see Bergstrom (2009) for a discussion of the links between moral rules and utility.

[20] The model developed in this chapter can also be thought of as capturing the idea of two different preference sets to some extent. However, combining two preference sets into a conventional model usually assumes that one supersedes the other or individuals follow certain rules (see Thaler (1980) for an example). The model in this chapter assumes that individuals trade-off moral preferences against utility preferences depending on their propensity to act morally.

[21] Sen (1977) also points out that the resulting behaviour may well be the same as the utility-maximising one, but that this may not have been the individual's motivation.

[22] White (2003) provides a good summary of the essentials of Kant's philosophy as relevant to an economist and discusses linking it to the traditional model of economic behaviour.

[23] Such an approach (a version of which is also used in our model) is generally inspired by Kant's Formula of Universal Law (a version of the categorical imperative) which states: "Act only according to that maxim whereby you can at the same time will that it become a universal law." (Kant (1875), p. 421).

anticipate that everybody else will indeed behave the same way, meaning that ultimately individuals still maximise their personal utility. The model developed in this chapter differs to that of Laffont (1975) in that individuals have no expectations regarding the behaviour of others and they assume that all others continue to behave in the utility-maximising way. Individuals therefore recognise that acting morally means they will suffer a loss of personal utility. The central aspect of this approach is that individuals assess the intrinsic 'moral value' of their action by comparing the utility they actually get to the utility they would get if everyone else were to make the same choice as this individual.[24] Individuals then trade off this hypothetical 'moral value' of the action against the associated Utility Cost when making their consumption choice.

Ulph (2006) sets up a simple model of consumption choice where individuals have some propensity to act morally. It assumes that people determine the intrinsic 'rightness' of their action by comparing the utility they actually get to the utility they would get if everyone else were to make the same consumption choice as this individual. In that, individuals recognise that their action will not influence the behaviour of others in any way and thus the individual will get a loss in utility compared to the conventional economic choice. In making the consumption decision, individuals put some weight to the 'rightness' of the choice and trade this off against the utility loss incurred from deviating from the utility-maximising level. The model developed here is based on and an extension of this approach. It generalises the approach taken by Ulph (2006) and includes a measure for the size of the population whereas Ulph (2006) normalised the population to 1. This is done to highlight the impact of the population size on total emissions and therefore the damage caused. Furthermore, the propensity to act morally and the way an individual with propensity to act morally makes the consumption choice is captured in a different way compared to Ulph (2006).

In addition to the concept of 'moral value', the model developed in this chapter also considers that individuals may have altruistic concern for others' utility (Pure Altruism). In line with the definition by Hammond (1987), the model assumes that people may exhibit altruistic concern for the utility of all other individuals as a collective. Specifically, the model assumes that an individual's utility function is the sum of their direct utility from consumption and the total utility of all other individuals weighted by the degree of altruistic concern.[25] While Johansson (1997) also models this form of Pure Altruism,

[24] This type of 'Kantian' calculation is similar to the one used by Brekke et al. (2003), although they link the moral value of this Kantian behaviour to a concern for self-image instead.

[25] Although for simplicity reasons it is assumed that the degree of altruistic concern is the same for all individuals in the population, it is possible to generalise the result to allow for varying degrees of

the model developed here explicitly models a continuum of individuals in order to capture the atomistic nature of individuals' consumption choice in the context of large-scale environmental problems such as climate change. Such an individual would recognise the atomistic nature of their choice and hence, when making their consumption decision, takes average and total levels of consumption, welfare as well as damage from the externality as given. It is noteworthy that both the discrete and continuous approaches are widely used in the literature depending on the issue to be addressed. Finally, by modelling a combination of two different types of altruism, 'Pure Altruism' (the concern for others' utility) and 'Genuine Altruism' (the propensity to act morally through non-utility-maximising behaviour), the model captures the idea that people may be concerned about the welfare of others but may also have some propensity to act morally despite their inability to influence the impact of the environmentally harmful good.

Of course a distinction between Kant's categorical imperative and this model is that Kant sees moral duties as absolute that completely supersede any considerations of utility. The model developed in this chapter, on the other hand, does not use such a binary approach, but allows the individual's propensity to act morally to determine in how far they are driven by the moral value of the action relative to the associated loss in utility. Only individuals with full propensity to act morally will completely disregard their own utility and act in a fully Kantian fashion. At the same time, individuals with no propensity to act morally will not take into account the moral value at all and thus act in a standard utility-maximising way. The propensity to act morally is assumed to be an exogenous parameter that is distributed across different types of individuals in the economy. Since no assumption is made regarding what this distribution may look like, this approach does not argue that people necessarily have a certain propensity to act morally but the aim is simply to evaluate the consequences for the socially optimal tax on an environmentally harmful good if some people do act this way.

3.3 The Model

3.3.1 Standard Theory

Start with a population consisting of a continuum of potentially different types of individuals. These types are indexed by $k \in [0,1], 0 \leq k \leq 1$. The distribution is given by

altruism for different types of individuals without affecting the main conclusions from our analysis.

the density function

$$f(k) > 0, \quad k \in [0,1]; \qquad \int_0^1 f(k)dk = 1.$$

Each individual has the same initial endowment of income $y > 0$ and chooses between a 'clean' good x and a 'dirty' good z. The clean good is a numeraire good with a price of 1 and therefore represents the expenditure on all other goods but z. Its consumption is assumed to generate no externalities. On the other hand, the dirty good generates one unit of emissions per unit consumed, which is a negative externality to all individuals. Further, let $\zeta(.)$ denote the function that assigns to an individual of type k their chosen consumption of the dirty good, $\zeta(k) \geq 0$, and let $\bar{z}$ denote the average consumption of the dirty good. The size of the population is measured by $M > 0$ and therefore total emissions E are

$$E = M\bar{z}, \quad \text{where} \quad \bar{z} = \int_0^1 z_k f(k)dk. \tag{3.1}$$

Each individual derives utility from the personal consumption of the two goods. For simplicity we assume that preferences over the two goods are the same across all types and as such all individuals have identical utility functions.[26] Additionally, utility is assumed to be linear in consumption of the clean good in order to avoid issues of income distribution.[27] This means as long as consumers always consume a positive amount of the clean good, the marginal utility of income is constant and welfare losses that may arise are not due to inequality but inefficiencies. Utility derived from the consumption of the two good takes the following form:

$$\begin{aligned} & u(x,z,E) = x + \phi(z) - D(E), \\ & \text{where} \quad \phi'(z) > 0; \quad \phi'' < 0; \\ & \text{and} \quad \forall E > 0, \quad D'(E) > 0; \quad D''(E) \geq 0. \end{aligned} \tag{3.2}$$

The damage from emissions experienced by individuals is captured by the damage function $D(E)$. It is a strictly increasing and convex function for all positive levels of emissions. Furthermore, $\phi(z)$ describes the private gross benefit gained from consumption of

[26] Although individuals have identical utility functions, different types of individuals will later be defined by their propensity to act morally.

[27] Income distribution is of course an important element in the analysis of environmental externalities. However, in order to isolate the effect of altruism and moral behaviour on individuals' consumption choice, issues of income distribution have been kept out of this analysis.

the dirty good and is strictly increasing and strictly concave in z.

The dirty good is produced by a perfectly competitive industry operating with constant unit costs $c > 0$. At the same time the government imposes an emissions tax $t \geq 0$ on the consumption of z. The tax revenue is redistributed to the individuals through a lump-sum transfer σ that is identical for all individuals. Therefore the government budget constraint is defined by

$$\sigma = t\bar{z}. \tag{3.3}$$

Using the definition of emissions provided in (3.1) and the government budget constraint, we can now derive the utility from personal consumption and emissions as a function that only depends on the individual's consumption level of the dirty good, and the average consumption of the dirty good across the population:

$$u(z; \bar{z}, t) = (y + t\bar{z}) - (c + t)z + \phi(z) - D(M\bar{z}) \tag{3.4}$$

In line with the idea of atomistic consumption, in all stages of the analysis we use the standard Nash assumption that the individual always takes the consumption by everyone else as given when making the choice over consumption of the dirty good. This means that the individual treats average emissions as independent of z because their choice of consumption of the dirty good has no influence on average and total emissions. Indeed, based on the definition of total emissions as shown in (3.1), even if every other individual of the same type as this one were to change their consumption simultaneously, it would still not impact the level of average or total emissions.

Using this basic setup the individual chooses z to maximise their direct personal utility as given by (3.4). Under standard utility-maximising behaviour the chosen level of consumption of the dirty good $\tilde{z}(t)$ can be derived from

$$\tilde{z}(t) = \underset{z}{ArgMax} \quad u(z; \bar{z}, t). \tag{3.5}$$

Therefore the chosen level of consumption for the utility-maximising individual is characterised by

$$\phi'[\tilde{z}(t)] = c + t. \tag{3.6}$$

As we would expect, the left hand side of (3.6) describes the marginal gross benefit of consumption of the dirty good and the right hand side describes the private marginal cost of consumption. Further, as we would expect given the atomistic nature of the consumption choice in this model, the damage incurred from the dirty good is not a factor in determining the chosen consumption level under standard theory.

Given this we can determine that for any tax rate t and any dirty good assignment function $\zeta(.)$, social utility across all individuals in the population is

$$S(\zeta, t) = M \int_0^1 u[\zeta(k), \bar{z}, t] f(k) dk, \qquad \text{where } \bar{z} = \int_0^1 \zeta(k) f(k) dk \tag{3.7}$$

Since individuals have identical, and strictly concave utility functions, the socially optimal consumption level requires that everyone consume the same amount of the dirty good. This common level, $\hat{z}$, is defined as

$$\hat{z} = \underset{z}{ArgMax}\Big[S(\zeta, t)\Big] = \underset{z}{ArgMax}\Big[y - cz + \phi(z) - D(Mz)\Big] \tag{3.8}$$

As we can see, the social planner takes account of the link between the taxes paid on the dirty good and the lump-sum transfer received by consumers through the government budget constraint. This means that the socially optimal level of consumption is independent of the tax rate t. In addition, the social planner also takes into account the connection between the consumption of the dirty good and total emissions. As we would expect, this means that the social planner can fully internalise the externality. The socially optimal level of consumption of the dirty good $\hat{z}$ is then implicitly defined by

$$\phi'(\hat{z}) = c + MD'(M\hat{z}), \quad \text{where} \quad \hat{z} < \tilde{z}(0). \tag{3.9}$$

This shows that, from a social planner's perspective, the social marginal cost of consumption consists of the direct marginal cost of production of the dirty good, as well as the social marginal damage created by the emissions externality, $MD'(M\hat{z})$, but not the tax on the dirty good since the revenues are entirely redistributed to the individuals.

From (3.6) and (3.9) it is also straightforward to see that the government can achieve the socially optimal consumption level by setting the tax on the dirty good equal to the social marginal damage of consumption (the standard Pigovian tax). Denoting this optimal tax rate by $\hat{t}$, it follows that

$$\hat{t} = MD'(M\hat{z}) \tag{3.10}$$

So far we have only described the basic model setup and laid out the results under standard theory. These will serve as counterfactual for the remaining analysis.

3.3.2 Pure Altruism

As discussed in the introduction, Pure Altruism in this model means that individuals may put some weight on the total utility of all other individuals in the population. Individuals will therefore not only maximise their own private utility from consumption, but the (weighted) sum of their own private utility and the total utility of all others. For simplicity it is assumed that the degree of altruistic concern does not vary across different types of individuals.[28] The weight given to the total utility of others is denoted $\alpha \geq 0$ for all k. Then the total welfare of an individual who consumes an amount z of the dirty good is [29]

$$w(z; \zeta, t, \alpha) = U(z; \bar{z}, t) + \alpha S(\zeta, t), \qquad \text{where } \bar{z} = \int_0^1 \zeta(k) f(k) dk. \tag{3.11}$$

Given the atomistic nature of consumption, the total utility of all other individuals in the population is simply equal to the social utility as defined in (3.7). Furthermore, individuals still treat the consumption of all other individuals as given when making their consumption choice over the dirty good. This means individuals treat the assignment function $\zeta(k)$, and hence the average consumption level $\bar{z}$, as well as the level of social utility $S(.)$, as given. Indeed this means that individuals do not just take as given the damage they themselves suffer from the externality, but also the damage that everyone else suffers. It follows that, with atomistic consumption, if the individual chooses the level of consumption of the dirty good that maximises total individual welfare, then, independent of the level of altruistic concern, consumption will be the same as the level

[28] This assumption makes it more straightforward to isolate the key drivers of behaviour when we develop the model of moral behaviour. It is, however, possible to allow the degree of altruism to vary across different types without altering the key results of this chapter.

[29] The utility that includes the effect of altruistic concern is referred to as 'welfare'. This is simply done to separate the direct type of utility from the altruistic type.

that maximises their utility as characterised by (3.6).

Since we assumed a constant level of α for all types of individuals, we still have identical personal welfare functions for all individuals. Using (3.11) it is then straightforward to derive that, for any given tax rate t and assignment function $\zeta(.)$, the total welfare of all individuals, or social welfare, is given by

$$W(\zeta, t, \alpha) = (1 + \alpha M)S(\zeta, t). \tag{3.12}$$

This shows that the level of altruistic concern across the population only scales up the total level of welfare, but the socially optimal level of consumption is still one in which everyone consumes the same amount of the dirty good as determined by standard theory. Consequently the level of $\hat{z}$ is still defined by (3.9) and the optimal tax inducing everyone to consume the socially optimal level is still described by (3.10). This result leads us to the first proposition.

Proposition 1 *In a world of atomistic consumption, with every individual choosing their consumption by maximising their utility, altruistic concern for the utility of others has no impact on the optimal level of consumption for the individual, the socially optimal level of consumption or the socially optimal tax level.*[30]

Proof: By maximising individual welfare as shown in (3.11), it is straightforward to derive that $\phi'(z) = c + t$, which is the same as derived without altruism in (3.6). Similarly, maximising the social welfare function as shown in (3.12), we derive that $\phi'(\hat{z}) = c + MD'(M\hat{z})$,the same level as under standard theory in (3.9). Combining these two findings it is also evident that $\hat{t} = MD'(M\hat{z})$ induces everyone to consume the same socially optimal amount, the same as shown under standard theory without altruism.

So far this chapter has established the results with and without altruistic concern for the utility of others based on standard utility-maximising behaviour. The next section will now turn to the key part of this chapter and develop the theory of moral behaviour.

[30] Note that this result does depend on the functional form used for the individual's welfare function described in (3.11). However, this result can be generalised to other functional forms, for example the case where individuals' welfare takes the form $w(z; \zeta, t, \alpha) = U(z; \bar{z}, t)[S(\zeta, t)]^{\alpha}$.

3.3.3 Moral Behaviour

We start by assuming that initially everyone consumes $\tilde{z}(t)$, which is the chosen consumption level consistent with standard theory as derived in Section 3.3.1. And since the aim of the model is to investigate whether moral behaviour can make up for a shortfall in government policy, we further begin with the assumption that the tax on the dirty good is below the socially optimal level, i.e. $t < \hat{t}$, and therefore we also have $\tilde{z}(t) > \hat{z}$. The model of moral behaviour developed here assumes that individuals recognise that if they deviate from the consumption level $\tilde{z}(t)$, they will incur a loss in personal welfare. This means that individuals still know that the rational choice for them and everyone else would be to choose $\tilde{z}(t)$ (given the atomistic nature of consumption). However, they may also be driven by moral concerns for the environment and recognise that the overall outcome will be far from optimal for themselves and everyone else as long as the government continues to set the wrong tax. This may induce some individuals to act differently. Such an individual may measure a hypothetical moral value of deviating from the standard consumption level without expecting any personal benefit from this moral action.[31]

The model measures the hypothetical moral value of the choice through a type of 'Kantian' calculation as discussed in Section 3.2. This means that an individual with some propensity to act morally will assess how much better off they and everyone else would be if they and everyone else were to choose the same level of consumption, z.[32] As already mentioned, the individual recognises that deviating will entail a cost to personal welfare, but may balance off this cost against the moral value, depending on their propensity to act morally. Let us assume that different types of individuals may differ in their propensity to act morally. Some individuals may only be concerned with the moral value of their action and not take any account of the loss of personal welfare they will incur. Yet other individuals may give no weight to the moral value but fully take account of the personal welfare cost associated with deviating from $\tilde{z}(t)$. In addition to the propensity to act morally, the model recognises that individuals may still exhibit altruistic concern for others' utility as modelled in the previous section. The previous results have shown that altruistic concern does not have any influence on the consumption choice under utility-maximising behaviour so the question arises whether this changes when individuals have

[31] It is crucial to emphasize that moral value does not mean the individual derives any benefit (or utility) from the moral action.

[32] This is of course a simplification of Kant's Formula of Universal Law. For the purposes of this analysis, and in line with other literature using a Kantian approach as discussed in Section 3.2, we translate the categorical imperative in the sense that a Kantian calculation asks what would be optimal if everyone acted the same way (also see Laffont (1975); Brekke et al. (2003)).

some propensity to act morally.

The first step is to quantify the level of welfare loss associated with deviating from the initial consumption level. Let $\zeta(t)$ be the assignment function that assigns everyone the initial level of consumption $\tilde{z}(t)$. Therefore the average consumption of the dirty good is also $\tilde{z}(t)$. For an individual of type k who gives altruistic weight $\alpha \geq 0$ to the utility of other individuals, this gives an initial welfare level of

$$\begin{aligned} w(\tilde{z}(t);\zeta(t),t,\alpha) =& v(\tilde{z}(t),t) + \alpha M\Big[y - c\tilde{z}(t) + \phi(\tilde{z}(t)) - D(M\tilde{z}(t))\Big] \\ =& \Big[(y + t\tilde{z}(t)) - (c+t)\tilde{z}(t) + \phi(\tilde{z}(t)) - D(M\tilde{z}(t))\Big] \\ & + \alpha M\Big[y - c\tilde{z}(t) + \phi(\tilde{z}(t)) - D(M\tilde{z}(t))\Big] \\ =& (1+\alpha M)\Big[y - c\tilde{z}(t) + \phi(\tilde{z}(t)) - D(M\tilde{z}(t))\Big]. \end{aligned} \tag{3.13}$$

However, if the individual chooses a different level of consumption such that $\hat{z} \leq z \leq \tilde{z}(t)$, but everyone else carries on consuming $\tilde{z}(t)$ this generates the following level of welfare for the individual:

$$\begin{aligned} w(z;\zeta(t),t,\alpha) =& U(z;\zeta(t),t) + \alpha M\Big[y - c\tilde{z}(t) + \phi(\tilde{z}(t)) - D(M\tilde{z}(t))\Big] \\ =& \Big[(y + t\tilde{z}(t)) - (c+t)z + \phi(z) - D(M\tilde{z}(t))\Big] \\ & + \alpha M\Big[y - c\tilde{z}(t) + \phi(\tilde{z}(t)) - D(M\tilde{z}(t))\Big]. \end{aligned} \tag{3.14}$$

Combining (3.13) and (3.14) leads us to the loss in personal welfare of choosing a different level of consumption of the dirty good, z, as opposed to the initial level, $\tilde{z}(t)$:

$$\begin{aligned} C(z;\tilde{z}(t),t) &= w(\tilde{z}(t);\zeta(t),t,\alpha) - w(z;\zeta(t),t,\alpha) \\ &= \Big[\phi(\tilde{z}(t)) - (c+t)\tilde{z}(t)\Big] - \Big[\phi(z) - (c+t)z\Big]. \end{aligned} \tag{3.15}$$

Let us denote this as the Utility Cost. It is noteworthy that in the calculation of the Utility Cost everything that depends on what the other individuals do, and therefore anything associated with the level of altruistic concern, has cancelled out. In addition, since an individual's choice has no impact on the total level of emissions, the damage function has also cancelled out. The Utility Cost is purely driven by the difference in the private benefit and cost elements of the dirty good. Were the individual to minimise this cost they would choose $\phi'(z) = c + t$ and then we would be back at the standard level of consumption where $z = \tilde{z}(t)$. Intuitively, in order to minimise the cost of deviating from the initial consumption level, one would not deviate at all.[33]

The next step is to derive a measurement of the hypothetical moral value of choosing a different level of consumption. As already mentioned, this moral value is assessed by evaluating the welfare the individual would obtain if they and everyone else were to choose that same level of consumption. For an individual who places an altruistic weight $\alpha \geq 0$ on the welfare of all other individuals this would yield

$$w(z, \zeta^K(z); t, \alpha) = (1 + \alpha M)\Big[y - cz + \phi(z) - D(Mz)\Big], \tag{3.16}$$

where $\zeta^K(z)$ is the 'Kantian' function that assigns everyone else the same level of consumption of the dirty good as that chosen by the individual. There are two factors the individual will incorporate when making this calculation. First, the individual sees that if they and everyone else choose the same level of z, then the lump-sum transfer to the individual from the government tax revenues will equal the tax paid on the dirty good, rendering the tax rate irrelevant for the level of welfare obtained. Second, the individual takes account of the impact the choice of z has on total emissions and therefore the damage associated with it. Since it is a specific calculation of the level of welfare if everyone were to choose the same level of the dirty good, the individual effectively makes the same calculation a social planner would make. This means that the individual does not just evaluate the moral value for themselves, but also the benefit to everyone else in the population. Therefore, the hypothetical moral value of choosing a level of consumption, z, as opposed to the initial level, $\tilde{z}(t)$, is

$$\begin{aligned} B(z; \tilde{z}(t), t, \alpha) =& w(z, \zeta^K(z); t, \alpha) - w(\tilde{z}(t); \zeta(t), t, \alpha) \\ =& (1 + \alpha M)\Big\{ \Big[y - cz + \phi(z) - D(Mz)\Big] \\ & - \Big[y - c\tilde{z}(t) + \phi(\tilde{z}(t)) - D(M\tilde{z}(t))\Big]\Big\}. \end{aligned} \tag{3.17}$$

Let us denote this as the hypothetical Moral Benefit of deviating from the initial level. From here it is easy to see that were an individual to simply maximise the hypothetical Moral Benefit, the individual would choose the socially optimum level of consumption $\hat{z}$

[33] Even if the entire population acted entirely moral and consumed the social optimum regardless of the tax level, a single utility-maximising individual would minimise the Utility Cost defined in (3.15) and consume $\phi'(z) = c + t$ regardless of what others do. This is due to atomistic nature of the consumption decision.

regardless of their level of altruistic concern for the utility of others.

We can now put together the hypothetical Moral Benefit and the Utility Cost of deviating from the initial level by assuming an individual will make the consumption choice putting some weight on the moral value of deviating and balancing this off against the associated loss in personal welfare. Let this propensity to act morally be measured by μ.[34] To begin, the model will be developed with a linear decision function where individuals choose their consumption by maximising the weighted difference between the moral value and the associated personal welfare cost:

$$\mu B - (1-\mu)C, \qquad 0 \leq \mu \leq 1. \tag{3.18}$$

Substituting (3.15) and (3.17) into (3.18) we see that for an individual with $0 \leq \mu \leq 1$, consumption of the dirty good will be chosen to

$$\underset{z}{MAX} \quad \mu(1+\alpha M)\Big[y - cz + \phi(z) - D(Mz)\Big] + (1-\mu)\Big[\phi(z) - (c+t)z\Big]. \tag{3.19}$$

At this point it is noteworthy that, due to the linear nature of the choice function, the initial level of consumption $\tilde{z}(t)$ is irrelevant to the consumption choice and the initial level of consumption could indeed be at any level without influencing the resulting consumption choice under Kantian behaviour. When we explore the generalised choice function later in this chapter and a couple of non-linear examples, we will see that this will change under different assumptions. Now, if we let

$$k = \frac{\mu(1+\alpha M)}{1+\mu\alpha M}, \qquad 0 \leq k \leq 1, \tag{3.20}$$

then it is straightforward to derive from (3.19) that the individual's consumption choice can be characterised by

$$\phi'(z) = c + \Big[kMD'(Mz) + (1-k)t\Big]. \tag{3.21}$$

[34] This propensity to act morally is assumed to be an exogenously given parameter. Therefore, in this setting the degree of morality is not influenced by what others do and it is not a parameter that is chosen with utility considerations. Indeed, it captures the individual's willingness to sacrifice utility in order to do 'the right thing' regardless of what others do. However, as highlighted in the discussion in Section 3.8, it may be interesting to look more closely at what drives the level of μ and how it may change over time or be influenced by policy choices.

The parameter k is a combined parameter capturing both the individual's propensity to act morally μ, and the level of altruistic concern for others' utility α. This parameter can be thought of as the overall level of 'virtue' individuals may exhibit, where $0 \leq k \leq 1$. Examining (3.21), we see that if $k = 1$ the individual will choose a level of consumption equal to the socially optimal level $z = \hat{z} = c + MD'(Mz)$. On the other hand, if $k = 0$ the individual will choose a consumption level of the dirty good equal to the conventional choice, $z = \tilde{z}(t) = c + t$. Of course the level of k can be anywhere between 0 and 1 and the level of z will vary accordingly between $\tilde{z}(t)$ and $\hat{z}$. Figure 3.1 illustrates an example level of consumption for an individual with a value of k such that $0 < k < 1$.

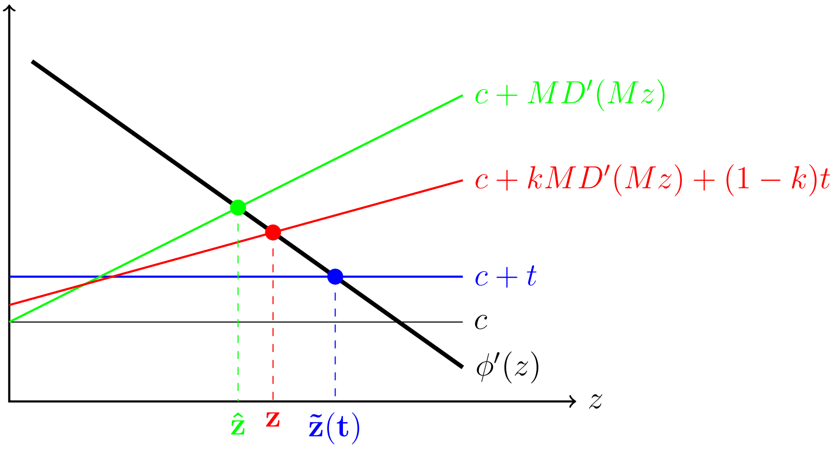

Figure 3.1: Consumption Choice for Individuals with Some Propensity to Act Morally

Next let us take a closer look at the determinants of k. Suppose the individual has no altruistic concern for others' utility ($\alpha = 0$). Then it follows that $k = \mu$. This means that for an individual with no altruistic concern, behaviour only depends on their propensity to act morally. However, if $\mu = 1$ the individual will still cut down consumption of the dirty good to $\hat{z}$ regardless of the level of α. Indeed we find that even if $\alpha = 0$, an individual with $\mu > 0$ will choose a lower consumption level of the dirty good relative to the standard level. It follows that altruistic concern for others' utility is not a necessary component of this type of behaviour.

Now suppose that the individual has no propensity to act morally ($\mu = 0$). Then, it is straightforward to see that $k = 0$ regardless of the value of α. Effectively, if $\mu = 0$ we are back to the case where individuals choose consumption of dirty good by maximising their welfare function which means that individuals will choose $\tilde{z}(t)$ regardless of their level of altruistic concern. Using these insights we can set out the next proposition.

Proposition 2 *Altruistic concern for others' utility is neither necessary nor sufficient to induce people to cut back consumption of the dirty good, but a propensity to act morally is both necessary and sufficient.*

Proof: From (3.21) we know that z is a decreasing function of k for any $t < \hat{t}$. From (3.20) we can further determine that if $\mu = 1 \rightarrow k = 1$, if $\mu = 0 \rightarrow k = 0$, and $\frac{\partial k}{\partial \mu} = \frac{1+\alpha M}{(1+\mu\alpha M)^2} > 0$ for any value of α. Therefore α is neither necessary nor sufficient, but μ is both necessary and sufficient to induce a decrease in z.

Let us now look at the case where individuals have some, but not full propensity to act morally ($0 < \mu < 1$). Then k is an increasing function of α. This means that the level of altruistic concern can affect the consumption choice even though people recognise that their consumption of the dirty good has no effect on total emissions and anybody else's welfare. This is because in calculating the hypothetical Moral Benefit of the action, individuals take into account their level of altruistic concern for others' utility and therefore the impact their choice will have on everybody else's utility, but the associated Utility Cost of the consumption choice is independent of the level of altruistic concern. Yet we should also note that, even if $\alpha = 1$, as long as $\mu < 1$, an individual will never cut down consumption of the dirty good all the way to $\hat{z}$. This leads us to the next proposition:

Proposition 3 *If individuals have some, but not full propensity to act morally, an individual with a higher level of altruistic concern for others will reduce consumption of the dirty good if the government sets the tax on the dirty good too low, but never all the way to the socially optimal level.*

Proof: From (3.20) we can derive that $\frac{\partial k}{\partial \alpha} = \frac{\mu(1-\mu)M}{(1+\mu\alpha M)^2} > 0$ for any $0 < \mu < 1$. Therefore k is an increasing function of α. Furthermore, we know from (3.21) that z is a decreasing function of k for any $t < \hat{t}$. Therefore an increase in α will lead to a reduction in the consumption of the dirty good for any $0 < \mu < 1$. However, to achieve $z = \hat{z}$, we require $k = 1$. As per the definition of k in (3.20), this is only the case when $\mu = 1$ and therefore $z = \hat{z}$ can never be achieved for any $0 < \mu < 1$ regardless of the value of α.

Next, suppose that the individual has some level of altruistic concern and some propensity to act morally (i.e. $\alpha > 0$, and $\mu > 0$). Then we can see that k is also an increasing function of the size of the population, M. It shows that the larger the population of people affected, the more people will cut back their consumption of the dirty good. The size of the population does not just influence the level of k though. Taking a closer look at (3.21) it is easy to see that for individuals with some propensity to act morally (i.e. $k > 0$), there is another channel through which an increase in M will cause individuals to cut back their consumption of the dirty good. An increase in M increases the social marginal damage of consumption and therefore reduces the socially optimal level of consumption $\hat{z}$.[35] The population size M impacts the social marginal damage because, (a) it increases the number of people affected, and (b) it increases the total level of emissions which leads to increasing marginal damage since damage is convex in z (i.e. $D''(.) \geq 0$). We can now state the following general proposition.

Proposition 4 *If the government fails to set the tax that would be optimal given the usual economic behaviour and the usual conception as to what constitutes welfare, then, to the extent that individuals have some propensity to act morally, private action will to some extent compensate for the lack of government action and drive consumption of the dirty good down towards the optimal level.*

Proof: For any $0 < \mu \leq 1$ we have $0 < k \leq 1$ as per the definition of k in (3.20). Using this and comparing (3.21) with (3.6) and (3.9), it is straightforward to see that we must have $\phi'(\hat{z}) \geq \phi'(z) > \phi'[\tilde{z}(t)]$ and therefore we also know that $\hat{z} \leq z < \tilde{z}(t)$.

We further know that although people may behave according to a different set of rules, individuals' welfare is still determined by the altruistic welfare function as defined in (3.11). The government recognises this and therefore still uses the same social welfare function given in (3.12) to determine the socially optimal allocation of welfare. Similarly, the social utility function $S(.)$ is also still the same as described in (3.7). It is then straightforward to see that the socially optimal tax level for the government is still the standard Pigovian tax rate $\hat{t}$, the same that would be optimal if the population had no propensity to act morally (i.e. $\mu = 0$ for all types). Indeed, by examining (3.21) while plugging in the Pigovian tax as shown in (3.10), we can easily determine that each individual will choose consumption level $\hat{z}$ regardless of the level of k. This means with a tax rate equal to $\hat{t}$ we can still achieve the first-best solution where everyone consumes the same amount of the dirty good and the damage is fully internalised. This leads us to the final proposition.

[35] This impact of M on the socially optimal level of consumption is of course not an exclusive feature of moral behaviour as modelled here, but can also be observed in the characterisation of the socially optimal level of consumption as shown in (3.9).

Proposition 5 *Although individuals may to some extent compensate for the government's failure to set the optimal tax, this does not imply that the government should not set the optimal tax.*[36]

Proof: From (3.21) we can derive that by increasing t to $\hat{t} = MD'(M\hat{z})$ everyone will consume $\hat{z}$ as characterised by (3.9). This raises social welfare because (a) aggregate consumption of the dirty good is socially optimal and (b) consumption is equalised across consumers.

3.3.4 Government Sets the Tax Too High

Initially we assumed that the government sets the tax on the dirty good below the optimal level (i.e. $t < \hat{t}$). Now suppose that the government actually sets the tax rate too high, so $t > \hat{t}$ and therefore $\tilde{z}(t) < \hat{z}$. To illustrate this, consider Figure 3.2:

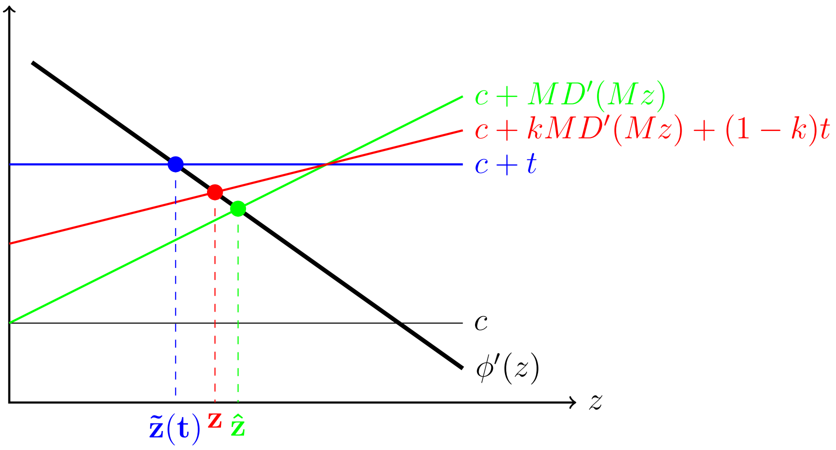

Figure 3.2: Consumption Choice with a Tax Rate $t > \hat{t}$

Because the initial level of consumption is now lower than the socially optimal level, an individual with some propensity to act morally will actually increase their consumption of the dirty good in order to bring it closer to the socially optimal level. The benchmark for an individual with some propensity to act morally is always the socially optimal level of consumption and depending on the weight they give to the hypothetical Moral Benefit, they will choose a consumption level of the dirty good that moves closer to $\hat{z}$. This point

[36] In the special case that the entire population has full propensity to act morally ($\mu = 1$ for the entire population), the tax would not be required since everybody would consume the socially optimal level regardless of the tax. However, even in that case, setting the optimal tax would not create any harm.

illustrates that the propensity to act morally does not necessarily imply a reduction in the consumption of an environmentally harmful good. Rather, an individual with some propensity to act morally will change consumption in order to get closer to the social optimum in the absence of the correct tax on the dirty good.[37]

Result 1 *If the government sets the tax too high (i.e. $t > \hat{t}$), individuals with propensity to act morally will increase their consumption of the dirty good toward the socially optimal level, $\hat{z}$.*

3.3.5 Ignoring the Government Budget Constraint

In the model developed so far individuals were able to take account of the government budget constraint when evaluating the hypothetical Moral Benefit of deviating from the initial consumption level. This meant that individuals recognise in the calculation of the hypothetical Moral Benefit that when they and everyone else were to choose z, this would render the tax rate irrelevant, because - as in the case of the social planner's optimum - if everybody consumes the same amount, the tax redistributed is exactly the same as the tax individuals have to pay. However, now suppose that individuals do not make this connection when evaluating the hypothetical Moral Benefit, but only assess the effect the choice will have on total and average emissions and the damage associated with this. This means they take the price of the dirty good $p = c + t$, as well as the lump-sum transfer from the government σ as given. Individuals then choose the consumption level of the dirty good by maximising

$$\underset{z}{MAX} \quad \phi(z) - (c+t)z - \Big[kD(Mz)\Big], \tag{3.22}$$

and therefore the consumption choice for the individual is described by

$$\phi'(z) = (c+t) + kMD'(Mz). \tag{3.23}$$

An individual who does not take account of the government budget constraint and has no propensity to act morally ($\mu = 0 \rightarrow k = 0$) will still not deviate from the initial consumption level $\tilde{z}(t)$. However, an individual with full propensity to act morally ($\mu = 1 \rightarrow k = 1$) will actually choose a consumption level of the dirty good lower than the socially optimal level ($z < \hat{z}$). Thus an individual with $k = 1$ actually overcompensates

[37] Buchholz et al. (2012) also analyse a case of public good provision where the subsidy may be too high.

in their consumption choice of the dirty good. This raises the question of what level of k is required to bring the consumption level to the socially optimal level. We can derive this level of k by looking at

$$\begin{aligned} c + MD'(Mz) &= (c+t) + kMD'(Mz), \\ k &= 1 - \frac{t}{MD'(Mz)}. \end{aligned} \tag{3.24}$$

The above shows that the required level of k depends on the ratio between the tax rate and the marginal damage. This is important because if the tax rate were to be equal to the Pigovian tax (i.e. $t = \hat{t}$), then only an individual with no propensity to act morally ($k = 0$) would consume $\hat{z}$. Any individual for whom $k > 0$ would choose a consumption level lower than the socially optimal level and therefore the individual's propensity to act morally leads to sub-optimal behaviour. Figure 3.3 illustrates this case:

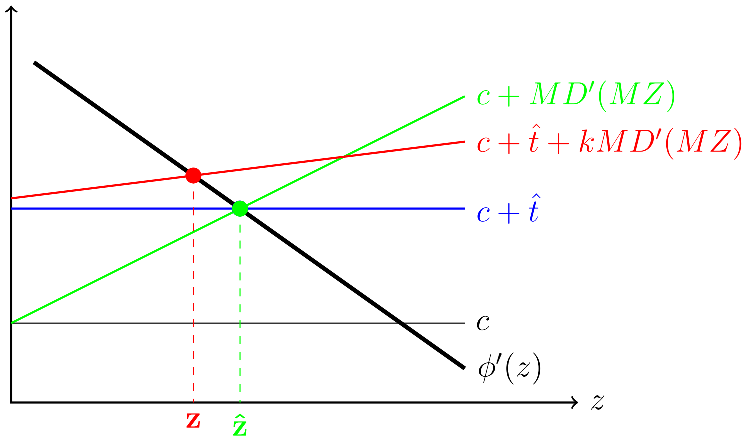

Figure 3.3: Consumption Choice when Individuals Ignore the Government Budget Constraint

The question arises what the optimal tax level would be in this case. From equation (3.24) we can see that for any given individual the tax rate that would induce an individual to choose the social optimum $\hat{z}$ is given by

$$\hat{t}_g = (1-k)MD'(Mz). \tag{3.25}$$

The optimal tax rate now is a function of the parameter k and we of course know that this can be different for different types of individuals. This means that because individuals fail to make the correct calculation in their assessment of the hypothetical Moral Benefit we now have to deal with two imperfections and as such the government can no longer achieve the first-best solution of everyone consuming $\hat{z}$ through a single tax on the dirty good.

Result 2 *If individuals fail to correctly take account of the government budget constraint in the calculation of the hypothetical Moral Benefit, individuals with full propensity to act morally will over-compensate and consume less than the socially optimal level. Furthermore, if the government sets the tax at* $t = \hat{t}$ *any individual with* $\mu > 0$ *will consume less than the socially optimal level.*

The first distortion is because individuals no longer consume the same amount of the dirty good even when $t = \hat{t}$ and the second distortion is because average consumption of the dirty good is not at the socially optimal level when $t = \hat{t}$. Indeed setting the tax rate at $\hat{t}$ could actually reduce social welfare compared to welfare under the initial tax rate $t < \hat{t}$. Additionally, the very fact that individuals aim to act morally but fail to do in the correct way, could lead to lower welfare than if individuals had no propensity to act morally. Given that in the presence of the behavioural failure the tax is not able to completely correct for the market failure (the emissions externality), the question arises whether the initial tax level or no tax at all is actually a preferred solution compared to setting the Pigovian tax. Shogren and Taylor (2008) discuss whether behavioural failures lead a behavioural-environmental second-best problem and whether policy interventions could realistically correct for both market and behavioural failures at the same time. They argue that it would be practically impossible to design such incentives since the policy designer would require significant further information (often at the individual level). Furthermore, there may be a range of behavioural failure some of which may separable, but not necessarily all of them. This implies that one would possible have to design a different incentive structure for each type and degree of behavioural failure, which is practically impossible. While we only have one behavioural failure in our example, it is still evident that a policy maker trying to correct for both the market and behavioural failure would require significant additional information, in this case information about the distribution of k, the parameter capturing how much 'virtue' an individual exhibits. This in turn would require information about the distribution of μ, the individuals' propensity to act morally. Note that we have not made any assumptions about the distribution to this point since it was not necessary to do so. The Pigovian tax was able to induce optimal behaviour regardless of individuals' propensity to act morally and there was no behavioural failure even though individuals are not acting in a utility-maximising way. This chapter will not conduct an analysis of the second-best solution as it is outside the scope of this chapter. The purpose of this example was to illustrate the importance of individuals' ability to make the correct calculation with regard to the hypothetical Moral Benefit and the Utility Cost associated with deviating from the utility-maximising level. Note that in this example the behavioural failure was with regard to the government budget constraint and the redistribution of the tax revenues to individuals. Given the difficulty of correcting both a market failure and a behavioural failure (or even identifying behavioural failures in the first place), Shogren and Taylor (2008) argue that it may be best to use instruments to correct market failures that are less likely to induce behavioural failures, for example a marketable permit system. While not a straightforward approach to use at an individual level, it may have the advantage of not requiring the individual to take into account the government's budget constraint and therefore avoiding the behavioural failure in the first place.

3.4 Generalised Choice Function

3.4.1 Analysis of Generalised Choice Function

The model of moral behaviour developed in the previous section uses a simple linear choice function where individuals maximise the weighted difference between the hypothetical Moral Benefit and the Utility Cost of deviating from the standard utility-maximising choice. This section generalises this choice function and analyses how far the previous results may change. However, it is still assumed that an individual chooses consumption of the dirty good by maximising some function of the hypothetical Moral Benefit of deviating from the standard utility-maximising consumption level as described in (3.17), and the associated Utility Cost as described in (3.15). Therefore the choice function is given by

$$F(C,B), \qquad \text{where} \qquad \frac{\partial^2 F(B,C)}{\partial z^2} < 0, \tag{3.26}$$

and

$$B = (1+\alpha M)\Big\{\Big[y + \phi(z) - cz - D(Mz)\Big] - \Big[y + \phi(\tilde{z}(t)) - c\tilde{z}(t) - D(M\tilde{z}(t))\Big]\Big\}, \tag{3.27}$$

$$C = \Big[\phi(\tilde{z}(t)) - (c+t)\tilde{z}(t)\Big] - \Big[\phi(z) - (c+t)z\Big]. \tag{3.28}$$

Obviously we require the choice function to be strictly concave in z and this means the second partial derivative with respect to z has to be negative. Knowing this will help to interpret some of the comparative statics analysis later in this section. The specific conditions that need to hold for the choice function to be concave are detailed in Section 3.4.2. Furthermore, it is straightforward to determine from the above that, consistent with the baseline model, we have

$$\frac{\partial B}{\partial z} = (1+\alpha M)\Big[\phi'(z) - c - MD'(Mz)\Big] \leq 0 \quad \forall \quad z \geq \hat{z}, \tag{3.29}$$

and

$$\frac{\partial C}{\partial z} = -\Big[\phi'(z) - (c+t)\Big] \leq 0 \quad \forall \quad z \leq \tilde{z}(t). \tag{3.30}$$

Of course we know that the individual will maximise the choice function over z in order to determine the consumption level. The required first-order condition therefore is given by

$$\frac{\partial F}{\partial B}\frac{\partial B}{\partial z} + \frac{\partial F}{\partial C}\frac{\partial C}{\partial z} = 0. \tag{3.31}$$

For notational simplicity, let

$$\frac{\partial F}{\partial B} = F^B, \qquad \text{and} \qquad \frac{\partial F}{\partial C} = F^C.$$

Determining F^B and F^C for the baseline linear choice function used in Section 3.3.3, we can immediately see how the generalisation links to the baseline model (i.e. $F^B = \mu$ and $F^C = -(1-\mu)$). Next, plugging (3.29) and (3.30) into (3.31) we get

$$F^B(1+\alpha M)\Big[\phi'(z) - c - MD'(Mz)\Big] = F^C\Big[\phi'(z) - (c+t)\Big]. \tag{3.32}$$

Now, analogous to the baseline model, if we let

$$k_g = \frac{F^B(1+\alpha M)}{F^B(1+\alpha M) - F^C} \qquad \text{where} \qquad 0 \leq k_g \leq 1, \tag{3.33}$$

we find that (3.32) becomes

$$\phi'(z) = c + k_g M D'(Mz) + (1-k_g)t. \tag{3.34}$$

From this it is immediately observable that (3.33) is the equivalent of (3.20) from the baseline model and (3.34) is the equivalent of (3.21). In general terms the 'virtue' parameter k_g determines how much weight in the consumption choice is given to the social marginal damage caused by the externality relative to the weight given to the tax which has a direct effect on utility derived. It is also straightforward to see that $k_g = 0$ when $F^B = 0$ and $k_g = 1$ when $F^C = 0$. This makes intuitive sense. For example, when the impact of a marginal change in the hypothetical Moral Benefit on the choice function is zero, then the individual will not take this benefit into account in making the consumption choice and thus the level of 'virtue' is zero, which in turn leads to a consumption level of $z = \tilde{z}(t)$. On the other hand, if the marginal impact of a change in the Utility Cost on the choice function is zero, the individual will not take account of the Utility Cost associated with the consumption choice and only use the hypothetical Moral Benefit. Thus $k_g = 1$ and the individual will choose $z = \hat{z}$. In the baseline model the parameter μ captured the individual's propensity to act morally, which was the key determinant of k and therefore of how much consumption would move towards the social optimum. The generalised choice function does not have a specific parameter to measure the individual's propensity to act morally. However, while the standalone values of F^B and F^C are of lesser importance, it is the relative levels of F^B and F^C that determine the weight given to the social optimum and therefore are the reflection of an individual's willingness to depart from the utility-maximising choice, or in other words, the individuals propensity to act morally. Note that in the baseline model we had $F^B = \mu$ and $F^C = -(1-\mu)$, and therefore F^B and F^C were directly related to each other with $-\frac{F^B}{F^C} = \frac{\mu}{1-\mu}$. In a more general choice function this does not necessarily have to be the case, but it is still the relative level that determines how far consumption will be changed towards the social optimum and therefore the term $-\frac{F^B}{F^C}$ captures the individual's propensity to act morally for any $F^B > 0$ and $F^C < 0$. At the same time, an individual with $F^C = 0$ and $F^B > 0$

has full propensity to act morally while and individual with $F^B = 0$ and $F^C < 0$ has no propensity to act morally.

Looking at (3.33) we can also see that the level of altruistic concern for the welfare of others is neither necessary nor sufficient for an individual to cut back consumption. Similar to the baseline model we have $k_g = 1$ if $F^B > 0$ and $F^C = 0$ regardless of the value of α. At the same time, if $F^B = 0$ and $F^C < 0$ we have $k_g = 0$ regardless of the value of α.[38] In the baseline model we also determined that if an individual has some but not full propensity to act morally, a higher level of altruistic concern would reduce consumption of the dirty good to some degree. However, with the more general setup we cannot unambiguously determine that the virtue parameter k_g is an increasing function of α. The reason for this will be addressed in more detail in Section 3.4.6. With regard to Proposition 3 from the baseline model we can only determine with certainty that even if k_g is an increasing function of α and therefore altruistic concern for the utility of others reduces consumption of the dirty good, consumption will not be cut to $\hat{z}$ regardless of the value of α.

Proposition 6 *Proposition 2 holds under any concave general choice function of the hypothetical Moral Benefit and associated Utility Cost, $F(B, C)$. Altruistic concern is neither necessary nor sufficient to induce people to reduce consumption of the dirty good, but a propensity to act morally is both necessary and sufficient. However, Proposition 3 does not necessarily hold in its entirety. Depending on the specification of the choice function it is possible that altruistic concern can either increase or decrease consumption of the dirty good. However, even when altruism reduces consumption, it is never sufficient to reduce consumption all the way to the socially optimal level regardless of the value of α.*

Proof: From (3.34) we know that z is a decreasing function of k_g for any $t < \hat{t}$. From (3.33) we can further determine that if $F^B > 0$ and $F^C = 0 \rightarrow k_g = 1$; and if $F^B = 0$ and $F^C < 0 \rightarrow k = 0$. In addition, if we define $R_C^B = -\frac{F^B}{F^C}$ for $F^B > 0$ and $F^C < 0$ we can rewrite (3.33) as $k_g = \frac{R_C^B(1+\alpha M)}{R_C^B(1+\alpha M)+1}$. Then we find $\frac{\partial k_g}{\partial R_C^B} = \frac{1}{[R_C^B(1+\alpha M)-1]^2} > 0$ for any value of α. Therefore α is neither necessary nor sufficient, but R_C^B is both necessary and sufficient to induce a decrease in z. Furthermore, from (3.33) we can derive that $\frac{\partial k_g}{\partial \alpha} = \frac{-F^C[F^B B \frac{\partial B}{\partial \alpha} + F^B M]}{[F^B(1+\alpha M)-F^C]^2}$. For any $F^B > 0$ and $F^C < 0$ this is only positive if $F^B > -F^{BB}B$ for any $\hat{z} < z < \tilde{z}(t)$. Therefore k is not unambiguously an increasing function of α and we cannot determine that an increase in α will necessarily lead to a

[38] As will be evident from the concavity conditions in Section 3.4.2, the definition of k_g in (3.33) also illustrates that at least one of F^B and F^C has to be non-zero.

reduction in the consumption of the dirty good when $F^B > 0$ and $F^C < 0$. However, to achieve $z = \hat{z}$ we require $k_g = 1$. As per the definition of k in (3.33), this is only the case when $F^C = 0$ and $F^B > 0$, and therefore $z = \hat{z}$ can never be achieved for any $F^B > 0$ and $F^C < 0$ regardless of the value of α.

Given the insights we have derived from the generalised model so far, it is easy to see from (3.34) that Proposition 4 also holds under a generalised choice function and indeed in the absence of the government setting the right tax, private action may to some degree compensate for the government's failure. Now if we let $t = \hat{t}$ as defined in (3.10), we can see that (3.34) reduces to $\phi'(z) = c + MD'(Mz) = \phi'(\hat{z})$ independent of the value of k_g. Thus a tax at the socially optimal level of $\hat{t} = MD'(Mz)$ induces the first-best solution where every individual consumes the same, and socially optimal, amount of the dirty good regardless of their propensity to act morally. This holds regardless of the functional form used for the concave choice function as long as both the hypothetical Moral Benefit and the Utility Cost are calculated in the correct way. Furthermore, it is straightforward to see that if either the Moral Benefit or the Utility Cost is not included in the choice function, we still obtain the same result whenever $t = \hat{t}$. This is an important result because it shows that the result of Proposition 5 is not dependent on the specific functional form of the choice function.

Proposition 7 *Both Proposition 4 and Proposition 5 hold under any concave general choice function of the hypothetical Moral Benefit and associated Utility Cost, $F(B, C)$. If the government fails to set the optimal tax, then individuals who put some weight on the Hypothetical Moral Benefit in their choice function will to some extent compensate for the wrong tax and reduce their consumption of the dirty good. However, although individuals may to some extent compensate for the government's failure to set the optimal tax, this does not imply that the government should not set the optimal tax, which is the standard Pigovian tax defined in (3.10).*

Proof: For any $F^B > 0$ and $F^C \leq 0$ we have $0 < k_g \leq 1$ as per the definition of k in (3.33). Using this and comparing (3.34) with (3.6) and (3.9), it is straightforward to see that we must have $\phi'(\hat{z}) \geq \phi'(z) > \phi'[\tilde{z}(t)]$ and therefore we also know that $\hat{z} \leq z < \tilde{z}(t)$. Furthermore, from (3.34) we can derive that by increasing t to $\hat{t} = MD'(M\hat{z})$ everyone will consume $\hat{z}$ as characterised by (3.9).

3.4.2 Conditions for Concavity of Choice Function

As stated above we require that $F(B, C)$ is a strictly concave function in z in order to obtain sensible results. In other words, we require that $\frac{\partial^2 F(B,C)}{\partial z^2} < 0$. We already know that the first-order condition is given by $F^B \frac{\partial B}{\partial z} + F^C \frac{\partial C}{\partial z} = 0$. From this we can derive the second-order condition as

$$F^{BB}\Big(\frac{\partial B}{\partial z}\Big)^2 + F^B \frac{\partial^2 B}{\partial z^2} + F^{CC}\Big(\frac{\partial C}{\partial z}\Big)^2 + F^C \frac{\partial^2 C}{\partial z^2} + (F^{BC} + F^{CB})\frac{\partial B}{\partial z}\frac{\partial C}{\partial z} < 0. \qquad (3.35)$$

Therefore the following conditions are sufficient to ensure strict concavity of the choice function for any $\hat{z} < z < \tilde{z}(t)$:

1. $F^B \geq 0$, $F^C \leq 0$, and $F^B F^C \neq 0$.
2. $F^{BB} \leq 0$, $F^{CC} \leq 0$.
3. $F^{BC} \leq 0$, $F^{CB} \leq 0$.
4. $\frac{\partial B}{\partial z} < 0$ and $\frac{\partial^2 B}{\partial z^2} < 0$.
5. $\frac{\partial C}{\partial z} < 0$ and $\frac{\partial^2 C}{\partial z^2} > 0$.

The first condition determines the that choice function is increasing in the hypothetical Moral Benefit and decreasing in the Utility Cost, one of which has to be strictly true. The next condition ensures that the choice function is concave in both the hypothetical Moral Benefit and the Utility Cost. The third condition ensures that the cross-partial derivatives of the choice function are also non-positive. This means that the rate at which the choice function increases in the hypothetical Moral Benefit is decreasing in the Utility Cost and that the rate at which the choice function decreases in the Utility Cost is increasing in the hypothetical Moral Benefit. Furthermore, from the earlier definitions of the hypothetical Moral Benefit and Utility Cost we already know that the last two conditions hold for any $\hat{z} < z < \tilde{z}(t)$, i.e. the hypothetical Moral Benefit is decreasing in z and strictly concave and the Utility Cost is decreasing in z and strictly convex.[39]

[39] This derives from the strict concavity of the private gross benefit of consumption $\phi(.)$ and the convexity of the damage function $D(.)$ as defined in (3.2).

3.4.3 Impact of a Change in the Initial Consumption Level

In Section 3.3.3 we saw that due to the linear nature of the choice function the initial level of consumption - in that case assumed to be the utility-maximising level $\tilde{z}(t)$ - had no impact on the chosen level of consumption. Indeed the initial level could be any value without any impact. However, this may not be true with a non-liner choice function. It is important to note here that while $\tilde{z}(t)$ will simply be referred to as the initial consumption level, it is also the level the individual assumes all others in the economy will consume. It is therefore the determinant of the damage experienced which in turn is a key determinant of the size of the hypothetical Moral Benefit. To explore the impact a change in the initial level $\tilde{z}(t)$ may have, we look at the partial derivate of z with respect to $\tilde{z}(t)$, which is given by

$$\frac{\partial z}{\partial \tilde{z}(t)} = \Big[-\frac{1}{SOC}\Big]\Big\{\Big[F^{BB}\frac{\partial B}{\partial z} + F^{CB}\frac{\partial C}{\partial z}\Big]\frac{\partial B}{\partial \tilde{z}(t)}\Big\} \geq 0 \qquad \forall \quad \hat{z} < z < \tilde{z}(t), \tag{3.36}$$

where SOC is the second derivative of the choice function with respect to z as shown in (3.35). From the concavity requirement of the choice function we know that $-\frac{1}{SOC} > 0$ and $F^{BB}\frac{\partial B}{\partial z} + F^{CB}\frac{\partial C}{\partial z} \geq 0$ for any $\hat{z} < z < \tilde{z}(t)$. Furthermore, it is straightforward to see from (3.27) that $\frac{\partial B}{\partial \tilde{z}(t)} > 0$ for any $\tilde{z}(t) > \hat{z}$ and thus we can determine that the chosen consumption level is increasing in the initial consumption level. Intuitively this is because the higher the initial level of consumption, the greater the hypothetical Moral Benefit given a chosen consumption level z (which is lower than the initial level but greater than the social optimum). This means that if the initial consumption level increases, the individual can proportionately achieve the required level of hypothetical Moral Benefit (given their specific choice function) with a higher consumption level.

However, we also see that the influence of $\tilde{z}(t)$ on z is only driven through a change in the hypothetical Moral Benefit, and not through the associated Utility Cost. This is because $\frac{\partial C}{\partial \tilde{z}(t)} = \phi'(\tilde{z}(t)) - (c+t) = 0$, since $\tilde{z}(t)$ is defined as the optimal consumption level under standard theory and thus $\phi'(\tilde{z}(t)) = c+t$. Intuitively this is because we know that a change in $\tilde{z}(t)$ has to be accompanied by a corresponding change in t (assuming c is held fixed) which will offset any change in the Utility Cost. Note that this is only the case because the initial level of consumption is assumed to be $\tilde{z}(t)$. If the initial level of consumption were at a different level, say at a level $z_0 \neq \tilde{z}(t)$, then we would have $\frac{\partial C}{\partial z_0} \neq 0$ and the impact of a change in the initial consumption level on the chosen level would be given by

$$\begin{aligned}\frac{\partial z}{\partial z_0} =&\Big[-\frac{1}{SOC}\Big]\Bigg\{\Big[F^{BB}\frac{\partial B}{\partial z}+F^{CB}\frac{\partial C}{\partial z}\Big]\frac{\partial B}{\partial z_0}\\ &+\Big[F^{BC}\frac{\partial B}{\partial z}+F^{CC}\frac{\partial C}{\partial z}\Big]\frac{\partial C}{\partial z_0}\Bigg\}\geq 0 \qquad \forall \quad \hat{z}<z, z_0 \leq \tilde{z}(t).\end{aligned} \tag{3.37}$$

We see from the above that the impact of an increase in z_0 is definitively to increase the chosen consumption level z of the dirty good as long as $\hat{z} < z, z_0 \leq \tilde{z}(t)$. Indeed, in this case, the effect of a change in the initial level is even stronger as there are now two components driving a change in the consumption choice. As before, an increase in the initial level has an impact on the chosen level through the hypothetical Moral Benefit, where a higher initial level allows the same level of Moral Benefit to be achieved with a higher level of consumption. In addition, a higher level of initial consumption - as long as it is less than the utility-maximising level - requires the individual to increase the chosen level in order to maintain the same level of Utility Cost (or the other way around, an increase in the initial level for a given level of the chosen level below the initial level, increases the Utility Cost associated with that choice). If, however, the initial level were above the utility-maximising level ($z_0 > \tilde{z}$), then $\frac{\partial C}{z_0} < 0$ and it is no longer clear in which direction the chosen level will move as the effect through the hypothetical Moral Benefit works towards increasing the chosen level, but the Utility Cost effect works towards reducing the chosen level.

Result 3 *With a choice function non-linear in the hypothetical Moral Benefit, an increase in in the initial consumption level (and consumption level of all others) can increase the chosen consumption level.*

Let us now return to the main assumption that the initial consumption level is at the utility-maximising level, $\tilde{z}(t)$. If we hold the cost of production c constant, we know that we can't get a change in $\tilde{z}(t)$ without a change in t. Therefore the next step is to evaluate how a change in the tax would influence the chosen consumption level with a generalised choice function.

3.4.4 Impact of a Change in the Emissions Tax

In the linear case we have seen that an increase in the tax leads to a decrease in the chosen consumption level, but this was only driven by the direct effect of the tax on the price of the good which in turn impacted the Utility Cost depending on the weight given to it. If the tax increases, the utility-maximising consumption level $\tilde{z}(t)$ decreases and therefore allows for a reduction in the chosen level without an increase in the associated Utility Cost. However, in a more general setting we have established that the initial consumption level - and by definition the consumption level of all other individuals - can also have an indirect impact on the choice through the hypothetical Moral Benefit and therefore there is another way the tax influences z. The relationship between the tax t and the chosen consumption level of the dirty good z is given by

$$\begin{aligned}\frac{\partial z}{\partial t} =& \Big[-\frac{1}{SOC}\Big]\Big\{\Big[F^{BB}\frac{\partial B}{\partial z}+F^{CB}\frac{\partial C}{\partial z}\Big]\frac{\partial B}{\partial t} \\ &+\Big[F^{CC}\frac{\partial C}{\partial z}+F^{BC}\frac{\partial B}{\partial z}\Big]\frac{\partial C}{\partial t}+F^{C}\Big\} \leq 0 \qquad \forall \quad \hat{z} < z < \tilde{z}(t).\end{aligned} \tag{3.38}$$

This expression is in part very similar to the one in (3.37). There are two components driving the consumption choice. First is the impact of the tax on the Moral Benefit through a change in the consumption of everybody else. Since $\frac{\partial B}{\partial t} < 0$ we know that $[F^{BB}\frac{\partial B}{\partial z}+F^{CB}\frac{\partial C}{\partial z}]\frac{\partial B}{\partial t} \leq 0$. Indeed, this effect is essentially the same as the effect described in (3.37). Because the tax only enters the calculation of the hypothetical Moral Benefit through its determination of the initial consumption level, we can replace $\frac{\partial B}{\partial t}$ with $\frac{\partial B}{\partial \tilde{z}(t)}\frac{\partial \tilde{z}(t)}{\partial t}$ and therefore the effect through the Moral Benefit channel is simply determined by how much a change in the tax changes the initial tax level.

Second, we have the impact of the tax on the Utility Cost of the chosen consumption level. When looking at a change in the initial consumption level only, we found that $\frac{\partial C}{\partial \tilde{z}(t)} = 0$ because $\phi'(\tilde{z}(t)) = c + t$. This is still the case here but the tax also enters the Utility Cost directly on both the calculation of utility under the initial level as well as the chosen level. Therefore we have $\frac{\partial C}{\partial t} = z - \tilde{z}(t) < 0$. And since we know from the conditions of concavity of the choice function that $F^C \leq 0$, we have $[F^{CC}\frac{\partial C}{\partial z} + F^{BC}\frac{\partial B}{\partial z}]\frac{\partial C}{\partial t} + F^C \leq 0$. Thus we can confirm that, as we would intuitively expect, we have $\frac{\partial z}{\partial t} \leq 0$.

Result 4 *With any concave choice function of the hypothetical Moral Benefit and associated Utility Cost, an increase in the tax reduces consumption of the dirty good if* $\hat{z} < z < \tilde{z}(t)$.

3.4.5 Impact of a Change in the Damage Experienced

Next let us evaluate how z is influenced by the total damage from consumption of the dirty good that individuals experience, $D(M\tilde{z}(t))$. Since the individual assumes that everybody else will continue to consume $\tilde{z}(t)$, and the individual's consumption change is negligible due to the atomistic nature of the consumption decision, this is the damage that individuals expect to experience from the emissions externality. The impact is

$$\frac{\partial z}{\partial D(M\tilde{z}(t))} = \Big[-\frac{1}{SOC}\Big]\bigg\{\Big[F^{BB}\frac{\partial B}{\partial z} + F^{CB}\frac{\partial C}{\partial z}\Big]\frac{\partial B}{\partial D(M\tilde{z}(t))}\bigg\} \geq 0 \qquad \forall \quad \hat{z} < z < \tilde{z}(t). \tag{3.39}$$

This is very similar to Equation (3.36) since the experienced damage only has an impact on the hypothetical Moral Benefit. It is easy to determine that $\frac{\partial B}{\partial D(M\tilde{z}(t))} > 0$ for any $\hat{z} < z < \tilde{z}(t)$ and therefore, as we would expect from the above intuition, z is increasing in the damage experienced $D(M\tilde{z}(t))$. As the damage experienced increases, an individual with some propensity to act morally can consume a higher level of the dirty good to achieve the same level of hypothetical Moral Benefit. Note that a change in the total damage caused by the dirty good does not affect the Utility Cost associated with the consumption choice and therefore the consumption choice is only increasing in total damage when the choice function is non-linear in the hypothetical Moral Benefit.

Result 5 *With a choice function non-linear in the hypothetical Moral Benefit, an increase in the level of damage the individual experiences increases the chosen consumption level if $\hat{z} < z < \tilde{z}(t)$.*

3.4.6 Impact of a Change in Altruism

So far we have established how the tax, the initial consumption level as well as the damage experienced may impact the chosen consumption level z. For completeness let us now look at the level of altruistic concern for others' utility, α. The impact of a change in α on the chosen consumption level z is given by

$$\frac{\partial z}{\partial \alpha} = \Big[-\frac{1}{SOC}\Big]\bigg\{F^{BB}\frac{\partial B}{\partial z}\frac{\partial B}{\partial \alpha} + F^{B}\frac{\partial^2 B}{\partial z \partial \alpha}\bigg\}. \tag{3.40}$$

From here we can establish that $\frac{\partial z}{\partial \alpha} < 0$ only when $F^B > -F^{BB}B$ for any $\hat{z} < z < \tilde{z}(t)$. This may seem counter-intuitive since we would expect any altruism to have an effect to reduce consumption if it has an effect at all. However, there are two factors working in opposite direction given a non-linear choice function. The first-order effect captured by $F^B\frac{\partial^2 B}{\partial z \partial \alpha} \leq 0$ works towards reducing consumption of the dirty good since the marginal Moral Benefit of z is decreasing in α. However, the second-order effect $F^{BB}\frac{\partial B}{\partial z}\frac{\partial B}{\partial z} \geq 0$ works towards increasing the consumption since the higher the level of α the higher the level of the Hypothetical Moral Benefit for a given difference between z and $\tilde{z}(t)$. This effect does not exist in the linear setting as the initial level does not impact the choice of consumption in that case.

Result 6 *With a choice function non-linear in the hypothetical Moral Benefit, an increase in the degree of altruistic concern for others only reduces consumption of the dirty good when $F^B > -F^{BB}B$.*[40]

This section has so far provided some insights about how the chosen consumption level may be more sensitive to the initial conditions within a non-linear choice function. To illustrate these insights further, we will now go through two examples of a non-linear choice function, one that is linear in the hypothetical Moral Benefit but non-linear in the Utility Cost, and one that is linear in the Utility Cost but non-linear in the hypothetical Moral Benefit.[41] The purpose of these examples is to help highlight some of the effects described in the more general analysis of this section.

3.4.7 Example 1: $\mu B - (1-\mu)C^2$

In this first example we will analyse a simple version of the choice function that is still linear in the hypothetical Moral Benefit B, but quadratic in the associated Utility Cost C. This specific example will help to isolate some of the individual components influencing the consumption choice that have been established in Sections 3.4.3 - 3.4.6. The choice function in this example is specified such that an individual with propensity to act morally $0 \leq \mu \leq 1$ will choose consumption of the dirty good to maximise

$$\mu B - (1-\mu)C^2. \tag{3.41}$$

It is straightforward to verify that this choice function is strictly concave in z as required.

[40] This condition can be rearranged to $\frac{B}{F^B}F^{BB} > -1$ and could be interpreted as saying that the elasticity of the marginal choice function value with respect to the hypothetical Moral Benefit has to be less than one in absolute terms.

[41] The two example funcitonal forms used in Section 3.4.7 and Section 3.4.8 are not suggestions that these are the 'correct' ones or one is more applicable that the other. They were chosen as example functions that are simple to analyse while fulfilling the concavity condition and at the same time provide one example where the choice function is linear in the hypothetical Moral Benefit and non-linear in the Utility Cost, and one example where the choice function is non-linear in the hypothetical Moral Benefit and linear in the Utility Cost.

Using the results from Section 3.4.1 we find that consumption of the dirty good is defined by

$$\phi'(z) = c + k_{C^2} M D'(Mz) + (1 - k_{C^2})t, \tag{3.42}$$

where

$$k_{C^2} = \frac{\mu(1+\alpha M)}{\mu(1+\alpha M) + 2(1-\mu)C}. \tag{3.43}$$

As in the linear case, when $\mu = 1$ we have $k_{C^2} = 1$ and when $\mu = 0$ we have $k_{C^2} = 0$. Furthermore, as already determined in more generality, the result that the standard Pigovian tax induces everyone to consume the social optimum still holds in this case as well. Looking at the comparative statics, it is straightforward to establish that

$$\frac{\partial z}{\partial D(M\tilde{z}(t))} = 0, \tag{3.44}$$

since the choice function is linear in the hypothetical Moral Benefit and the level of damage experienced does not affect the Utility Cost. This means that, as in the baseline model, the level of damage experienced $D(M\tilde{z}(t))$ has no impact on the chosen consumption level. Next we can determine that the impact of a change in the initial consumption level $\tilde{z}(t)$ is given by

$$\frac{\partial z}{\partial \tilde{z}(t)} = \Big[-\frac{1}{SOC}\Big]\Big\{2(1-\mu)\Big[\phi'(\tilde{z}(t)) - (c+t)\Big]\Big[\phi'(z) - (c+t)\Big]\Big\} = 0, \tag{3.45}$$

where

$$\begin{aligned} SOC =& \mu(1+\alpha M)\Big[\phi''(z) - M^2 D''(Mz)\Big] + 2C(1-\mu)\phi''(z) \\ &- 2(1-\mu)\Big[\phi'(z) - (c+t)\Big]^2 < 0. \end{aligned}$$

As already discussed in Section 3.4.3, this result is driven by the fact that $\phi'(\tilde{z}(t)) = (c+t)$, which is a result of the individual assuming that all other individuals will make their consumption choice optimally in line with standard theory. Given that the initial consumption level is at the utility-maximising $\tilde{z}(t)$, we know that the only channel through which an exogenous change in $\tilde{z}(t)$ might impact the chosen level is through the hypothetical Moral Benefit, and since the choice function is still linear in $B(.)$, the marginal impact is zero. Next, the relationship between z and the tax t is given by

$$\frac{\partial z}{\partial t} = \Big[-\frac{1}{SOC}\Big]\Big\{2(1-\mu)\Big[\phi'(z)-(c+t)\Big][z-\tilde{z}(t)]-2(1-\mu)C\Big\} < 0 \qquad \forall \quad \hat{z} < z < \tilde{z}(t). \tag{3.46}$$

As we can see from the above, the effect of the tax is still just the direct effect on the Utility Cost. The only difference is that this effect is stronger compared to the standard linear case because of the quadratic increase in the Utility Cost from a change in t. As we already know from the result in (3.45), there is no further effect through the impact of a change in t on the consumption of all other individuals. Finally let us look at the impact of a change in α, the altruistic concern for others' utility, on the chosen consumption level of the dirty good. This is given by

$$\frac{\partial z}{\partial \alpha} = \Big[-\frac{1}{SOC}\Big]\mu M\Big[\phi'(z)-c-MD'(Mz)\Big] < 0 \qquad \forall \quad \hat{z} < z < \tilde{z}(t). \tag{3.47}$$

In the general setting we were not able to determine that $\frac{\partial z}{\partial \alpha} \leq 0$ would generally hold for any $\hat{z} < z < \tilde{z}(t)$. However, in this example of the choice function it is straightforward to establish that it does hold and so the chosen consumption level is indeed a decreasing function of the level of altruism; as it is in the baseline model and as we would intuitively expect. There is no upward effect of altruism in this case because the choice function is still linear in the hypothetical Moral Benefit. Indeed the numerator of (3.47) is the same as with the linear choice function. The only difference is in $\frac{1}{SOC}$. The rate at which the choice function changes as z increases is stronger with this choice function compared to the linear function due to the quadratic influence on the choice function and so the effect of an increase in α on the chosen consumption level is actually stronger than it is in the baseline model.[42]

3.4.8 Example 2: $\mu B^{\gamma} - (1-\mu)C$, $0 < \gamma < 1$

As a second example let us now look at a choice function linear in the Utility Cost, but non-linear in the hypothetical Moral Benefit. To maintain simplicity let us assume that an individual with propensity to act morally $0 \leq \mu \leq 1$ chooses consumption of the dirty

[42] This can be determined by comparing the second derivative with respect to z of the linear choice function and the non-linear example used here.

good to maximise

$$\mu B^{\gamma} - (1-\mu)C, \qquad \text{where} \quad 0 < \gamma < 1. \tag{3.48}$$

Again, it is straightforward to verify that the choice function is indeed strictly concave in z as required, which is ensured by $0 < \gamma < 1$. And again we can simply derive the equation that defines the consumption choice of the dirty good as

$$\phi'(z) = c + k_{\gamma} M D'(Mz) + (1-k_{\gamma})t, \tag{3.49}$$

where

$$k_{\gamma} = \frac{\mu\gamma B^{\gamma-1}(1+\alpha M)}{\mu\gamma B^{\gamma-1}(1+\alpha M) + (1-\mu)}. \tag{3.50}$$

As already established for the general case, we again see that when $\mu = 1$ we have $k_{\gamma} = 1$ and when $\mu = 0$, we have $k_{\gamma} = 0$. Furthermore, of course the standard Pigovian tax as defined in (3.10) still induces every individual to consume the socially optimal amount of the dirty good.

With regards to comparative statics, as a first step we can determine that the non-linearity in the hypothetical Moral Benefit implies that the consumption choice of the dirty good does now depend on the consumption choice of all other individuals and therefore also on the damage experienced as a result of the emissions caused by consumption of the dirty good. The impact of the damage experienced on the chosen consumption level is given by

$$\begin{aligned}\frac{\partial z}{\partial D(M\tilde{z}(t))} =& \Big[-\frac{1}{SOC}\Big]\Big\{\mu\gamma(\gamma-1)B^{\gamma-2}(1+\alpha M)^2 \\ & \Big[\phi'(z) - c - MD'(Mz)\Big]\Big\} > 0 \qquad \forall \quad \hat{z} < z < \tilde{z}(t),\end{aligned} \tag{3.51}$$

where

$$\begin{aligned}SOC =& \mu\gamma(\gamma-1)B^{\gamma-2}(1+\alpha M)^2\Big[\phi'(z) - c - MD'(Mz)\Big]^2 \\ & + \mu\gamma B^{\gamma-1}(1+\alpha M)\Big[\phi''(z) - M^2 D''(Mz)\Big] + (1-\mu)\phi''(z) \quad < 0.\end{aligned}$$

It is straightforward to verify that z is indeed an increasing function of $D(M\tilde{z}(t))$ for any $\hat{z} < z < \tilde{z}(t)$. An increase in the damage experienced actually enables the individual to consume a higher amount of the dirty good and still achieve the same level of hypothetical Moral Benefit. Similarly, we can confirm that z is an increasing function of the consumption of all other individuals $\tilde{z}(t)$ by looking at

$$\begin{aligned}\frac{\partial z}{\partial \tilde{z}(t)} =& \Big[-\frac{1}{SOC}\Big]\Big\{-\mu\gamma(\gamma-1)B^{\gamma-2}(1+\alpha M)^2\Big[\phi'(z)-c-MD'(Mz)\Big]\\ &\Big[\phi'(\tilde{z}(t))-c-MD'(M\tilde{z}(t))\Big]\Big\} \quad > 0 \qquad \forall \quad \hat{z} < z < \tilde{z}(t).\end{aligned} \tag{3.52}$$

An increase in the consumption of all other individuals increases the benchmark level of the Moral Benefit which in turn means that the individual can consume a higher amount of the dirty good to achieve the same level of Moral Benefit. Unlike the first example, we can also see now that an increase in the tax on the dirty good does not just impact the consumption choice through the direct effect, but also through the shift in the consumption of all other individuals. This is given by

$$\begin{aligned}\frac{\partial z}{\partial t} =& \Big[-\frac{1}{SOC}\Big]\Big\{\mu\gamma(\gamma-1)B^{\gamma-2}(1+\alpha M)^2\Big[\phi'(z)-c-MD'(Mz)\Big]\\ &\Big[\phi'(\tilde{z}(t))-c-MD'(M\tilde{z}(t))\Big]\frac{\partial \tilde{z}(t)}{\partial t}-(1-\mu)\Big\} < 0 \qquad \forall \quad \hat{z} < z < \tilde{z}(t).\end{aligned} \tag{3.53}$$

The above confirms that z is a decreasing function of the tax rate t. The first part in the braces captures the indirect effect from a shift in the consumption of all other individuals while the second part - $(1-\mu)$ - captures the direct linear effect on the Utility Cost. Finally, the impact of altruism on the chosen level is given by

$$\frac{\partial z}{\partial \alpha} = \Big[-\frac{1}{SOC}\Big]\mu\gamma^2 MB^{\gamma-1}\Big[\phi'(z)-c-MD'(Mz)\Big] \leq 0 \qquad \forall \quad \hat{z} < z < \tilde{z}(t). \tag{3.54}$$

Therefore z is also a decreasing function of the level of altruistic concern for others' utility with this example of a choice function. In this case there is both a factor driving to decrease consumption as well as the factor driving to increase consumption as described in Section 3.4.6, but the downward effect outweighs the upward factor with this particular setup.

This section has provided some insights into how the factors driving the consumption decision under the alternative theory of moral behaviour may be affected by using a non-linear choice function and confirmed that Propositions 2, 4 and 5 also hold with a more general choice function while Proposition 3 only holds under specific conditions. The next section will now extend the analysis further to analyse the case when the individual has to choose consumption of a range of different dirty goods rather than just one.

3.5 Multiple Goods

The baseline model assumed that there is one dirty good causing emissions and damage, as well as the clean numeraire good covering expenditure on all other goods. This section aims to extend this framework to the case of multiple dirty goods, each of which may have a different emissions rate and be subject to a different tax. Note though that each of the dirty goods causes the same type of emissions and it is the total of all emissions from all the dirty goods that cause the damage experienced. This approach is in line with capturing the global climate change problem, with countless different goods causing the GHG emissions which lead to climate change. The primary purpose of this section is to describe the conditions for determining the chosen consumption levels with moral behaviour when there are multiple dirty goods and analyse the effect of a change in the relative price (through the tax) of one dirty good on the chosen consumption level of another dirty good.

3.5.1 Analysis of Choice over Multiple Goods

The setup of this model is mostly identical to the baseline model, but we now have a vector $\mathbf{z}$ of $n \geq 1$ different dirty goods, where $\mathbf{z} = [z_1, \ldots, z_n]$. Each good has a cost of production c_i, $1 \leq i \leq n$, which is captured by the vector $\mathbf{c} = [c_1, \ldots, c_n]$, as well as a corresponding emissions rate e_i captured by the vector $\mathbf{e} = [e_1, \ldots, e_n]$. The cost of production and emissions rate may be different for each good, but doesn't necessarily have to be. Of course if two or more goods have exactly the same cost of production and emissions rate they can be regarded as one and the same good for the purposes of this analysis. But two goods may, for example, have the same cost of production but differ in their emissions rate and vice versa. Finally, each of the n dirty goods may be subject to a different tax imposed by the government.[43] The tax for good z_i is given by t_i and all the

[43] In practice one would expect there to be a tax on each unit of emissions rather than a different tax on each dirty good. The model could be changed to that effect without any significant changes to the results described in this section. However, the comparative statics analysis in Section 3.5.2 aims to show

tax rates are captured by the vector $\mathbf{t} = [t_1, \ldots, t_n]$. Given this setup the hypothetical Moral Benefit of deviating from the standard consumption level is

$$B = (1+\alpha M)\Big\{\Big[y + \phi\big(\mathbf{z}\big) - (\mathbf{c}\cdot\mathbf{z}) - D\big(M(\mathbf{e}\cdot\mathbf{z})\big)\Big] \\ - \Big[y + \phi\big(\tilde{\mathbf{z}}(\mathbf{t})\big) - (\mathbf{c}\cdot\tilde{\mathbf{z}}(\mathbf{t})) - D\big(M(\mathbf{e}\cdot\tilde{\mathbf{z}}(\mathbf{t}))\big)\Big]\Big\}, \tag{3.55}$$

and the Utility Cost from deviating is

$$C = \Big[\phi\big(\tilde{\mathbf{z}}(\mathbf{t})\big) - (\mathbf{c}+\mathbf{t})\cdot\tilde{\mathbf{z}}(\mathbf{t})\Big] - \Big[\phi\big(\mathbf{z}\big) - (\mathbf{c}+\mathbf{t})\cdot\mathbf{z}\Big]. \tag{3.56}$$

For any consumption decision driven by the maximisation of a choice function $F(B, C)$, there are effectively two steps an individual takes in making their consumption choice. The first is a matter of efficiency which means that the individual needs to choose consumption such that the hypothetical Moral Benefit is maximised subject to any given level of Utility Cost. This will provide the efficient frontier of consumption. The second is the choice of consumption along that efficient frontier, which is determined by the weight given to the Moral Benefit and the Utility Cost in the choice function.

As a first step we aim to determine the efficient frontier of consumption, that is to find the maximum level of the hypothetical Moral Benefit given a level of Utility Cost, $\bar{C}$. Assuming linearity in both the hypothetical Moral Benefit and the Utility Cost, this maximum level of the hypothetical Moral Benefit is given by

$$B(\bar{C}) = \underset{\mathbf{z}}{MAX} \quad (1+\alpha M)\Big[\phi\big(\mathbf{z}\big) - (\mathbf{c}\cdot\mathbf{z}) - D\big(M(\mathbf{e}\cdot\mathbf{z})\big)\Big] - l \quad \text{s.t.} \quad C \leq \bar{C}, \tag{3.57}$$

where

$$l = (1+\alpha M)\Big[\phi\big(\tilde{\mathbf{z}}(\mathbf{t})\big) - (\mathbf{c}\cdot\tilde{\mathbf{z}}(\mathbf{t})) - D\big(M(\mathbf{e}\cdot\tilde{\mathbf{z}}(\mathbf{t}))\big)\Big]$$

captures the component of the hypothetical Moral Benefit level that represents everything

how a change in the relative price of one dirty good affects the chosen consumption levels of another dirty good. For this it is useful to have a policy instrument (the tax) that changes the relative prices of the two goods without having to assume a change in the cost of production or emissions rate. Note further that it is not the primary purpose of this section to determine the optimal tax on each good or a unit of emissions but rather to describe the conditions for determining the chosen consumption levels with moral behaviour when there are multiple dirty goods.

to do with the initial level of consumption.[44] The optimisation problem shown in (3.57) is a straightforward multivariate maximisation with an inequality constraint. We can transform this problem into the following Lagrangian:

$$L = (1+\alpha M)\Big[\phi\big(\mathbf{z}\big) - (\mathbf{c}\cdot\mathbf{z}) - D\big(M(\mathbf{e}\cdot\mathbf{z})\big)\Big] - l + \lambda\Big[\bar{C} - C\Big]. \tag{3.58}$$

The Kuhn-Tucker conditions then define the chosen level of consumption for each of the dirty goods z_i, $1 \leq i \leq n$. Denoting $B_i = \frac{\partial B}{\partial z_i}$ and $C_i = \frac{\partial C}{\partial z_i}$, the conditions are

$$\frac{B_i}{C_i} \leq \lambda, \qquad z_i \geq 0, \qquad z_i\frac{B_i}{C_i} = \lambda, \qquad \forall \quad 1 \leq i \leq n; \qquad \text{and} \tag{3.59}$$

$$\bar{C} - C \geq 0, \qquad \lambda \geq 0, \qquad \lambda(\bar{C} - C) = 0, \tag{3.60}$$

where both (3.59) and (3.60) are of course complementary slackness conditions. Also note that $\frac{\partial L}{\partial C} = -\lambda$. The value of λ represents the marginal effect of the Utility Cost constraint on the optimal Moral Benefit, which can also be thought of as the shadow price of the Utility Cost. Of course, if $\lambda = 0$ then no weight would be given to the Utility Cost and the individual would simply maximise the hypothetical Moral Benefit (i.e. $B_i = 0$ $\forall\ 1 \leq i \leq n$). For analysis purposes let us therefore assume that $\lambda > 0$ which implies $\bar{C} = C$. From the conditions in (3.59) and (3.60) we can then see that for any positive consumption levels of any of the dirty goods (i.e. for any $z_i > 0$) we require that

$$\frac{B_i}{C_i} = \frac{(1+\alpha M)\Big[\phi_i'\big(\mathbf{z}\big) - c_i - Me_iD'\big(M\mathbf{e}\cdot\mathbf{z}\big)\Big]}{-\Big[\phi_i'\big(\mathbf{z}\big) - (c_i + t_i)\Big]} = \lambda, \tag{3.61}$$

where $\phi_i'\big(\mathbf{z}\big) = \frac{\partial\phi(\mathbf{z})}{\partial z_i}$. If we assume that there is positive consumption levels of all the dirty goods for any particular individual then we know that consumption of the dirty goods is determined by

$$\frac{B_1}{C_1} = \frac{B_2}{C_2} = \ldots = \frac{B_n}{C_n} = \lambda \tag{3.62}$$

The condition in (3.62) is analogous to the optimality condition in any standard utility

[44] The initial consumption level is irrelevant in the maximisaiton problem because - as in the baseline model - we again have assumed a linear setup for simplicity.

maximisation problem where the marginal rates of substitution of the various goods have to be equal. Indeed we can think of $\frac{B_i}{C_i}$ as a marginal rate of 'moral' substitution, the marginal Moral Benefit relative to its marginal Utility Cost. As such it is intuitive that we require that the marginal rate of moral substitution is equal for each of the dirty goods, which in turn have to be equal to λ. Given we don't yet know the value of λ we can plot the efficient frontier of consumption as follows:

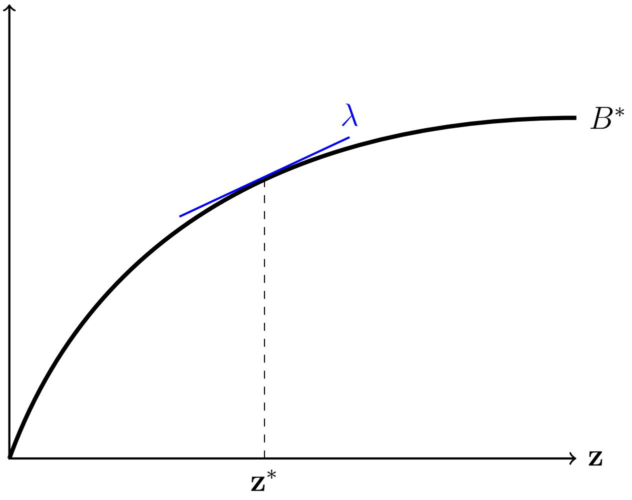

Figure 3.4: Efficient Frontier of Consumption

This curve represents the maximum level of the hypothetical Moral Benefit given any level of the associated Utility Cost. At the same time, each point along this curve represents the optimal consumption level for each different level of λ. The level of λ, and therefore the actual choice of consumption, then of course depends on the specification of the choice function. For any choice function $F(B, C)$ as specified in Section 3.4, we know that analogous to (3.31) the consumption choice is determined by the first-order condition

$$\frac{\partial F}{\partial B} B_i + \frac{\partial F}{\partial C} C_i = 0 \qquad \forall \quad 1 \leq i \leq n. \tag{3.63}$$

For the linear setup used in this section it is then straightforward to determine that for $F(B, C) = \mu B - (1 - \mu)C$, the value of λ is given by

$$\lambda = \frac{1 - \mu}{\mu}. \tag{3.64}$$

Therefore we require the marginal rate of 'moral' substitution to be equal to $\frac{1-\mu}{\mu}$ for each good which has a positive consumption level. If we assume that $n = 1$ then of course this reduces the baseline model where $\mu B_1 = (1 - \mu)C_1$.

So far we have looked at the first-order conditions that define the chosen levels of consumption for each of the dirty goods. In addition, let us assume that all the necessary second-order conditions hold to ensure the stationary points are maxima. If we have only two dirty goods, both of which have a positive consumption level, then it is straightforward to derive that the second-order condition for maximum is given by

$$C_1C_2L_{12} + C_1C_2L_{21} - C_2^2L_{11} - C_1^2L_{22} > 0. \tag{3.65}$$

It is straightforward to determine that this condition is fulfilled if the following two inequalities hold:

$$C_1L_{22} - C_2L_{12} > 0, \tag{3.66}$$

and

$$C_2L_{11} - C_1L_{21} > 0. \tag{3.67}$$

We will assume that the conditions above hold. These will help us in evaluating the results of the comparative statics analysis that follows.

3.5.2 Comparative Statics

Given that we have established the required conditions for determining the consumption choice when there are multiple dirty goods, let us now investigate how consumption of each good is affected by the tax imposed on any of the dirty goods. To simplify the analysis, let us look at the case of two goods. Then it is straightforward to derive from the workings in Section 3.5.1 that the choice problem is characterised by the following three equations:

$$\bar{C} - C = 0, \tag{3.68}$$

$$L_1 = B_1 - \lambda C_1 = 0, \tag{3.69}$$

$$L_2 = B_2 - \lambda C_2 = 0. \tag{3.70}$$

For illustration we look at the effect of a change in the tax imposed on good 1. From the above system of equations we can derive

$$\begin{bmatrix} 0 & -C_1 & -C_2 \\ -C_1 & L_{11} & L_{12} \\ -C_2 & L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} \frac{\partial \lambda}{\partial t_1} \\ \frac{\partial z_1}{\partial t_1} \\ \frac{\partial z_2}{\partial t_1} \end{bmatrix} = \begin{bmatrix} (z_1 - \tilde{z}_1) \\ \lambda \\ 0 \end{bmatrix}. \tag{3.71}$$

Furthermore, let us define

$$\mathbf{J} = \begin{bmatrix} 0 & -C_1 & -C_2 \\ -C_1 & L_{11} & L_{12} \\ -C_2 & L_{21} & L_{22} \end{bmatrix}.$$

From the second-order conditions established in (3.65) we know that $|\mathbf{J}| > 0$. To first investigate how a change in the tax on good 1 affects the consumption of good 1, using Cramer's rule we can determine from (3.71) that

$$\begin{aligned} \frac{\partial z_1}{\partial t_1} &= \frac{1}{|\mathbf{J}|} \begin{vmatrix} 0 & (z_1 - \tilde{z}_1) & -C_2 \\ -C_1 & \lambda & L_{12} \\ -C_2 & 0 & L_{22} \end{vmatrix} \\ &= \frac{-(z_1 - \tilde{z}_1)}{|\mathbf{J}|} \begin{vmatrix} -C_1 & L_{12} \\ -C_2 & L_{22} \end{vmatrix} + \frac{\lambda}{|\mathbf{J}|} \begin{vmatrix} 0 & -C_2 \\ -C_2 & L_{22} \end{vmatrix} < 0. \end{aligned} \tag{3.72}$$

As we would intuitively expect, and in line with the results of the single good case in Section 3.4.4, consumption of good 1 will decrease with an increase in the tax on good 1. The above also shows us that, analogous to any standard consumption optimisation, we were able to split the effect of an increase in the tax rate (which is an increase in the effective price of good 1) into the equivalent of an income and substitution effect. To clarify this, note that (3.72) is the same as

$$\frac{\partial z_1}{\partial t_1} = -(z_1 - \tilde{z}_1)\frac{\partial z_1}{\partial \bar{C}} + \frac{\lambda}{|\mathbf{J}|} \begin{vmatrix} 0 & -C_2 \\ -C_2 & L_{22} \end{vmatrix} < 0. \tag{3.73}$$

We can think of the first component not as an income effect but as a 'Utility Cost effect' and the second component as the substitution effect. The Utility Cost effect captures the change in consumption as a result of an increase in the Utility Cost level (where we know that $\frac{\partial z_1}{\partial \bar{C}} < 0$) and is weighted by the difference between the chosen consumption level and the optimal choice under standard utility-maximising behaviour. Note also that

$-(z_1 - \tilde{z}_1) = -\frac{\partial C}{\partial t_1}$, which is the marginal Utility Cost of an increase in the tax on good 1. Since we know that when an individual maximises the hypothetical Moral Benefit subject to the Utility Cost constraint, we must have $z_i \leq \tilde{z}_i$ and therefore we know that $-(z_1 - \tilde{z}_1) \geq 0$. This in turn means that the Utility Cost effect of an increase in the tax on good 1 is negative as we would intuitively expect. On the other hand, the substitution effect captures the change in consumption purely due to the change in relative prices between the two goods, holding the level of Utility Cost constant, and of course this substitution effect is negative in line with standard theory.

However, the more interesting question is to ask what happens to consumption of good 2 as a results of an increase in the tax on good 1. This is characterised by

$$\begin{aligned}
\frac{\partial z_2}{\partial t_1} &= \frac{1}{|\mathbf{J}|}\begin{vmatrix} 0 & -C_1 & (z_1 - \tilde{z}_1) \\ -C_1 & L_{11} & \lambda \\ -C_2 & L_{21} & 0 \end{vmatrix} \\
&= \frac{(z_1 - \tilde{z}_1)}{|\mathbf{J}|}\begin{vmatrix} -C_1 & L_{11} \\ -C_2 & L_{21} \end{vmatrix} + \frac{-\lambda}{|\mathbf{J}|}\begin{vmatrix} 0 & -C_1 \\ -C_2 & L_{11} \end{vmatrix} \qquad (3.74) \\
&= -(z_1 - \tilde{z}_1)\frac{\partial z_2}{\partial \bar{C}} \qquad\qquad - \frac{\lambda}{|\mathbf{J}|}\begin{vmatrix} 0 & -C_1 \\ -C_2 & L_{11} \end{vmatrix}.
\end{aligned}$$

Again we are able to split the derivative into a substitution effect and a Utility Cost effect. The second part of the above (i.e. the substitution effect) is positive as we would expect. However, just as for good 1, the Utility cost effect is negative. Since the total Utility Cost caused by deviating from the standard consumption level is driven by the consumption of both goods, a change in the tax rate for one good affects the Utility Cost for both goods. This means that it is not clear whether consumption of good 2 will increase or decrease as a result of an increase in the tax on good 1. For the consumption of good 2 to increase, i.e. $\frac{\partial z_2}{\partial t_1} > 0$, the substitution effect needs to outweigh the Utility Cost effect. For this to be the case we require that

$$\frac{C_1 C_2}{C_2 L_{11} - C_1 L_{21}}\lambda > -(z_1 - \tilde{z}_1). \qquad (3.75)$$

Intuitively we know that the closer z_1 and $\tilde{z}_1$ are to begin with the lower the marginal Utility Cost (i.e. C_1) is at z_1. At the same time, the left hand side of (3.75) is decreasing in C_1 and increasing in C_2. Of course a higher marginal Utility Cost for good 1 also goes hand in hand with a larger value for $-(z_1 - \tilde{z}_1)$. But we can say that the substitution effect can outweigh the Utility Cost effect if the marginal Utility Cost of good 2 is sufficiently small (less negative) and at the same time the weight given to the Utility Cost (i.e. the value of λ) is sufficiently large. A larger weight given to the Utility cost is equivalent to a lower propensity to act morally. Therefore, the lower the propensity to act morally the higher the value of λ and thus the more likely it is that an increase in the tax on good 1 increases consumption of good 2. This makes intuitive sense since the lower the propensity to act morally, the more weight is given to utility considerations, and therefore the more flexible the individual is to substitute consumption. In addition, we know that the higher the consumption level of good 2 (i.e. the closer it is to the utility-maximising level), the smaller (less negative) the marginal Utility Cost of good 2. Therefore, the higher the consumption level of good 2 is to begin with the more likely it is that the substitution effect outweighs the Utility Cost effect. This is consistent with the effect of λ since the lower the propensity to act morally, the closer the consumption level of each good will be to the utility-maximising level.

Result 7 *An increase in the tax of a dirty good will decrease consumption of that dirty good but can increase the consumption of another dirty good if the marginal Utility Cost of the second good is sufficiently small and λ is sufficiently large.*

This section has demonstrated how the consumption choice is made when there are multiple dirty goods the individual has to choose over. In addition, this section has demonstrated the interdependencies between the goods, specifically how the tax imposed on one good can either increase or decrease the chosen consumption level of another dirty good.

3.6 Heterogeneous Preferences

A central simplification element of the baseline model was the assumption that private preferences for the dirty good are identical for all individuals in the population. This also meant that in the social optimum all individuals would consume the same amount of the dirty good and the only factor distinguishing individuals in the theory of moral behaviour was their propensity to act morally captured by the parameter μ. This section will attempt to relax this assumption and evaluate what would happen if individuals have heterogeneous preferences. Specifically the aim is to analyse if the result under Proposition 5, which states that the optimal tax is still the standard Pigovian tax as defined in (3.10) under moral behaviour, still holds with heterogeneous preferences. This section assumes that the government is only able to set one tax for all individuals and therefore is not able to set a different tax for individuals of different preference types. For this analysis we will return to the case of just one dirty good and one clean good and furthermore to the linear choice function as used in the baseline model. Furthermore, for simplicity reasons, this section will also assume that there are only two different preference types and there is no altruistic concern for others' utility (i.e. $\alpha = 0$ for the entire population). As before, we will first analyse the model under standard utility-maximising behaviour and then compare this to the alternative theory of moral behaviour.

3.6.1 Standard Theory

There are two types of consumers in the population; those with a 'high' preference for the dirty good and those with a 'low' preference. The private gross benefit from consumption of the dirty good for the high and low type is given by $\phi_H(z)$ and $\phi_L(z)$ respectively, where $\phi_H(z) > \phi_L(z)$ and $\phi'_H(z) > \phi'_L(z)$ for any given z. Then the utility for a consumer of type i, where $i = H, L$, is

$$u_i(z; \bar{z}, t) = \phi_i(z) - (c+t)z + (y + t\bar{z}) - D(M_T\bar{z}), \qquad i = H, L. \tag{3.76}$$

Note that $\bar{z}$ denotes the average consumption of the dirty good across the entire population and, as in the baseline model, due to the atomistic nature of the consumption decision the individual takes this as given with regard to the tax transfer from the government as well as the damage experienced from the externality. The parameter M_T captures the total mass of the entire population and therefore total emissions in the atmosphere are captured by $E = M_T\bar{z}$. Given this setup it is straightforward to determine that, similar to (3.5), we know that the utility-maximising level of consumption of the dirty good for an individual of type i is given by $\tilde{z}_i(t)$ and can be derived from

$$\tilde{z}_i(t) = \underset{z}{ArgMax} \quad u_i(z; \bar{z}, t) \qquad \forall \quad i = H, L. \tag{3.77}$$

From here it is easy to show that utility-maximising consumption of the dirty good for an individual of type i is characterised by

$$\phi'_i[\tilde{z}_i(t)] = c + t \qquad \forall \quad i = H, L. \tag{3.78}$$

As we can see, this is very similar to the case of identical preferences in the baseline model. The individual consumes the amount where the marginal private gross benefit is equal to the marginal private cost of consumption. In doing so, as in the baseline model, the individual ignores both the lump-sum transfer from the government to the individual as well as the damage experienced from the emissions externality. Of course from this we also see that the high type will consume more of the dirty good compared to the low type, i.e. $\tilde{z}_H(t) > \tilde{z}_L(t)$.

When looking at the social optimum under heterogeneous preferences we first have to recognise that we no longer have a case where everybody consumes the same level of the dirty good. As private preferences differ, the socially optimal level of consumption will also differ for the different types of individuals. The social planner maximises total welfare across all individuals. Denoting the consumption level of the dirty good for the high and low types by z_H and z_L respectively, the social utility function is given by

$$\begin{aligned} S(.) =&\sigma_H M_T\Big[\phi_H(z_H) - (c+t)z_H + (y+t\bar{z}) - D(M_T\bar{z})\Big] \\ &+ (1-\sigma_H)M_T\Big[\phi_L(z_L) - (c+t)z_L + (y+t\bar{z}) - D(M_T\bar{z})\Big], \end{aligned} \tag{3.79}$$

where

$$\bar{z} = (\sigma_H M_T z_H + (1-\sigma_H)M_T z_L)/M_T.$$

The parameter σ_H captures the share of the population that is of the high type, while $(1-\sigma_H) = \sigma_L$ represents the share of the population that is of the low type. Therefore $\sigma_H M_T$ captures the mass of the high type population and $(1-\sigma_H)M_T$ captures the mass of the low type population.

The first point that becomes obvious from (3.79) is that because socially optimal consumption is no longer the same across the population, we have $\bar{z} \neq z_H \neq z_L$, and therefore the social planner recognises that the individual will receive a different amount in lump-sum transfer from the government than the individual pays in tax. The social planner also recognises that the total level of emissions is a function of the consumption of both types of individuals and therefore simultaneously maximises the social utility function for the entire population with regard to both the optimal consumption level of each preference type. Given this, the first-order condition for the high type is

$$\begin{aligned} \frac{\partial S(.)}{\partial z_H} =&\sigma_H M_T\Big[\phi_H'(z_H) - (c+t) + \sigma_H t - \sigma_H M_T D'(M_T\bar{z})\Big] \\ &+ (1-\sigma_H)M_T\Big[\sigma_H t - \sigma_H M_T D'(M_T\bar{z})\Big] = 0. \end{aligned} \tag{3.80}$$

This can be re-arranged to

$$\begin{aligned}\frac{\partial S(.)}{\partial z_H} =&\sigma_H M_T \Big[\phi'_H(z_H) - c - M_T D'(M_T \bar{z})\Big] - \sigma_H M_T t + \sigma_H^2 M_T t - \sigma_H^2 M_T^2 D'(M_T \bar{z}) \\ &+ \sigma_H M_T t - \sigma_H^2 M_T t + \sigma_H^2 M_T^2 D'(M_T \bar{z}) = 0. \end{aligned} \tag{3.81}$$

From this it is then easy to see that the socially optimal level of consumption for the high type is characterised by

$$\phi'_H(\hat{z}_H) = c + M_T D'(M_T \hat{z}_T), \tag{3.82}$$

where

$$\hat{z}_T = (\sigma_H M_T \hat{z}_H + (1 - \sigma_H) M_T \hat{z}_L)/M_T. \tag{3.83}$$

Similarly, the socially optimal level of consumption for the low type is characterised by

$$\phi'_L(\hat{z}_L) = c + M_T D'(M_T \hat{z}_T). \tag{3.84}$$

As one would intuitively expect, the socially optimal levels for each type are defined by the point where the private marginal gross benefit of consumption is equal to the social marginal cost of consumption. Individuals of different types are not consuming the same amount in the social optimum and we have $\hat{z}_H > \hat{z}_L$. However, the difference in consumption levels is purely driven by the difference in preferences. It is also noteworthy that the socially optimal level of consumption for a particular type is a function of the socially optimal consumption level of the other type and therefore it is a function of the preference level of the other type. The optimal level of the other type influences the total emissions level and therefore marginal damage. To illustrate this let us determine specifically how optimal consumption of the high type is affected by a change in the optimum for the low type. This is given by

$$\frac{\partial \hat{z}_H}{\partial \hat{z}_L} = \frac{(1 - \sigma_H) M_T^2 D''(M_T \hat{z}_T)}{\phi''_H(\hat{z}_H) - \sigma_H M_T^2 D''(M_T \hat{z}_T)} < 0. \tag{3.85}$$

From (3.85) we see that if the optimum for the low type increases - for example as the result of an upward shift in preferences - the social optimum for the high type is reduced.[45] This is because the increase in demand from the low type increases their optimal consumption level which in turn increases total emissions and therefore the marginal damage for both types. The increase in marginal damage of course dampens the optimal increase in consumption for the low type compared to the case if there were no externality associated with the good. However, it increases the marginal damage for the high type as well, and this will reduce their optimal consumption level. Note that the degree to which consumption changes depends on the share of the population that is of the low type, as this determines how much marginal damage increases as a result of an increase in the optimum for the low type (i.e. the numerator of the expression in (3.85)). The change in optimal consumption for the high type is therefore determined by the increase in the marginal damage relative to the difference between the marginal change in private gross benefit derived from the dirty good and the change in marginal damage as a result of a change in consumption of the high type. It is also important to note that while individual utility-maximising behaviour for each type is completely independent of any other preference type, the social optimum for each type is linked to the preferences of all other types. While the atomistic nature of the consumption decision means that individuals do not take into account the emissions externality in their utility-maximising choice, the social planner does of course takes this into account and it is consumption of all types that determines total emissions and therefore the damage experienced from consumption of the dirty good.

We have determined that the socially optimal level of consumption of the dirty good for each type will be different. However, it also becomes evident from (3.82) and (3.84) as well as (3.78), that for each type the optimal tax that induces the socially optimal level of consumption is still given by

$$\hat{t}_H = \hat{t}_L = \hat{t} = M_T D'(M_T \hat{z}_T). \tag{3.86}$$

Therefore we still have a single tax rate at the standard Pigovian level of social marginal damage as defined in (3.10). This tax induces the socially optimal level of consumption for both types despite the differences in consumption preferences. Intuitively this

[45] Note that this is only the case with a strictly convex damage function. if the damage function were linear in emissions, i.e. $D''(.) = 0$, then the socially optimal consumption level would be unaffected by a preference change of the other type.

makes sense as the purpose of the Pigovian tax is to bring the individual to internalise the damage caused by consumption of the dirty good and the marginal damage across the population is the same as both types are equally affected by the total emissions. Of course this doesn't mean that, unlike the baseline model, individuals consume the same amount of the dirty good, but each individual will consume their socially optimal level depending on their preferences.

3.6.2 Moral Behaviour

The Utility Cost an individual incurs from deviating from the utility-maximising choice is, as before, given by the difference in utility between the chosen level given that everybody else continues to behave in a utility-maximising way and the level obtained under utility-maximising behaviour. Therefore the Utility Cost for an individual of preference type i is

$$\begin{aligned} C_i(z;\zeta(t);t) =& u_i(\tilde{z}_i(t);\zeta(t),t) - u_i(z;\zeta(t),t) \\ =& \Big[\phi_i(\tilde{z}_i(t)) - (c+t)\tilde{z}_i(t)\Big] - \Big[\phi_i(z_i) - (c+t)z_i\Big] \qquad \forall \quad i = H, L. \end{aligned} \tag{3.87}$$

Of course this is essentially the same as the Utility Cost of the baseline model as defined in (3.15). The only difference between different types is in the private gross benefit derived from consumption of the dirty good, $\phi_i(z_i)$, and the corresponding utility-maximising level of consumption $\tilde{z}_i(t)$. As we know from the development of the baseline model, the atomistic nature of the consumption decision determines that individuals ignore the effect of their consumption on others and take the consumption of all others as given. This in turn means that the calculation of the loss in utility as a result of deviating from the utility-maximising choice is also independent of the choices of all others, and therefore independent of other preference types. As in the case of identical preferences, it is also straightforward to determine from (3.87) that an individual of type i who chooses consumption to minimise the loss in utility will simply choose the utility-maximising level defined in (3.78).

In the baseline model with homogeneous preferences the Kantian question the individual asked was what would be optimal if the individual and everybody else were to consume the same amount of the dirty good. This rule worked in the case of homogeneous preferences since the social optimum required all individuals to consume the same amount of the dirty good. However, now that we have introduced heterogeneous preferences we need to refine the Kantian question to reflect the variation in preferences. The following description given by Brekke et al. (2003) gives a good way of doing so: "In a model with heterogeneous consumers, the adequate question would be: 'Which general rule of action would maximise social welfare, as I perceive it, given that everyone acted according to the same general rule as I?' The morally ideal action would then be a function of one's own individual characteristics, for example income and/or preferences." (p. 1972, footnote 8).

Using this we can say that a fully Kantian individual would need to determine the social optimum in some way. The hypothetical Moral Benefit used in the baseline model asked what the utility gain would be for the individual if everybody were to act the same way compared to standard utility-maximising behaviour. Since everyone was homogeneous, maximising the utility this individual would get if everybody else acted the same way, was the same as maximising the social welfare function. However, now the social welfare function is not just the individual's utility multiplied by the population size, it is the aggregate of all the different preference types across the population. And as we have established in the analysis under standard theory, the social optimum is different for each type and is also a function of the social optimum of other types. This is important in the way we define the hypothetical Moral Benefit. If we were to determine the hypothetical Moral Benefit by looking at the gain in overall social welfare if everybody were to act in the socially optimal way compared to the utility-maximising way, then the maximisation of this would indeed lead individuals to choose the morally ideal (or socially optimal) level of consumption. However, in this alternative theory of moral behaviour the individual trades off the hypothetical Moral Benefit against the loss of utility incurred by deviating from the utility-maximising choice. If we defined the hypothetical Moral Benefit based on aggregate social welfare, then the choice function would make an inconsistent trade-off (i.e. gain in aggregate social welfare versus loss in individual utility). Furthermore, if we then tried to correct for this by making the utility-cost calculation also on the basis of the aggregate across the population, then we would have removed the consumption choice entirely from the individual's perspective. The alternative theory of moral behaviour that has been developed in Section 3.3.3 is concerned with the trade-off between an individual's propensity to act morally and their individual loss in utility from acting morally. This is a central element and should be maintained under heterogeneous preferences as well. Therefore we need to find a way in which the individual's hypothetical Moral Benefit is still determined by assessing the utility gain if that individual and everybody else acted in the same way (but not necessarily consumed the same amount of the dirty good).

We know that the individual's moral choice is a choice of consumption at or somewhere between the utility-maximising level and the social optimum. We can therefore represent the chosen level for an individual of type i as

$$z_i = \theta \hat{z}_i + (1-\theta)\tilde{z}_i(t) \qquad \forall \quad i = H, L, \tag{3.88}$$

where the parameter θ captures the weight given towards the social optimum and determines how far the individual moves from the utility-maximising level towards the social optimum. Given this we can stipulate that the hypothetical Moral Benefit is determined by looking at the gain in utility the individual would have if they and everybody else in the population would move towards the social optimum by the same degree. In other words this means that the hypothetical Moral Benefit is determined by asking what would my utility gain be if I and everybody else were to choose the same θ. Therefore the hypothetical Moral Benefit of deviating from the utility-maximising choice for an individual of type i is therefore given by

$$B_i(\theta; t, \tilde{z}_i(t), \hat{z}_i) = u_i(\theta, \zeta(\theta); t, \tilde{z}_i(t), \hat{z}_i) - u_i(\tilde{z}_i(t), \zeta(t); t), \qquad \forall \quad i = H, L, \tag{3.89}$$

where $\zeta(\theta)$ is an assignment function that assigns everybody else in the population the level of consumption consistent with a level between their utility-maximising and socially optimal level (which differs depending on preference type) based on the weight θ as defined in (3.88). Furthermore, $\zeta(t)$ is the assignment function that assigns everybody else in the population their utility maximising level of consumption. As an example, the hypothetical Moral Benefit for an individual of the high type is

$$\begin{aligned} B_H(.) = &\Big[\phi_H(z_H) - [c+t]z_H + t[\sigma_H z_H + (1-\sigma_H)z_L] \\ &- D(\sigma_H M_T z_H + (1-\sigma_H) M_T z_L)\Big] \\ &- \Big[\phi_H(\tilde{z}_H(t)) - [c+t]\tilde{z}_H(t) + t[\sigma_H \tilde{z}_H(t) + (1-\sigma_H)\tilde{z}_L(t)] \\ &- D(\sigma_H M_T \tilde{z}_H(t) + (1-\sigma_H) M_T \tilde{z}_L(t))\Big], \end{aligned} \tag{3.90}$$

where

$$z_H = \theta \hat{z}_H + (1-\theta)\tilde{z}_H(t), \qquad \text{and} \quad z_L = \theta \hat{z}_L + (1-\theta)\tilde{z}_L(t).$$

If the individual's consumption were only driven by the hypothetical Moral Benefit, the individual would maximise the above with respect to θ. This yields the following first-order condition:

$$\phi'_H(z_H) = c + M_T D'(.) + \frac{\big(\tilde{z}_H(t) - \hat{z}_H\big) - \big(\tilde{z}_L(t) - \hat{z}_L\big)}{\tilde{z}_H(t) - \hat{z}_H}(1 - \sigma_H)\Big[t - M_T D'(.)\Big]. \quad (3.91)$$

From (3.91) it is straightforward to determine that this will only lead to a consumption choice at the social optimum (i.e. $\phi'_H(z_H) = c + M_T D'(.)$), if we have $\hat{z}_H - \tilde{z}_H(t) = \hat{z}_L - \tilde{z}_L(t)$ when $t \neq \hat{t}$. However, due to the differences in the preferences between the high and low type the difference between their social optimum and utility-maximising level may well be different.[46] Therefore, contrary to the baseline model with identical individuals, maximisation of the hypothetical Moral Benefit no longer necessarily leads to consumption of the social optimum. This is because the individual is not actually able to determine the 'true' moral optimum (i.e. the utility derived if they and everybody else were to consume the social optimum) but rather makes an estimate of this by assuming that everybody else would move towards their social optimum by the same degree. The effect we observe is driven by the fact that if the individual of type H chooses θ, this moves the assumed consumption choice for individuals of type L different to that of type H. But at the same time the consumption level of type L influences type H through both the lump-sum transfer from the government and the total emissions leading to the experienced damage. We can see from (3.91) that the degree to which the damage is internalised depends on the weighted difference between the utility-maximising and socially optimal levels across the population relative to the difference for this particular preference type. This makes intuitive sense since it reflects the assumption that everybody else moves towards the social optimum by the same factor and so the individual will internalise the movement of the entire population relative to his preference type.

Another critical difference to the baseline model is that maximising the hypothetical Moral Benefit leads the individual to take account of the tax imposed on consumption of the dirty good. In the baseline model the individual's benchmark was that everybody consumes the same amount and therefore the tax paid on the dirty good would equal the lump-sum transfer from the government, rendering the tax irrelevant. However, given that the government only sets one tax across the entire population and we have heterogenous preferences, the tax paid by a particular preference type will no longer match the transfer from the government since the average consumption level across the population is not the same as the consumption level of that preference type. From (3.91) we also see

[46] And if there were more than two types it would almost be certain that the differences would not be the same.

that whether an individual maximising the hypothetical Moral Benefit alone, consumes more or less than the social optimum depends on the tax relative to the Pigovian level and whether the difference between the social optimum and the utility-maximising level is larger for the high or low type.[47] For example, assuming a $t < \hat{t}$ an individual of the high type will consume less than the social optimum if the difference between the social optimum and the utility-maximising level is grater for the high type compared to the low type. On the other hand, if the difference of the low type is sufficiently large relative to the high type, then the individual may consume more than the social optimum.

The key here is that even if the individual only maximises the hypothetical Moral Benefit, the tax plays a critical role in determining consumption. However, we can also determine from (3.91) that if the government imposes the Pigovian tax (i.e. $\hat{t} = M_T D'(.)$), this reduces (3.91) to $\phi'_H(z_H) = c + M_T D'(.)$ and therefore means that maximisation of the hypothetical Moral Benefit with the Pigovian tax in place leads to consumption of the socially optimal level. Looking at (3.91) we can see that if $t = \hat{t}$ the distortion through the damage effect exactly cancels out the distortion through the tax channel. Note that while this is shown here for the high preference type, the equivalent also holds for the low preference type. Intuitively this makes sense because the Pigovian tax eliminates any difference between the utility-maximising level and the social optimum for all preference types as shown in the analysis of standard behaviour in Section 3.6.1.

Given we have set up both the Utility Cost of alternative behaviour as well as the calculation of the hypothetical Moral Benefit, we can now combine them in the individual's choice function determining their consumption choice. Just as in the baseline model we will assume that the individual makes the consumption decision by maximising the basic linear choice function as defined in (3.18). Therefore the individual of type i will choose consumption by maximising

$$\mu B_i - (1-\mu)C_i, \qquad 0 \leq \mu \leq 1, \qquad i = L, N. \tag{3.92}$$

As in the baseline model, the parameter μ captures the individual's propensity to act morally. This propensity to act morally is also distributed across the population but entirely independent of the distribution of preferences. Using preference type H for the further analysis, and substituting (3.87) and (3.90) into (3.92) it is straightforward to

[47] The case where the individual only maximises the hypothetical Moral Benefit is equal to the case of $\mu = 1$ when we use the linear choice function.

determine from the first-order condition that the consumption choice for an individual of type H with propensity to act morally μ is characterised by

$$\begin{aligned}\phi'_H(z_H) = c + \mu M_T D'(.) + (1-\mu)t \\ + \mu\frac{\big(\tilde{z}_H(t) - \hat{z}_H\big) - \big(\tilde{z}_L(t) - \hat{z}_L\big)}{\tilde{z}_H(t) - \hat{z}_H}(1-\sigma_H)\Big[t - M_T D'(.)\Big].\end{aligned} \tag{3.93}$$

Similarly, the consumption choice for an individual of type L with propensity to act morally μ is characterised by

$$\begin{aligned}\phi'_L(z_L) = c + \mu M_T D'(.) + (1-\mu)t \\ + \mu\frac{\big(\tilde{z}_L(t) - \hat{z}_L\big) - \big(\tilde{z}_H(t) - \hat{z}_H\big)}{\tilde{z}_L(t) - \hat{z}_L}\sigma_H\Big[t - M_T D'(.)\Big].\end{aligned} \tag{3.94}$$

We know that an individual with full propensity to act morally (i.e. $\mu = 1$) puts no weight to the Utility Cost associated with deviating and therefore only maximises the hypothetical Moral Benefit. From the earlier analysis of this case we already know that maximisation of the hypothetical Moral Benefit does not on its own lead to consumption of the social optimum and, unlike the baseline model, is still a function of the tax rate. This of course translates directly into the results determined by (3.93) and (3.94). Assuming that $t \neq \hat{t}$, we know that an individual with $\mu = 1$ will consume their best estimation of the morally ideal effort, but not the actual social optimum. On the other hand, an individual with no propensity to act morally (i.e. $\mu = 0$) will minimise the Utility Cost without regard for the hypothetical Moral Benefit and therefore, as in the baseline model, simply consume the utility-maximising level as defined in (3.78). However, we can also determine from (3.93) and (3.94) that if the government imposes the Pigovian tax, the individual will consume the social optimum regardless of their propensity to act morally. Therefore the baseline model result that the standard Pigovian tax is still the optimal tax under the alternative theory of behaviour also holds under heterogenous preferences. The reason for this that (a) the Pigovian tax exactly offsets the distortion caused through the damage factor and the tax factor from the imperfect assessment of the morally ideal consumption level as described earlier and (b) the Pigovian tax makes everybody consume the social optimum regardless of their propensity to act morally as demonstrated in the baseline model results. Therefore, even when individuals make the calculation of the morally ideal benchmark level in an imperfect way in the presence of heterogenous preferences, the optimal tax induces everyone to consume the socially optimal level of the dirty good regardless of differences in preferences and differences in individuals' propensity to act morally.[48]

Proposition 8 *If there are two types of individuals with different preferences over consumption of the dirty good, where at the same time individuals may exhibit some propensity to act morally, the Pigovian tax still induces everyone in the population to consume the socially optimal level of the dirty good.*

Proof: Plugging the Pigovian tax defined in (3.10) into the characterisation of consumption of both the high and low types as defined in (3.93) and (3.94) respectively, these reduce to their socially optimal levels as defined in (3.82) and (3.84).

So far we developed a number extensions around the baseline model and shown that these do not change the optimal tax the government should set on consumption of the dirty good. The next section will develop one further extension, combining the theory of moral behaviour with a model of desire for conformity.

3.7 Moral Behaviour and a Desire for Conformity

This section aims to combine the theory of moral behaviour with the model of desire for conformity developed in Ulph and Ulph (2014) and also introduced as part of Dasgupta et al. (2015). The model developed by Ulph and Ulph (2014) captures the idea that individuals might change their consumption choice away from the standard level in order to get closer to a norm that is established by the consumption choices of other individuals. This approach is different from the modelling of Veblen effects in that individuals are not trying to match their consumption to that of an aspirational group, but rather value the conformity with similar individuals as such, which in turn establishes a consumption norm. This also means that, in contrast to competitive consumption, it may induce the individual to consume less of good in order to conform to the norm. At the same time an individual chooses whether to adhere to norm or not, and will only do so if this yields a net utility benefit. For more details on this see Ulph and Ulph (2014).[49]

When combining this model of desire for conformity with the alternative theory of moral behaviour it is important to determine at what stage morality enters into the individual's consideration. One option is that the moral consideration enters at the consumption decision stage, another is that it enters at the stage where individuals choose to adhere to a

[48] While one can intuitively expect this result to also hold when there are more then two preference types, further analysis is required to formally show this.

[49] For a brief overview of the literature on social norms see Section 3.2.2.

norm or not. Furthermore, when including morality it is important to consider whether the norm itself should have some normative value or whether norms simply emerge as a result of the consumption decisions as done in Ulph and Ulph (2014). In order to be as consistent as possible with the two models this section is based on, the analysis will focus on the case where the norm simply emerges and morality only enters at the consumption decision level. This is consistent with the approach in the baseline model. Note that this means that the Kantian optimum is therefore independent of the norm that emerges since the Kantian optimum already assumes that everybody consumes the same level and therefore everybody consumes at that level which would also be the emerging norm if everybody were to act in a fully Kantian fashion. This also means that an individual's choice whether to adhere to a norm or not is a purely utility-maximising process and individuals will take into account their propensity to act morally and the effect it has on their utility when they choose whether to adhere to a norm or not. Note that the analysis done in this section will not analyse the emergence of norms in detail but will only do so to evaluate whether Proposition 5 - which says that the government should set the standard Pigovian tax - still holds when individuals have a desire for conformity.

3.7.1 Model Setup

The basic model setup is the same as in the baseline model developed in Section 3.3. However, for notational simplicity we will now denote average consumption of the dirty good by z_A instead of $\bar{z}$. In order to extend the model to capture adherence to a norm with a desire for conformity in line with Ulph and Ulph (2014) we simply add two elements to the utility function defined in (3.4). First is the "strength of the desire for conformity" (Ulph and Ulph 2014, p. 4) which is captured by ω. Second, we require the "individual's strength of adherence to the norm" (Ulph and Ulph 2014, p. 4), which is captured by γ. Therefore utility for an individual who is adhering to a norm is

$$u(z; z_A, z_N, \gamma, \omega) = \phi(z) - (c+t)z + (y + tz_A) - D(Mz_A) - \gamma|z - z_N| + \omega, \tag{3.95}$$

where z_N is the norm level of consumption that this individual has chosen to adhere to. While ω is simply a benefit the individual derives from adhering to a norm, the parameter γ determines the reduction in utility if an individual has chosen to adhere to a norm but doesn't consume exactly the norm level.[50] As will become clearer in the analysis,

[50] While it is not straightforward to make predictions about the levels of the strength of desire for conformity ω and the strength of adherence to the norm γ, one may expect that an individual with

the individual has a choice whether to adhere to a norm in the first place or not. If the individual chooses not to adhere to a norm then the utility function reduces to the standard setup and the model reduces to the baseline case analysed in Section 3.3. As usual, we will briefly develop the key results under standard utility-maximising behaviour to serve as counterfactual before moving on to the main analysis of moral behaviour when individuals also value conformity.

3.7.2 Utility-Maximising Behaviour

This section will only repeat the analysis of Ulph and Ulph (2014) at a very simple level in order to build a counterfactual for the analysis of moral behaviour. For detailed results of consumption choices when people value conformity and act in a standard utility-maximising way, see Ulph and Ulph (2014) and Dasgupta et al. (2015). As is explained in those papers, there are three stages to analysing the consumption problem. First is the individual's choice whether to adhere to a norm or not. If the individual chooses not to adhere to a norm, then they will simply maximise their 'standard' utility as shown in (3.4) and the individual's consumption choice will be characterised by $\phi'(z) = c + t$ as usual. The second stage is the determination of the norm and the third stage is the individual's consumption decision. Note that for now we will ignore a further stage where the government decides on the level of the tax on the dirty good, but of course this stage would precede the others. Of course it is most effective to conduct the model analysis backwards through the various stages - as done in Ulph and Ulph (2014) - and therefore we start with the consumption choice given the individual has chosen to adhere to a norm z_N. Note also that for analysis of Stage 2 & 3 we can ignore the fixed benefit ω that individuals get if they choose to adhere to a norm, since it is not a function of the chosen level or the norm level of consumption and only has significance when the individual decides whether to adhere to a norm or not.

Stage 3 - Consumption Choice

The maximisation of the utility function has to be done in two stages to deal with the absolute value of the difference between consumption and the norm level. Therefore we

a high desire for conformity also has a higher strength of adherence to the norm. However, it is also plausible that an individual may care a great deal about being a conformist without being too strict about how close the consumption choice then actually is to the norm level. At the same time, there may be individuals whose desire for conformity is not very strong, but once they have decided to conform to a norm they are very strict about consuming close to the norm level.

have two constrained maximisation problems. In the first one the consumer chooses z to

$$\underset{z}{MAX} \quad y + \phi(z) - (c+t)z + tz_A - D(Mz_A) - \gamma(z - z_N) \qquad \text{s.t.} \quad c \geq c_N, \tag{3.96}$$

and in the second one the consumer chooses z to

$$\underset{z}{MAX} \quad y + \phi(z) - (c+t)z + tz_A - D(Mz_A) - \gamma(z_N - z) \qquad \text{s.t.} \quad c \leq c_N. \tag{3.97}$$

The solutions to the above problems are

$$z = z_N \Leftrightarrow z_N \geq \underline{z}; \qquad z = \underline{z} \Leftrightarrow z_N < \underline{z}, \tag{3.98}$$

and

$$z = z_N \Leftrightarrow z_N \leq \overline{z}; \qquad z = \overline{z} \Leftrightarrow z_N > \overline{z}, \tag{3.99}$$

respectively. As such the solutions in (3.98) and (3.99) define the norm-consistent interval of consumption

$$\Big[\phi'(\underline{z}) = c + t + \gamma \quad , \quad \phi'(\overline{z}) = c + t - \gamma\Big]. \tag{3.100}$$

If the norm lies within that interval the individual will choose the norm level, and if it lies outside the norm-consistent interval, the individual will choose the boundary level closest to the norm (i.e. either the upper or lower bound of the interval).

Stage 2 - Equilibrium Norms

The analysis in this section will use the same definition of an equilibrium norm as defined in Ulph and Ulph (2014), which states that:

A norm, z_N is an equilibrium norm if[51]

1. it is the average of the consumption decisions of al the individuals who adhere to that norm, as determined in Stage 3.

[51] Ulph and Ulph (2014), p. 7

2. there is more than one norm in existence then the norm to which any individual adheres is that which generates the highest level of indirect utility for that individual.

So far we only have identical individuals in the analysis. In that case there will be a single equilibrium norm that can take any value in the norm-consistent interval of consumption. If there were heterogeneity in the individuals' strength of adherence to the norm, γ, but everybody would be identical otherwise, there will be a single norm within the norm-consistent interval of the individual with the lowest strength of adherence to the norm. Heterogeneity in other factors may lead to norms outside anyone's norm-consistent interval and may also lead to multiple norms. We will explore this further when it becomes relevant in the analysis of moral behaviour. For more detail on the norms that may emerge under standard utility-maximising behaviour see Ulph and Ulph (2014).

Stage 1 - Decision over adherence to Norm

An individual will choose to adhere to a norm if the utility derived from adhering to it, as defined in (3.95), is greater than the utility derived if the individual chooses not to adhere to a norm, as defined defined in (3.4). This means an individual will choose to adhere to a norm if

$$\omega > \gamma |z - z_N|. \tag{3.101}$$

Social Planner

When maximising social welfare we know that since individuals have identical utility functions, all individuals would consume the same amount of the dirty good and therefore the social optimum is characterised by

$$\phi'(\hat{z}) = c + MD'(M\hat{z}). \tag{3.102}$$

This level is of course the same as in the baseline model. Furthermore, since all individuals would consume the same level, this social optimum would also be the norm level and everybody would choose to adhere to that norm.

3.7.3 Moral Behaviour

Let us now implement the moral behaviour element and, as in the baseline model, the analysis will assume that individuals may have some propensity to act morally $0 \leq \mu \leq 1$, and choose consumption by maximising the choice function

$$\mu B - (1-\mu)C,$$

where, for an individual who has chosen to adhere to a norm, $B(.)$ is the hypothetical Moral Benefit given by

$$\begin{aligned} B(.) = &\Big[y + \phi(z) - cz - D(Mz)\Big] \\ &- \Big[(y + tz_A) + \phi(z_0) - (c+t)z_0 - D(Mz_A) - \gamma|z_0 - z_N|\Big], \end{aligned} \tag{3.103}$$

and $C(.)$ is the associated Utility Cost given by

$$C(.) = \Big[\phi(z_0) - (c+t)z_0 - \gamma|z_0 - z_N|\Big] - \Big[\phi(z) - (c+t)z - \gamma|z - z_N|\Big]. \tag{3.104}$$

The term z_0 captures the chosen consumption level under utility-maximising behaviour as analysed in Section 3.7.2.

Stage 3 - Consumption Choice

Applying the two-stage analysis analogous to the one in Section 3.7.2, we find that the norm-consistent interval for an individual with $0 \leq \mu \leq 1$ is

$$\begin{aligned} \Big[\phi'(\underline{z}_\mu) &= c + \mu M D'(M\underline{z}_\mu) + (1-\mu)t + (1-\mu)\gamma, \\ \phi'(\overline{z}_\mu) &= c + \mu M D'(M\overline{z}_\mu) + (1-\mu)t - (1-\mu)\gamma\Big]. \end{aligned} \tag{3.105}$$

For individuals with full propensity to act morally, i.e. $\mu = 1$, we find that $\phi'(\underline{z}_\mu) = \phi'(\overline{z}_\mu) = c + MD'(M\hat{z})$, which is the socially optimal level. This means that people with full propensity to act morally will always choose the social optimum regardless of their strength of adherence to the norm. As such they will ignore the social norm in favour of the moral norm. However, if the individual with full propensity to act morally has nevertheless chosen in Stage 1 to seek conformity, their choice will impact the norm that emerges and therefore pull the consumption of individuals with a lower propensity to act morally towards the social optimum. The extent to which this happens will of course depend on where the norm lies, and how strong others' adherence to that norm is.

On the other hand, for $\mu = 0$ we have $\phi'(\underline{z}) = c + t + \gamma$ and $\phi'(\overline{z}) = c + t - \gamma$, which is identical to (3.100) and represents the norm-consistent interval of consumption under utility-maximising behaviour. This is of course a function of the desire for conformity, but not a function of the social optimum. As before, the individual will consume the norm level if the norm lies within the interval and will consume the boundary level if the norm lies outside the interval.

For simplicity let us now assume that the tax is zero, i.e. $t = 0$. We can plot this as follows:

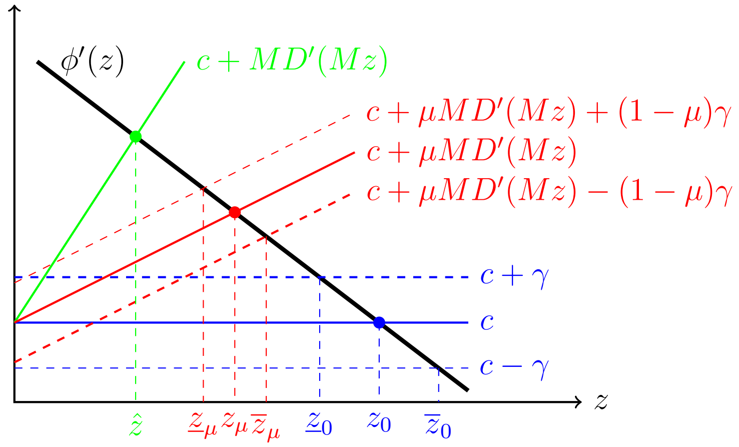

Figure 3.5: Norm-Consistent Intervals of Consumption under Moral Behaviour

Figure 3.5 shows the norm-consistent intervals for an individual with $\mu = 0$ (blue) and for an individual with $0 < \mu < 1$ (red). Note also that for an individual with $\mu = 1$ (green), there is no norm-consistent interval and, as already mentioned, this individual will always consume the socially optimal level. Critically, we can see from (3.105) and Figure 3.5 that the norm-consistent interval of consumption becomes narrower as μ increases. This is because the higher the individual's propensity to act morally, the less weight is given to the individual's utility relative to the Kantian optimum and therefore effectively less weight is given to adherence to the norm.

Result 8 *The higher an individual's propensity to act morally, the narrower the norm-consistent interval of consumption given a certain strength of adherence to the norm.*

An individual with full propensity to act morally has no norm-consistent interval and will consume the social optimum regardless of where the norm lies and the individual's strength of adherence to the norm.

Stage 2 - Equilibrium Norms

Since we don't have identical individuals anymore it is less straightforward to determine what the emerging norm looks like. Let us proceed by first assuming that individuals are identical in all regards except their propensity to act morally. Looking back at Figure 3.5 assume that there are only two types of individuals, one with no propensity to act morally ($\mu = 0$, marked in blue), and individuals with some, but not full propensity to act morally ($0 < \mu < 1$, marked in red). Further assume that, as drawn in Figure 3.5, the norm-consistent intervals of consumption for the two types do not overlap and that there is only one norm.[52] Then we know from Ulph and Ulph (2014) that the norm will lie between the red interval and the blue interval. This in turn means that the red individuals consume the upper bound of their interval and the blue individuals consume the lower bound of their interval. Therefore adhering to the norm actually increases consumption for the individual with some propensity to act morally, and decreases it for the individual with no propensity to act morally relative to the level that would have been chosen if the individuals were not adhering to a norm. However, because the norm-consistent interval for the individuals with some propensity to act morally is narrower than the one for individuals without propensity to act morally, their increase in consumption from the level that would have been chosen if they weren't adhering to a norm, is smaller than the decrease in consumption for the individuals with no propensity to act morally relative to their consumption that would have been chosen if they weren't adhering to a norm. Because an individual with some propensity to act morally is incurring a Utility Cost, the increase in consumption from adhering to the norm actually increases utility derived from consumption (excluding the direct effect from having chosen to conform in Stage 1). This is contrary to the results under utility-maximising behaviour, where deviation from the standard level yields a decrease in direct utility derived from consumption.

Result 9 *If an individual with some, but not full propensity to act morally adheres to a norm that lies at a higher consumption level, adherence will increase consumption and utility as it reduces the Utility Cost associated with the individual's propensity to act morally.*

[52] Of course in this example we could also have two norms emerging, and one would lie in the norm-consistent interval of the blue individuals and the other in the norm-consistent interval of the red individuals.

Example of Specific Functional Form

In order to understand how the effect just described affects average consumption of the dirty good across the population let us analyse a simplified example of the model using a specific functional form for the private gross benefit of consumption of the dirty good and the damage function. These are

$$\phi(z) = az - \frac{b}{2}z^2,$$

and

$$D(E) = \frac{d}{2}(Mz_A)^2.$$

Then the utility function described in (3.95) becomes

$$u(z; z_N, \gamma, \omega) = az - \frac{b}{2}z^2 - (c+t)z + (y + tz_A) - \frac{d}{2}(Mz_A)^2 - \gamma|z - z_N| + \omega. \tag{3.106}$$

Next let us assume that there are two types of individuals in the population and both have some propensity to act morally. The first type has a high propensity to act morally ($\mu_H > 0$) and the other type a low propensity to act morally ($\mu_H > \mu_L > 0$). Then we find that the norm-consistent interval for each type is defined as

$$[\underline{z}_{\mu_i}, \overline{z}_{\mu_i}] = \left[\frac{a-c}{b+\mu_i M^2 d} - \frac{(1-\mu_i)\gamma}{b+\mu_i M^2 d}, \frac{a-c}{b+\mu_i M^2 d} + \frac{(1-\mu_i)\gamma}{b+\mu_i M^2 d}\right], \qquad i = L, H, \tag{3.107}$$

where the level chosen without adherence to a norm is given by

$$z_{\mu_i} = \frac{a-c}{b+\mu_i M^2 d}, \qquad i = L, H. \tag{3.108}$$

From the above it becomes evident that the propensity to act morally does not just define the baseline level of consumption, but also influences the width of the norm-consistent interval. If the two norm-consistent intervals do not overlap and we only have a single norm, then we know that this norm will lie in between the two intervals at the average level of consumption, with the high type consuming $\overline{z}_{\mu_H}$ and the low type consuming $\underline{z}_{\mu_L}$. For the norm-consistent intervals not to overlap, we require that $\overline{z}_{\mu_H} < \underline{z}_{\mu_L}$. If we denote the share of the population that is 'high' type by $0 < \pi < 1$, the norm that emerges is

$$
\begin{aligned}
z_N &= \pi \overline{z}_{\mu_H} + (1-\pi)\underline{z}_{\mu_L} \\
&= \pi \left[\frac{a-c}{b+\mu_H M^2 d}\right] + (1-\pi)\left[\frac{a-c}{b+\mu_L M^2 d}\right] \\
&\quad + \pi \left[\frac{(1-\mu_H)\gamma}{b+\mu_H M^2 d}\right] - (1-\pi)\left[\frac{(1-\mu_L)\gamma}{b+\mu_L M^2 d}\right].
\end{aligned} \tag{3.109}
$$

Note that the first two terms in (3.109) describe the average consumption across the population if there were no desire for conformity. Let us denote that level by z_A. Then we can derive that $z_N < z_A$ if the sum of the last two terms above is negative. Therefore we have $z_N < z_A$ if

$$
\left[\frac{\pi}{1-\pi}\right]\left[\frac{1-\mu_H}{1-\mu_L}\right] < \frac{b+\mu_H M^2 d}{b+\mu_L M^2 d}. \tag{3.110}
$$

If we assume now that half of the population is high type and half is low type (i.e. $\pi = 0.5$), then it is easy to show that the above condition will always hold as long as $\mu_H > \mu_L$, which we have of course assumed at the start. This is turn means that we know that we always have

$$
\frac{1-\mu_H}{1-\mu_L} < \frac{b+\mu_H M^2 d}{b+\mu_L M^2 d}.
$$

Since the terms on both sides are positive and greater than one, we can also derive that the condition in (3.110) will always hold if $\frac{\pi}{1-\pi} \leq 1$. This is of course the case for any $\pi \leq 0.5$. This means that the smaller the share of the population with a high propensity to act morally, the more average consumption will decrease as a result of individuals' desire for conformity. On the other hand, this also means that if the share of the population with high propensity to act morally is large, the desire for conformity may actually increase average consumption compared to the case without desire for conformity. Note that this result is specific to the case where the norm-consistent intervals do not overlap. If they were to overlap the resulting norm could take any value in the overlap area and therefore we are not able to make more specific predictions about the size of the change consumption relative to the case where individuals do not adhere to a norm.

Result 10 *Adherence to a norm can decrease overall average consumption of the dirty good if the population has a relatively high share of individuals with low propensity to act morally. However, if the population has a high share of individuals with high propensity to act morally, adherence to a norm can increase overall average consumption.*

Stage 1 - Decision over adherence

As noted in the introduction to this section, morality only enters at the consumption decision stage and the decision whether to adhere to a norm or not is done purely on a utility-maximisation basis. This means the individual will also take into account the effect of the propensity to act morally and the consequences this has for the utility achieved. An individual with propensity to act morally, $0 \leq \mu \leq 1$, will choose to adhere to a norm if

$$\phi(z^{norm}) - (c+t)z^{norm} - \gamma|z^{norm} - z_N| + \omega > \phi(z_\mu) - (c+t)z_\mu, \tag{3.111}$$

where z^{norm} is the consumption level under adherence to a norm (given the individual's propensity to act morally) and z_μ is the consumption level without adherence to a norm. To explore this further, let us go back to the three cases depicted in Figure 3.5. First, an individual with $\mu = 1$ will consume $\hat{z}$ regardless of what the norm is and regardless of whether he adheres to a norm or not. As such, if there is a norm that is equal to $\hat{z}$, then the individual will always consume at that norm level and thus all individuals with $\mu = 1$ will choose to adhere to the norm. Then the individual gains ω in utility but has no cost to utility other than the one the individual occurs anyway due to their morality.

Result 11 *If there is a norm at the social optimum $\hat{z}$, all individuals with full propensity to act morally ($\mu = 1$) will choose to adhere to a norm.*

However, if there is no norm equal to the social optimum, the individual with full propensity to act morally will only choose to adhere to the norm if

$$\omega > \gamma|\hat{z} - z_N|. \tag{3.112}$$

As such, if the norm level is sufficiently far away from the social optimum, the individual may choose not to adhere to the norm. Next let us look at the case from Figure 3.5 where there are two types, one with no propensity to act morally (blue) and the other with some propensity to act morally (red). The norm-consistent intervals do not overlap and there is a single norm. Then the individual with no propensity to act morally ($\mu = 0$) will choose to adhere to the norm if

$$\omega - \gamma(\underline{z}_0 - z_N) > \phi(z_0) - \phi(\underline{z}_0) - (c+t)(z_0 - \underline{z}_0), \tag{3.113}$$

and the individual with some propensity to act morally will choose to adhere to the norm if

$$\omega - \gamma(z_N - \overline{z}_\mu) > \phi(z_\mu) - \phi(\overline{z}_\mu) - (c+t)(z_\mu - \overline{z}_\mu). \tag{3.114}$$

While these two conditions look similar, they have a significant difference. The right hand side of (3.113) is positive, but the right hand side of (3.114) is negative. This is because adhering to the norm increases consumption for the red type and this leads to an increase in the direct utility from consumption. This is important because it makes that type significantly more likely to choose to adhere to the norm. Of course our example assumed that both types are adhering to the norm, which created that norm in the first place. However, the key point to take away from this is that for individuals with propensity to act morally, adhering to a norm may increase their consumption of the dirty good and this will to some degree offset the Utility Cost they are incurring due to their morality. As such an individual who is fully aware of the fact that they will sacrifice utility in consumption may mitigate his morality through the utility-maximising choice of whether to adhere to a norm or not.

Result 12 *An individual with some, but not full propensity to act morally may consume more of the dirty good when adhering to a norm than if they weren't adhering to norm. Therefore an individual may choose to adhere to a norm in order to mitigate the effect of the individual's morality to some degree as it decreases the Utility Cost incurred compared to the case where the individual does not adhere to a norm.*

3.7.4 Optimal Tax

The final step is to evaluate what the existence of moral behaviour and adherence to a norm means for determining the optimal tax. Of course we know that for individuals who do not adhere to a norm, the tax inducing socially optimal consumption is the same standard Pigovian tax as defined in (3.10) for everybody regardless of their propensity to act morally. The question is whether this still holds when individuals adhere to a norm. To start, let us look at what happens to the norm-consistent interval of a moral individual when we impose the standard tax. The norm-consistent interval described in (3.105) then becomes

$$\Big[\phi'(\underline{z}_\mu) = c + MD'(M\hat{z}) + (1-\mu)\gamma \quad , \quad \phi'(\overline{z}_\mu) = c + MD'(M\hat{z}) - (1-\mu)\gamma\Big]. \tag{3.115}$$

This shows that with the tax the norm-consistent interval of consumption centres around the socially optimal level of consumption for all individuals regardless of their propensity to act morally. However, the width of the norm-consistent interval depends on the value of the individual's propensity to act morally as well as the strength of adherence to the norm. This is because the tax has not internalised the norm component from the Utility Cost calculation and therefore the propensity to act morally still matters in determining how much weight is given to the norm compared to the Kantian optimum (which ignores adherence to norm). Therefore we still have the case that the higher an individual's propensity to act morally, the narrower the norm-consistent interval of consumption. Also we can see from (3.115) that only individuals with full propensity will definitely consume the social optimum under this tax.

However, this also means that effectively all individuals are identical under this tax except for the size of their norm-consistent interval. We know from Ulph and Ulph (2014) that if individuals only differ in their strength of adherence to the norm, there will almost certainly be a single norm and that will be within the norm-consistent interval of the type with the lowest strength of adherence to the norm. In our case this means that there will be a single norm and it lies within the narrowest norm-consistent interval. If we assume that we have at least some individuals with $\mu = 1$, we know that the narrowest interval is no interval at all around the social optimum and therefore we know that the emerging norm will almost certainly be at the social optimum. This in turn means that everybody will consume at that norm level and therefore the first-best solution is achieved through the standard Pigovian tax as in the case without adherence to the norm. Therefore the same tax induces the first best solution for all individuals who adhere to the norm and those who do not adhere to a norm. Finally, since imposing the tax induces everybody who adheres to a norm to consume the single norm at the social optimum, we can also derive that all individuals will choose to adhere to the norm, as they will gain the private benefit derived from conformity, ω, but incur no cost compared to the case where they do not adhere to a norm.

Proposition 9 *If there is at least one individual with $\mu = 1$ and individuals only differ in the propensity to act morally and the strength of adherence to the norm, the standard Pigovian tax of $\hat{t} = MD'(M\hat{z})$ induces everyone to choose to adhere to a norm and consume at the single norm level that emerges at the socially optimal level of consumption. As such the tax induces the first-best solution for all individuals regardless of their propensity to act morally and the degree to which they value conformity.*

Proof: Plugging the Pigovian tax as defined in (3.10) into (3.105) the norm-consistent interval of consumption for an individual with propensity to act morally $0 \leq \mu \leq 1$ emerges as defined in (3.115). Therefore, for each individual the norm-consistent interval of consumption is centred around the socially optimal consumption level $\phi'(\hat{z}) = c + MD'(M\hat{z})$. The individual's propensity then only affects the size of the norm-consistent interval. For an individual with $\mu = 1$, we can see from (3.115) that the interval is reduced to only the socially optimal level. We also know that since the norm level is determined by average consumption of all individuals adhering to that norm that, if individuals only differ in their strength of adherence to the norm, there will be a single norm within the narrowest norm-consistent interval. Since the narrowest interval is the social optimum, the norm has to be at the social optimum and all individuals adhering to the norm will consume the norm level. Furthermore, since the consumption level under adherence to the norm is the same as the utility-maximising level given the tax imposed, all individuals will choose to adhere to the norm and increase their utility by their private benefit derived from conformity, ω.

3.8 Concluding Remarks

The aim of this chapter was to develop an alternative theory of moral behaviour and altruism in an environmental context and evaluate what implications such behaviour may have for environmental policy. It has shown that when people have some propensity to act morally they will cut back consumption of an environmentally harmful good. At the same time, altruistic concern for the utility of others can contribute to the amount that individuals cut consumption, but only if they also have some propensity to act morally. The optimal tax on the dirty good remains the same as under standard theory and the first-best solution can be achieved. A key reason for this is that, even though some individuals may cut consumption towards to socially optimal level for moral reasons, we are still only dealing with market failure, which the Pigovian tax can correct for all individuals. However, as demonstrated in Section 3.3.5 this result hinges on individuals correctly making the (sophisticated) calculation of the hypothetical moral value when acting in this genuinely altruistic way. If this is not the case there will be both a market and behavioural failure. Consequently the Pigovian tax will no longer induce the welfare-maximising social optimum and may indeed reduce welfare compared to the initial tax or even no tax on the dirty good at all. While this chapter has only briefly addressed such behavioural failures, further work may be useful to investigate what second-best solution could be achieved through an emissions tax alone. Furthermore, it may be interesting to analyse other policy instruments that may correct for, or avoid, behavioural failures such as the one in Section 3.3.5.

Section 3.4 has demonstrated that the main results of the baseline model also hold with any concave choice function of the hypothetical Moral Benefit and the Utility Cost.[53] In particular it has shown that the optimal tax is always the standard Pigovian tax. The intuition behind this is that the tax makes the social optimum the same as the utility-maximising choice and therefore eliminates the Utility Cost for an individual. At the same time it also makes the utility-maximising choice equal to the Kantian optimum and so everybody will choose the social optimum regardless of their propensity to act morally. When individuals choose consumption of a range of different dirty goods as analysed in Section 3.5, the consumption choice is not dissimilar from a standard consumption problem, but since the individual is not determining the choice through maximisation of the utility function, the consumption choices are determined by a marginal rate of moral substitution of the relative level of the marginal Moral Benefit and marginal Utility Cost of each good. However, there are still the usual price effects and an increase in the tax on one good will decrease consumption of that good. At the same time, whether an increase in the tax on one good will increase consumption of another good depends on whether the substitution effect outweighs the Utility Cost effect of increasing consumption of that good. Following this, Section 3.6 extended the model to the case of heterogenous preferences. This required an adjustment of the Kantian question an individual uses to determine the hypothetical Moral Benefit. With identical preferences the individual could simply ask what would be optimal if everybody consumed the same amount of the dirty good. However, with heterogenous preferences the equivalent calculation of a Kantian optimum would become social welfare maximisation problem rather than a hypothetical gain in the individual's utility. In order not to remove the consumption decision from the individual's perspective entirely, it is assumed that the individual simply asks what would be optimal if they and everybody else moved proportionately by the same amount from the utility-maximising level to the social optimum. While this imperfect calculation can be regarded as a type of behavioural failure, and affects the degree to which individuals with propensity to cut back their consumption in the absence of the optimal tax, it does not change the optimal tax that the government should set. As a final extension to the baseline model, Section 3.7 combined the model of moral behaviour with a model of desire for conformity as developed in Ulph and Ulph (2014). The analysis shows that if people exhibit this desire for conformity, individuals with propensity to act morally may adhere to a norm that makes them consume more of the dirty good than if they

[53] The only key result that is different is that that altruism does not necessarily have to lead to a reduction of consumption with a choice function non-linear in the hypothetical Moral Benefit.

were not adhering to a norm. And this may also influence some individuals' decision whether to adhere to a norm or not since the increase in consumption may mitigate some of the loss in utility from their propensity to act morally. However, if there is at least one individual with propensity to act morally and individuals only differ in their propensity to act morally and the strength of their desire for conformity, the standard Pigovian tax will still induce everyone to consume the social optimum.

A key assumption of the model developed in this chapter is that the propensity to act morally for each type of individual is a given and static parameter. There are two questions that arise from this observation. First, where does the propensity to act morally come from? This question is researched in a number of disciplines including psychology, sociology, neuroscience and evolutionary biology (see for example Heinrichs et al. 2013 for a variety of contributions on moral motivation from different fields of research). Extensions to this chapter could make more detailed links between the insights from those disciplines and the model developed in this chapter. And second, can the propensity to act morally be influenced by extrinsic incentives on the dirty good? This addresses whether the propensity to act morally can be crowded-in or crowded-out by extrinsic incentives. Crowding of intrinsic motivation has received a lot of attention in the literature on pro-social behaviour due to its potential policy relevance.[54] Both of these questions raise the issue of how the propensity to act morally could change over time and it may be interesting to develop a dynamic model of moral behaviour to address this. In this context it could be interesting to evaluate whether a temporary tax can crowd-in individuals' moral behaviour such that the tax becomes redundant at some point or whether the opposite will occur and the tax crowds out the propensity to act morally, thus making it even more important that the government continues to set the right tax. Furthermore this raises the issue of whether, if the government can undertake some action to increase the propensity to act morally in consumers, there is a case that this may be preferable compared to setting a higher tax. Of course to develop such a model with reasonable assumptions would require a thorough understanding of the psychological factors that can influence an individual's propensity to act morally. In addition, there may be scope to further develop the model of moral behaviour in combination with conformity and social norms. For example, it is conceivable that the propensity to act morally is a function of the share of the population exhibiting some propensity to act morally. This also relates to the issue of crowding of intrinsic motivation. For example, if the government sets the right tax this may over time erode the social norm that determines individuals' propensity to act morally and may make others also less inclined to act out of moral consideration.

[54] See Section 3.2.3 for an overview of the literature on crowding of intrinsic motivation.

At the same time, a temporary tax could create a norm of moral behaviour that may be able to sustain itself even when that tax is removed.[55]

[55] This is ine line with the theory proposed by Nyborg et al. (2006) as described in Section 3.2.1.

References

Abbott, A., Nandeibam, S., and Shea, L. O. (2013). Recycling : Social Norms and Warm-Glow Revisited. *Ecological Economics*, 90:10–18.

Akerlof, G. A. (1980). A Theory of Social Custom, of Which Unemployment May be One Consequence. *The Quarterly Journal of Economics*, 94(4):749–775.

Akerlof, G. A. and Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, CXV(3):715–753.

Allcott, H. (2011). Social norms and Energy Conservation. *Journal of Public Economics*, 95(9-10):1082–1095.

Altemeyer-Bartscher, M., Markandya, A., and Rübbelke, D. T. G. (2011). *The Private Provision of International Impure Public Goods: the Case of Climate Policy.* Basque Centre for Climate Change (BC3), Bilbao, Spain.

Altemeyer-Bartscher, M., Markandya, A., and Rübbelke, D. T. G. (2014). International Side-Payments to Improve Global Public Good Provision when Transfers are Refinanced through a Tax on Local and Global Externalities. *International Economic Journal*, 28(1):71–93.

Altemeyer-Bartscher, M., Rübbelke, D. T. G., and Sheshinski, E. (2010). Environmental Protection and the Private Provision of International Public Goods. *Economica*, 77(308):775–784.

Amigues, J.-P., Favard, P., and Moreaux, M. (1998). On the Optimal Order of Natural Resource Use When the Capacity of the Inexhaustible Substitute Is Limited. *Journal of Economic Theory*, 80:153–170.

Andreoni, J. (1988). Privately Provided Public Goods in a Large Economy: The Limits of Altruism. *Journal of Public Economics*, 35:57–73.

Andreoni, J. (1989). Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence. *The Journal of Political Economy*, 97(6):1447–1458.

Andreoni, J. (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal*, 100(401):464–477.

Archibald, G. C. and Donaldson, D. (1976). Non-Paternalism and the Basic Theorems of Welfare Economics. *Canadian Journal of Economics*, 9(3):492–507.

Aronsson, T. and Blomquist, S. (2003). Optimal taxation, global externalities and labor mobility. *Journal of Public Economics*, 87:2749–2764.

Aronsson, T. and Johansson-Stenman, O. (2014). When Samuelson Met Veblen Abroad: National and Global Public Good Provision when Social Comparisons Matter. *Economica*, 81(322):224–243.

Aronsson, T. and Löfgren, K.-G. (2001). Green Accounting and Green Taxes in the Global Economy. In Folmer, H., Gabel, H. L., Gerking, S., and Rose, A., editors, *Frontiers of Environmental Economics*, chapter 1, pages 12–35. Edward Elgar, Cheltenham.

Arrow, K. J. (1986). Rationality of Self and Others in an Economic System. *Journal of Business*, 59(4):385–399.

Arrow, K. J. and Dasgupta, P. (2009). Conspicuous Consumption, Inconspicuous Leisure. *The Economic Journal*, 119:F497–F516.

Azar, O. H. (2004). What Sustains Social Norms and How They Evolve? The Case of Tipping. *Journal of Economic Behavior & Organization*, 54:49–64.

Barrett, S. (1990). The Problem of Global Environmental Protection. *Oxford Review of Economic Policy*, 6(1):68–79.

Barrett, S. (1994). Self-Enforcing International Environmental Agreements. *Oxford Economic Papers*, 46:878–894.

Becker, G. S. (1974). A Theory of Social Interactions. *Journal of Political Economy*, 82(6):1063–1093.

Becker, G. S. (1981). Altruism in the Family and Selfishness in the Market Place. *Economica*, 48(189):1–15.

Bénabou, R. and Tirole, J. (2006). Incentives and Prosocial Behavior. *The American Economic Review*, 96(5):1652–1678.

Benchekroun, H. and Withagen, C. (2011). The Optimal Depletion of Exhaustible Resources: A Complete Characterization. *Resource and Energy Economics*, 33(3):612–636.

Benzion, U., Rapoport, A., and Yagil, J. (1989). Discount Rates Inferred from Decisions: An Experimental Study. *Management Science*, 35(3):270–284.

Bergstrom, T. (2009). *Ethics, Evolution, and Games Among Neighbors.* The Selected Works of Ted C Bergstrom, University of California, Santa Barbara.

Bergstrom, T., Blume, L., and Varian, H. (1986). On the Private Provision of Public Goods. *Journal of public economics*, 29:25–49.

Bernheim, B. D. (1994). A Theory of Conformity. *Journal of Political Economy*, 102(5):841–877.

Bernheim, B. D. and Rangel, A. (2007). Toward Choice-Theoretic Foundations for Behavioral Welfare Economics. *The American Economic Review*, 97(2):464–470.

Brekke, K. A., Kipperberg, G., and Nyborg, K. (2010). Social Interaction in Responsibility Ascription : The Case of Household Recycling. *Land Economics*, 86(4):766–784.

Brekke, K. A., Kverndokk, S., and Nyborg, K. (2003). An Economic Model of Moral Motivation. *Journal of Public Economics*, 87(9-10):1967–1983.

Brunner, S., Flachsland, C., and Marschinski, R. (2012). Credible Commitment in Carbon Policy. *Climate Policy*, 12(2):255–271.

Bruvoll, A. and Nyborg, K. (2004). The Cold Shiver of Not Giving Enough: On the Social Cost of Recycling Campaigns. *Land Economics*, 80(4):539–549.

Buchholz, W., Cornes, R., and Rübbelke, D. T. G. (2012). Matching as a Cure for Underprovision of Voluntary Public Good Supply. *Economics Letters*, 117(3):727–729.

Buchholz, W., Falkinger, J., and Rübbelke, D. T. G. (2014). Non-Governmental Public Norm Enforcement in Large Societies as a Two-Stage Game of Voluntary Public Good Provision. *Journal of Public Economic Theory*, 16(6):899–916.

Cabinet Office Behavioural Insights Team (2011). *Behaviour Change and Energy Use.* Cabinet Office, London.

Camerer, C. F., Loewenstein, G., and Rabin, M. (2004). *Behavioral Economics.* Princeton University Press, Ann Arbor.

Chakravorty, U. and Krulce, D. L. (1994). Heterogeneous Demand and Order of Resource Extraction. *Econometrica*, 62(6):1445–1452.

Chakravorty, U., Moreaux, M., and Tidball, M. (2008). Ordering the Extraction of Polluting Nonrenewable Resources. *American Economic Review*, 98(3):1128–1144.

Chaudhuri, A. (2011). Sustaining Cooperation in Laboratory Public Goods Experiments: a Selective Survey of the Literature. *Experimental Economics*, 14:47–83.

Cornes, R. and Sandler, T. (1984). The Theory of Public Goods: Non-Nash Behaviour. *Journal of Public Economics*, 23:367–379.

D'Arge, R. C. and Kogiku, K. C. (1973). Economic Growth and the Environment. *The Review of Economic Studies*, 40(1):61–77.

Dasgupta, P. (2000). Economic Progress and the Idea of Social Capital. In Dasgupta, P. and Serageldin, I., editors, *Social Capital: a Multi-faceted Perspective*, pages 325–424. TheWorld Bank, Washington, DC.

Dasgupta, P. (2008). Discounting Climate Change. *Journal of Risk and Uncertainty*, 37:141–169.

Dasgupta, P. and Heal, G. (1974). The Optimal Depletion of Exhaustible Resources. *The Review of Economic Studies*, 41(Symposium on the Economics of Exhaustible Resources):3–28.

Dasgupta, P. and Heal, G. (1979). *Economic Theory and Exhaustible Resources*. Cambridge University Press, Oxford.

Dasgupta, P., Southerton, D., Ulph, A., and Ulph, D. (2015). Consumer Behaviour with Environmental and Social Externalities: Implications for Analysis and Policy. *Environmental and Resource Economics*, pages 1–36.

Daube, M. and Ulph, D. (2016). Moral Behaviour, Altruism and Environmental Policy. *Environmental and Resource Economics*, 63(2):505–522.

Deci, E. L. (1971). The Effects of Externally Mediated Rewards on Intrinsic Motivation. *Journal of Personality and Social Psychology*, 18(1):105–115.

Easterlin, R. A. (1974). Does Economic Growth Improve the Human Lot? Some Empirical Evidence. In David, P. A. and Reder, M. W., editors, *Nations and Households in Economic Growth: Essays in Honor of Moses Abramowitz*, pages 89–125. Academic Press, New York.

Easterlin, R. A. (2001). Income and Happiness: Towards a Unified Theory. *The Economic Journal*, 111:465–484.

Ellingsen, T. and Johannesson, M. (2008). Pride and Prejudice: The Human Side of Incentive Theory. *The American Economic Review*, 98(3):990–1008.

Fontaine, P. (2008). Altruism, History of the Concept. In Durlauf, S. N. and Blume, L. E., editors, *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 2nd edition.

Forster, B. A. (1980). Optimal Energy Use in a Polluted Environment. *Journal of Environmental Economics and Management*, 7:321–333.

Frey, B. S. (1997). *Not Just for The Money : An Economic Theory of Personal Motivation*. Edward Elgar, Cheltenham.

Frey, B. S. (1999). Morality and Rationality in Environmental Policy. *Journal of Consumer Policy*, 22:395–417.

Frey, B. S. and Jegen, R. (2001). Motivation Crowding Theory. *Journal of Economic Surveys*, 15(5):589–611.

Frey, B. S. and Oberholzer-Gee, F. (1997). The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out. *The American Economic Review*, 87(4):746–755.

Galarraga, I. and Markandya, A. (2009). *Climate Change and Its Socioeconomic Importance*. Basque Centre for Climate Change (BC3), Bilbao, Spain.

Gaudet, G., Moreaux, M., and Salant, S. W. (2001). Intertemporal Depletion of Resource Sites by Spatially Distributed Users. *American Economic Review*, 91(4):1149–1159.

Gerlagh, R. (2011). Too Much Oil. *CESifo Economic Studies*, 57(1):79–102.

Hammond, P. J. (1987). Altruism. In Eatwell, J., Milgate, M., and Newman, P., editors, *The New Palgrave: A Dictionary of Economics*. Palgrave Macmillan, Basingstoke, 1st edition.

Hanley, N., Shogren, J. F., and White, B. (2013). *Introduction to Environmental Economics*. Oxford University Press, Oxford, 2nd edition.

Hargreaves Heap, S. P. (2013). Social Influences on Behaviour. In Mehta, J., editor, *Behavioural Economics in Competition and Consumer Policy*. Report from ESRC Centre for Competition Policy, University of East Anglia.

Harsanyi, J. C. (1955). Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *The Journal of Political Economy*, 63(4):309–321.

Harstad, B. (2012). Buy Coal! A Case for Supply-Side Environmental Policy. *Journal of Political Economy*, 120(1):77–115.

Heal, G. (1976). The Relationship between Price and Extraction Cost for a Resource with a Backstop Technology. *The Bell Journal of Economics*, 7(2):371–378.

Heinrichs, K., Oser, F., and Lovat, T., editors (2013). *Handbook of Moral Motivation: Theories, Models, Applications.* Vol. 1, Sense Publishers, Rotterdam.

Herfindahl, O. C. (1967). Depletion and Economic Theory. In Gaffney, M., editor, *Extractive Resources and Taxation*, pages 63–90. University of Wisconsin Press, Madison, WI.

Hoel, M. (1978). Resource Extraction and Recycling with Environmental Costs. *Journal of Environmental Economics and Management*, 5:220–235.

Hoel, M. and Kverndokk, S. (1996). Depletion of Fossil Fuels and the Impacts of Global Warming. *Resource and Energy Economics*, 18(2):115–136.

Holcomb, J. H. and Nelson, P. S. (1992). Another Experimental Look at Individual Time Preference. *Rationality and Society*, 4:199–220.

Holland, S. P. (2003). Set-up Costs and the Existence of Competitive Equilibrium when Extraction Capacity is Limited. *Journal of Environmental Economics and Management*, 46:539–556.

Hotelling, H. (1931). The Economics of Exhaustible Resources. *The Journal of Political Economy*, 39(2):137–175.

IPCC (2014). *Social, Economic, and Ethical Concepts and Methods.* Report of Working Group III.

Johansson, O. (1997). Optimal Pigovian Taxes under Altruism. *Land Economics*, 73(3):297–308.

Jones, S. R. G. (1984). *The Economics of Conformism.* Blackwell, Oxford.

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a Constraint on Profit Seeking: Entitlements in the Market. *The American Economic Review*, 76(4):728–741.

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

Kant, I. (1875). *Grounding for the Metaphysics of Morals.* Hackett, Indianapolis, 3rd edition.

Kemp, M. C. and Long, N. V. (1980). On Two Folk Theorems Concerning the Extraction of Exhaustible Resources. *Econometrica*, 48(3):663–673.

Kennett, D. A. (1980). Altruism and Economic Behavior, I: Developments in the Theory of Public and Private Redistribution. *American Journal of Economics and Sociology*, 39(2):183–198.

Laffont, J.-J. (1975). Macroeconomic Constraints, Economic Efficiency and Ethics: An Introduction to Kantian Economics. *Economica*, 42(168):430–437.

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *Quarterly Journal of Economics*, 112(2):443–477.

Lewis, T. R. (1982). Sufficient Conditions for Extracting Least Cost Resource First. *Econometrica*, 50(4):1081–1083.

Markandya, A. and Rübbelke, D. T. G. (2004). Ancillary Benefits of Climate Policy. *The Journal of Economics and Statistics*, 224(4):488–503.

Markandya, A. and Rübbelke, D. T. G. (2012). Impure Public Technologies and Environmental Policy. *Journal of Economic Studies*, 39(2):128–143.

Nolan, J., Schultz, P., Cialdini, R., Goldstein, N., and Griskevicius, V. (2008). Normative Social Influence is Underdetected. *Personality and Psychology Bulletin*, 34(7):914–923.

Nordhaus, W. D. (1994). *Managing The Global Commons: The Economics of Climate Change.* MIT Press, Cambridge, MA.

Nordhaus, W. D. (2006). After Kyoto: Alternative Mechanisms to Control Global Warming. *The American Economic Review*, 96(2):31–34.

Nordhaus, W. D. (2007). A Review of the Stern Review on the Economics of Climate Change. *Journal of Economic Literature*, XLV:686–702.

Nyborg, K., Howarth, R. B., and Brekke, K. A. (2006). Green Consumers and Public Policy: On Socially Contingent Moral Motivation. *Resource and Energy Economics*, 28(4):351–366.

Nyborg, K. and Rege, M. (2003). Does Public Policy Crowd out Private Contributions to Public Goods. *Public Choice*, 115(3):397–418.

Perman, R., Ma, Y., McGilvray, J., and Common, M. (2003). *Natural Resource and Environmental Economics*. Pearson Education, New York, 3rd edition.

Pigou, A. C. (1932). *The Economics of Welfare*. Macmillan & Co. Ltd., London, 4th edition.

Pittel, K. and Rübbelke, D. T. G. (2010). *Local and Global Externalities, Environmental Policies, and Growth*. Basque Centre for Climate Change (BC3), Bilbao, Spain.

Pollitt, M. G. and Shaorshadze, I. (2011). *The Role of Behavioural Economics in Energy and Climate Policy*. Cambridge Working Paper in Economics 1165, University of Cambridge.

Putnam, R. (2000). *Bowling Alone: The Collapse and Revival of American Community*. Simon and Schuster, New York.

Samuelson, W. and Zeckhauser, R. (1988). Status Quo Bias in Decision Making. *Journal of Risk and Uncertainty*, 1:7–59.

Schultz, P., Nolan, J., Cialdini, R., Goldstein, N., and Griskevicius, V. (2007). The Constructive, Destructive, and Reconstructive Power of Social Norms. *Psychological Science*, 18(5):429–434.

Schulze, W. D. (1974). The Optimal Use of Non-Renewable Resources: The Theory of Extraction. *Journal of Environmental Economics and Management*, 1:53–73.

Sen, A. K. (1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affairs*, 6(4):317–344.

Shogren, J. F. and Taylor, L. O. (2008). On Behavioral-Environmental Economics. *Review of Environmental Economics and Policy*, 2(1):26–44.

Simon, H. A. (1986). Rationality in Psychology and Economics. *Journal of Business*, 59(4):209–224.

Sinclair, P. (1992). High Does Nothing and Rising is Worse: Carbon Taxes Should Keep Declining to Cut Harmful Emissions. *The Manchester School*, 60(1):41–52.

Sinn, H. W. (2008). Public Policies Against Global Warming: A Supply Side Approach. *International Tax and Public Finance*, 15(4):360–394.

Smith, A. (1759). *The Theory of Moral Sentiments*. Oxford University Press, Oxford.

Smith, K. (2010). Stern, Climate Policy and Savings Rates. *Climate Policy*, 10(3):289–297.

Sobel, J. (2005). Interdependent Prand Reciprocity. *Journal of Economic Literature*, XLIII:392–436.

Solow, R. M. (1974). Intergenerational Equity and Exhaustible esources. *The Review of Economic Studies*, 41(Symposium on the Economics of Exhaustible Resources):29–45.

Solow, R. M. and Wan, F. Y. (1976). Extraction Costs in the Theory of Exhaustible Resources. *Bell Journal of Economics*, 7:359–370.

Stern, N. (2007). *Stern Review of the Economics of Climate Change*. HM Treasury, London.

Stern, N. (2008). The Economics of Climate Change. *American Economic Review*, 98(2):1–37.

Stiglitz, J. (1974). Growth with Exhaustible Natural Resources: Efficient and Optimal Growth Paths. *Review of Economic Studies*, 41(Symposium on the Economics of Exhaustible Resources):123–137.

Sugden, R. (1982). On the Economics of Philanthropy. *The Economic Journal*, 92(366):341–350.

Tahvonen, O. (1995). International CO2 Taxation and The Dynamics of Fossil Fuel Markets. *International Tax and Public Finance*, 2(2):261–278.

Tahvonen, O. (1997). Fossil Fuels, Stock Externalities, and Backstop Technology. *Canadian Journal of Economics*, 30(4):855–874.

Thaler, R. (1980). Toward a Positive Theory of Consumer Choice. *Journal of Economic Behavior & Organization*, 1:39–60.

Thaler, R. (1981). Some Empirical Evidence on Dynamic Inconsistency. *Economics Letters*, 8:201–207.

Thaler, R. (1999). Mental Accounting Matters. *Journal of Behavioral Decision Making*, 12:183–206.

Tirole, J. (2002). Rational Irrationality: Some Economics of Self-Management. *European Economic Review*, 46:633–655.

Ulph, A. and Ulph, D. (1994). The Optimal Time Path of a Carbon Tax. *Oxford Economic Papers*, 46(Special Issue on Environmental Economics):857–868.

Ulph, A. and Ulph, D. (2013). Optimal Climate Change Policies When Governments Cannot Commit. *Environmental and Resource Economics*, 56:161–176.

Ulph, A. and Ulph, D. (2014). *Consumption Decisions When People Value Conformity.* Discussion Paper No. 1414. School of Economics & Finance, University of St Andrews.

Ulph, D. (2006). *A Theory of Green Consumerism. (mimeo).* School of Economics & Finance, University of St Andrews.

van der Ploeg, F. and de Zeeuw, A. J. (1992). International Aspects of Pollution Control. *Environmental and Resource Economics*, 2(2):117–139.

van der Ploeg, F. and Withagen, C. (2012a). Is There Really a Green Paradox? *Journal of Environmental Economics and Management*, 64(3):342–363.

van der Ploeg, F. and Withagen, C. (2012b). Too Much Coal, Too Little Oil. *Journal of Public Economics*, 96(1-2):62–77.

Veblen, T. (1924). *The Theory of the Leisure Class: An Economic Study of Institutions.* George, Allen and Unwin, London.

White, M. D. (2003). Can Homo Economicus Follow Kant's Categorical Imperative? *The Journal of Socio-Economics*, 33(1):89–106.

Withagen, C. (1994). Pollution and Exhaustibiity of Fossil Fuels. *Resource and Energy Economics*, 16:235–242.