Optimal unselfishness in climate policy

Robert W. Hahn* Robert A. Ritz‡

2 October 2012

PRELIMINARY DRAFT — COMMENTS WELCOME

Abstract

Social preferences may play an important role in the provision of public goods, such as climate policy. We present a model of how unselfish a country optimally acts in its abatement effort, given an underlying preference for (a degree of) altruism. Due to free-riding and carbon leakage, a country almost always behaves more selfishly than its true preference. In some cases, it is optimal for a country that cares about global welfare to behave completely selfishly. Our analysis also highlights that it may be difficult to infer countries' true preferences for unselfishness from observed behaviour.

Keywords: Altruism, carbon leakage, climate policy, free-riding, social cost of carbon, social preferences, strategic delegation, unilateral commitment

JEL classifications: D03 (behavioural economics), H23 (externalities), H41 (public goods), Q58 (environmental policy).

*Director of Economics, Smith School of Enterprise and the Environment, University of Oxford; Chief Economist, Legatum Institute; Senior Fellow, Georgetown Center for Business and Public Policy. E-Mail: robert.hahn@smithschool.ox.ac.uk

†University Lecturer in Economics, University of Cambridge; Research Associate, Electricity Policy Research Group; Visiting Research Fellow, Oxford Institute for Energy Studies. E-Mail: rar36@cam.ac.uk

‡We would like to thank Toke Aidt, David Anthoff, Elizabeth Baldwin, Thomas Greve, Cameron Hepburn, Charles Mason, Robert Stavins, Alexander Teytelboym, Richard Tol, Alistair Ulph [*to be completed*] for helpful comments and advice, participants of the Energy & Environment seminar in Cambridge for a lively discussion, and Hemal Shah for valuable research assistance. The usual disclaimer applies.

1 Introduction

Recent years have witnessed a number of unilateral initiatives to combat climate change at the local, national, and regional levels. These initiatives aim to reduce greenhouse gas (GHG) emissions with varying levels of ambition. For example, the European Union has a program to reduce greenhouse gas emissions by 20 percent (relative to 1990 levels) by 2020, and the United Kingdom aims to cut emissions by 80 percent by 2050. More recently, in the United States, the state of California passed a law to reduce GHG emissions in 2020 to their 1990 level. Moreover, these initiatives have taken place in the absence of a binding global agreement by countries to jointly reduce emissions (for instance, by way of a global cap-and-trade scheme).

Similarly, there is an increasing use, in various jurisdictions, of the "social cost of carbon" in project evaluations and regulatory decision-making. The social cost of carbon reflects the marginal benefit to the world from reducing one ton of carbon dioxide (rather than only to an individual country or region). Several European countries, including the Netherlands, Finland, Italy and the United Kingdom, have already applied the social cost of carbon (Watkiss and Hope, 2012). The United States has also developed a measure of the social cost of carbon (Greenstone, Kopits and Wolverton, 2011), which has to date been applied in a limited program involving selected energy and environmental regulations. At the same time, however, many other countries do not incorporate the social cost of carbon in policymaking, and do not appear to have engaged in emissions abatement beyond "business-as-usual".

There is evidence that the domestic costs associated with some of these unilateral policies exceed their domestic benefits. For example, Tol (2011) conducts a cost-benefit analysis of the European Union's "20/20/2020" policy package, and finds that its benefit-cost ratio is less than unity across a range of scenarios. In a similar vein, the United Kingdom's Department of Energy and Climate Change's impact assessment of the 2008 Climate Change Act observes that "the benefits of UK action will be distributed across the globe," and also finds that "the economic case for the UK continuing to act alone where global action cannot be achieved would be weak" (DECC, 2009).

It is a central tenet of economics that individuals typically pursue their self-interest. It is hence difficult to reconcile such unilateral initiatives with the standard theory of environmental economics, including the theory of international environmental agreements (Barrett, 1994). Put simply, if unilateral action by local, national, or regional actors reduces their own domestic welfare, then why are they engaging in such initiatives? However, there is also a growing recognition that social (other-regarding) preferences may be important in explaining empirical observations on public good provision in general, and climate policy in particular (Kolstad, 2012). Indeed, some of the above initiatives appear to reflect "unselfish" or "altruistic" motives, in the sense of incorporating benefits that accrue outside the borders of the acting jurisdiction.

Based on this motivation, we study the role of unselfishness in a model of public good provision, characterized by free-riding and applied to the problem of climate policy.[1] In particular,

[1] We do not wish to claim that social preferences are the only possible way of explaining unilateral climate action. In some cases, other explanations, such as domestic political-economy considerations, may be important. However, it seems instructive to better understand the role that altruism can play in public-good provision.

given that a country has a preference for some degree of unselfishness, to what extent is this optimally reflected in its abatement efforts? To our knowledge, this is the first attempt in the literature to model and understand the notion of "optimal" unselfishness.

Our analysis is based on a simple two-country model of non-cooperative emissions abatement. A country's national welfare Π_i is given by the benefits it derives from global abatement minus the cost of its own national abatement effort. Global welfare is the sum of the two individual countries' net benefits, $W = (\Pi_1 + \Pi_2)$. The key feature of our model is that a country's true preference may depart from national self-interest. We assume that a country's true objective function $S_i = (1 - \theta_i)\Pi_i + \theta_i W$ can place weight on both its own national welfare and on global welfare, where $\theta_i \in [0, 1]$ represents its degree of unselfishness.[2]

Our main question is to what extent a country will actually wish to commit to engaging in abatement according to its true preference. To examine this, we introduce a strategic objective function for each country as $\Omega_i = (1 - \lambda_i)\Pi_i + \lambda_i S_i$, where $\lambda_i \in [0, 1]$ is its strategic preference. A country chooses its abatement effort according to its true preference if $\lambda_i = 1$, but otherwise acts more selfishly than would be its true preference. Our analysis focuses on determining a country's optimal commitment λ_i^* (at Date 1) to incorporate its unselfishness into abatement decision-making (at Date 2).[3]

A natural way of thinking about how a country can commit to pursuing abatement in a way that may depart from its true preferences is in terms of the theory of strategic delegation.[4] A country's true preference $\theta_i \in [0, 1]$ reflects its citizens' preferences towards global climate change, such as those of its median voter. Citizens delegate decision-making regarding abatement targets to politicians, where different politicians represent different abatement policies. For strategic reasons, as we shall see, citizens may wish to elect politicians whose climate-policy preference differs from their own, $\lambda_i \neq 1$.

Our model has two basic features. First, more unselfish behaviour by a country is associated with an increase in its national abatement effort (Lemma 1). Second, an increase in one country's abatement induces the other country to cut back its abatement effort—this is the standard free-riding effect that leads to carbon leakage (Lemma 2). However, the increase in national abatement does lead to an increase in global abatement (so its effort is partially offset but not overturned).[5]

We begin by showing that a small commitment towards unselfish behaviour (formally, $d\lambda_i > 0$ starting from $\lambda_1 = \lambda_2 = 0$), in equilibrium, has an ambiguous effect on a country's true objective S_i^* (Proposition 1). Intuitively, a unilateral commitment by country i increases the net benefit Π_j^* enjoyed by the other country j but acting unselfishly hurts its own net benefit Π_i^*.[6] The

[2] On a historical note, Edgeworth (1881) uses essentially the same formulation of altruism as ours, by writing $S_i = \Pi_i + \theta_i \Pi_j$ and calling θ_i the "coefficient of effective sympathy".

[3] In the literature on international environmental agreements, countries at Date 1 choose whether or not to join an agreement ("in or out"). By contrast, in our model, countries choose the *intensity* of their policy commitments. Our analysis shows that it is never optimal for a country to overstate its true unselfishness (ruling out $\lambda_i^* > 1$).

[4] See Roelfsema (2007) for a recent analysis of strategic delegation in environmental policy-making in a different setting; Vickers (1985) gives a seminal formalization of delegation in the context of the theory of the firm.

[5] While our results rely on leakage being strictly positive, their qualitative nature does not depend on whether leakage rates are "high" or "low".

[6] It is perhaps not immediately obvious that greater abatement reduces a country's net benefit Π_i^* since it

commitment thus enhances its own true objective S_i^* if (and only if) the former effect outweighs the latter. The positive effect will be large if country j's marginal benefit from abatement is large, and receives large weight according to country i's degree of unselfishness θ_i. The negative effect will be small if there is little free-riding, resulting in carbon leakage, by country j, and if country i's own marginal benefit from abatement is small. The same logic shows that it is also ambiguous whether more unselfish behaviour leads to an increase or decrease in global welfare W^* (Proposition 2).[7] These two results already make clear that whether incorporating unselfishness in decision-making is welfare-augmenting depends crucially on the details of the environment.

We then turn to the question of optimal commitments. We show that a full commitment by both countries to incorporating their unselfishness in decision-making ($\lambda_1^* = \lambda_2^* = 1$) is an equilibrium if and only if both countries also have entirely unselfish true preferences ($\theta_1 = \theta_2 = 1$). In this case, both countries engage in the first-best levels of abatement that incorporate the global benefit of their actions, so the first-best outcome is achieved (Proposition 3). However, if at least one country is partially selfish ($\theta_1 < 1$ *or* $\theta_2 < 1$), then it is optimal for *both* countries to not act according to their true preferences ($\lambda_1^* < 1$ *and* $\lambda_2^* < 1$).

For example, suppose that country 1's true preference is to be entirely unselfish, so $\theta_1 = 1$, while country 2 is unselfish only to some degree, so $\theta_2 < 1$. Then the *optimal* commitment by country 1 satisfies $\lambda_1^* < 1$, and a full commitment is dominated by a weaker policy. So the optimal way for country 1 to maximize global welfare W is to maximize a strategic objective $\Omega_1 = (1 - \lambda_1^*)\Pi_1 + \lambda_1^* W$ that is partially skewed towards it national welfare. In other words, *a country that genuinely wants to maximize global welfare actually does best by being at least somewhat selfish.* This result is a manifestation of the general insight due to Schelling (1960) that, in strategic settings, the maximum A may be achieved by maximizing B.

Intuitively, why can country 1 do better than playing according to its true, entirely unselfish preference? The key is that a small decrease in its own level of abatement (from the full commitment level) only leads to a second-order loss in global welfare. However, the resulting induced *increase* in the other country's abatement level leads to a first-order gain in global welfare (whenever the other country is not already choosing the first-best abatement). So the reason why a full commitment is almost never optimal is essentially "reverse leakage"—a weaker commitment reduces free-riding by the other country.

We also provide conditions under which a zero commitment is optimal for one or both countries (Proposition 4), and a characterization of the optimal interior commitments $(\lambda_1^*, \lambda_2^*) \in (0, 1)^2$ (Proposition 5). A reoccurring theme in our analysis is that it is difficult for a country to follow through on its unselfishness if it has a relatively high marginal benefit of abatement. In this case, its unilateral action hurts its own national welfare by a lot relative to the extent to

increases abatement costs but also increases benefits (since global abatement rises). However, the result becomes clear when instead considering the two negative effects at the margin: First, going beyond the selfish level of abatement entails a negative *marginal* net benefit; second, the other country is induced to decrease its abatement effort which also reduces benefits.

[7] Hoel (1991) obtains a similar result to our Proposition 2 in a model of unilateral commitment that does not incorporate social preferences.

which it helps the other country's welfare.

We highlight that *caution is required in inferring whether or not a country is "being selfish" from its observed behaviour.* Suppose it is observed or otherwise estimated that a country's abatement efforts appear to be entirely selfish. It does *not* follow that this country's underlying true preference is to be completely selfish. As already noted, little or no additional abatement can be consistent even with fully altruistic true preferences. So the question "How unselfish is country A?" is not that easy to answer based on its observed abatement behaviour. Moreover, we show that there may be a non-obvious relationship between countries' true preferences for unselfishness and their strategic actions. A country with a higher true preference for unselfishness may, in equilibrium, be the country whose abatement actions places *less* weight on global welfare (Proposition 6). So the question "Is country A more unselfish than country B?" also has no straightforward answer based on countries' observed behaviour.

The remainder of the paper proceeds as follows. Section 2 sets up the model and conducts some preliminary analysis. Section 3 examines the welfare impact of small commitments. Section 4 analyzes in detail countries' optimal commitments.

Section 5 presents some further results and extensions to the benchmark model. First, we show that, under some conditions, more unselfish abatement actions may be associated with *higher* rates of carbon leakage (Proposition 7). Put differently, unselfish behaviour here actually worsens the free-riding problem at the margin. This is somewhat unexpected as it seems natural to think that less selfish behaviour by countries will tend to mitigate carbon leakage. The more general point is that rates of carbon leakage—though useful and important—are not always reliable welfare indicators.

Second, we explore the impact of sequential commitments, where one country acts as a leader in terms of formulating its policy commitment. We show that, in general, it is ambiguous whether countries' commitments are strategic substitutes or complements (Lemma 5) and also whether sequential commitments increase or reduce welfare (relative to our benchmark case with simultaneous commitments).

Third, we suggest that an additional mechanism that guarantees cost-effectiveness of abatement at a global level can act as a substitute for other countries' unselfishness in terms of enabling strong unilateral commitment (Lemma 6). This provides a potential justification for a global cap-and-trade scheme even if a subset of countries in the scheme does not engage in any significant abatement. The purpose of the scheme is to encourage cost-effectiveness, as is standard, but also to enable unselfish countries to follow through on their preferences.

Fourth, we show that our main insights also go through with different formulations of countries' social preferences (that is, of the true objective function S_i) such as (i) the "warm glow" that may be associated with public good contributions (Andreoni, 1990) and (ii) settings in which a country is willing to "ignore" some of the abatement costs it incurs.

Section 6 offers concluding remarks and some directions for future research.[8]

[8] Most of our proofs are in the appendix.

2 Model

□ **Model setup**. Two countries, 1 and 2, can perform emissions abatement which we denote X_i for $i = 1, 2$. Country i's abatement benefits $B_i(X_1 + X_2)$ depend on the aggregate level of abatement by the two countries. The marginal benefit of abatement is positive but decreasing, $B_i'(\cdot) > 0$ and $B_i''(\cdot) < 0$. Different benefit functions could reflect, for example, differences in the consequences of climate change across countries or different emphasis on the scientific evidence for global warming.

The cost of emissions abatement $C_i(X_i)$ is country-specific. The marginal cost of abatement is positive and increasing, $C_i'(\cdot) > 0$ and $C_i''(\cdot) > 0$. Different cost functions reflect differences in production technologies across countries. To guarantee an interior solution for abatement levels, we assume that the cost of abatement satisfies $C_i(0) = C_i'(0) = 0$ for $i = 1, 2$.

Define a country's *net* benefits from abatement as

$$\Pi_i = B_i(X_1 + X_2) - C_i(X_i). \tag{1}$$

Global welfare is given by the sum of the two countries' net benefits,

$$W = \Pi_1 + \Pi_2. \tag{2}$$

In our model, each country's preferences towards climate change may be at least partly unselfish. In particular, country i's *true* objective function is given by

$$S_i = (1 - \theta_i)\Pi_i + \theta_i W, \tag{3}$$

where the parameter $\theta_i \in [0, 1]$ represents its true preference, that is, its degree of unselfishness. This objective function thus allows countries to place weight on their own national welfare as well as on global welfare.

The standard case where a country is completely self-interested is nested where $\theta_i = 0$, and so $S_i = \Pi_i$. The case in which a country's preference is entirely unselfish is nested where $\theta_i = 1$, and so $S_i = W$. The country's preference then reflects the full global benefit of abatement, $(B_1 + B_2)$. We can interpret this as an underlying preference to incorporate the "social cost of carbon" in abatement decisions. More generally, a higher value of θ_i represents a country that is less selfish and places greater weight on the other country's net benefit.

Our main question is to what extent a country will actually wish to commit to choosing its level abatement activity according to its true preference as given by S_i. To examine this, we introduce a *strategic* objective function for each country as

$$\Omega_i = (1 - \lambda_i)\Pi_i + \lambda_i S_i. \tag{4}$$

A strategic preference is a convex combination of a country's national welfare Π_i and of its true preference S_i, with a relative weight given by its strategic preference $\lambda_i \in [0, 1]$. If $\lambda_i = 0$, then the country's strategic objective is to be entirely selfish, so $\Omega_i = \Pi_i$ (regardless of its underlying true objective S_i). By contrast, if $\lambda_i = 1$, then the country's strategic objective is identical to its true objective, so $\Omega_i = S_i$. The parameters $(\lambda_1, \lambda_2) \in [0,1]^2$ are a modelling device with which we can represent to what extent countries will wish to follow through on their true preferences towards climate change in a strategic setting.

A natural way of thinking about how a country can commit to pursuing abatement in a way that departs from its true preferences is in terms of the theory of strategic delegation. A country's true preference $\theta_i \in [0,1]$ reflects its citizens' preferences towards global climate change, in particular those of its median voter. Citizens delegate decision-making regarding abatement targets to politicians, where different politicians represent different abatement policies. For strategic reasons, as we shall see, citizens may wish to elect politicians whose climate-policy preference differs from their own, that is $\lambda_i \neq 1$.

The timing of the model can be summarized as follows. At Date 0, each country is endowed with a benefit function and a cost function for abatement, $B_i(\cdot)$ and $C_i(\cdot)$, as well as with a true objective $S_i(\cdot)$ that reflects its degree of unselfishness, $\theta_i \in [0,1]$. Then, at Date 1, each country chooses its strategic preference $\lambda_i \in [0,1]$ to maximize its true objectives towards climate change, $\max_{\lambda_i} S_i$. Finally, at Date 2, each country chooses its level of abatement activity according to its strategic objective function, $\max_{X_i} \Omega_i$.[9]

Noting that $\Omega_i = \Pi_i + \lambda_i \theta_i \Pi_j$ $(j \neq i)$ reveals that the "first-best" benchmark is nested where (i) both countries have entirely unselfish true preferences, $\theta_1 = \theta_2 = 1$, *and* (ii) both countries commit to choosing their respective abatement levels accordingly, $\lambda_1 = \lambda_2 = 1$. In this case, the countries at Date 2 make abatement decisions to $\max_{X_i} W$, thus incorporating the full global benefit of their actions.

We focus on the subgame-perfect Nash equilibrium of the game, and follow the literature on strategic delegation in assuming that countries' strategic objective functions, Ω_1 and Ω_2, form credible commitments.[10]

□ **Preliminary analysis.** We begin by discussing some of the key properties of the model at Date 2, with a view to analyzing the welfare impact of strategic commitments by countries in the following two sections.

Country i's first-order condition for its level of abatement is

$$\frac{\partial \Omega_i}{\partial X_i} = (B_i' - C_i') + \lambda_i \theta_i B_j' = 0. \tag{5}$$

The two first-order conditions implicitly define country i's reaction function where its optimal abatement as a function of country j's abatement, $R_i(X_j)$, satisfies $\partial \Omega_i / \partial X_i = 0$. Differentiating

[9] This order of choices reflects the fact that a country's strategic position with respect to climate change is typically a long-term decision, while levels of abatement are short-term variables that are more easily changed.

[10] The latter assumption is essentially equivalent to assuming that countries' abatement levels (X_1, X_2) are observable. Given that their abatement benefit and costs functions are commonly known, this, in turn, is equivalent to country i's politicians knowing the other country j's $\lambda_j \theta_j$ when choosing their abatement policy. (We do not require that country i knows country j's true underlying preference θ_j for unselfishness.)

the first-order condition shows that the slope of the reaction function is given by

$$R_i'(X_j) = \frac{B_i'' + \lambda_i \theta_i B_j''}{\left(-B_i'' + C_i'' - \lambda_i \theta_i B_j''\right)}. \tag{6}$$

The countries' reaction functions will play an important role in our analysis. First, observe that reaction functions are downward-sloping, $R_i'(X_j) < 0$ since benefit functions are concave $B_i''(\cdot) < 0$ for $i = 1, 2$, so abatement levels are *strategic substitutes*. This captures the free-riding effect of abatement efforts: If one country increases abatement, this reduces the *marginal* benefit of abatement for the other country, which therefore responds by decreasing its abatement efforts. Second, although the other country responds by reducing abatement, the overall impact of an increase in one country's abatement effort is to increase global abatement (since $R_i'(X_j) > -1$).

In the context of climate policy, minus the value of the slope of country i's reaction function can be interpreted as the marginal rate of *carbon leakage* (see, e.g., IPCC, 2007) resulting from additional abatement by country j,

$$L_j \equiv \left[-R_i'(X_j)\right], \tag{7}$$

yielding the following result.

Lemma 1 *The marginal rate of carbon leakage resulting from additional abatement by country j is given by $L_j \equiv [-R_i'(X_j)] \in (0, 1)$ for any $(\theta_1, \theta_2) \in [0, 1]^2$* and any $(\lambda_1, \lambda_2) \in [0, 1]^2$.

Leakage rates in our model are thus always positive but less than 100%. Although there are some exceptions, this range reflects the large majority of existing theoretical and empirical work on carbon leakage. Many empirical estimates are derived from numerical simulations of multi-sector, general equilibrium models which focus on climate initiatives by OECD countries that result in carbon leakage to non-OECD countries. These models typically suggest leakage rates in the range 5–40%, with most estimates below 20%. Industry-level estimates of leakage for particular sectors (such as the cement and steel industries in the EU's emissions trading scheme) are often higher but only rarely exceed 100%.[11] In any case, while our results rely on leakage being strictly positive, their qualitative nature does not depend on whether leakage rates are "high" or "low".

The Nash equilibrium levels of abatement $(X_1^*(\lambda_1, \lambda_2), X_2^*(\lambda_1, \lambda_2))$ at Date 2 are defined by $X_1^* = R_1(X_2^*)$ and $X_2^* = R_2(X_1^*)$. The result from Lemma 1 ensures that the equilibrium is unique, stable, and exhibits well-behaved comparative statics.

We next establish the intuitive result that a more unselfish strategic objective by a country (higher λ_i) leads to an equilibrium increase in its own level of abatement activity (higher X_i^*).

[11] It is possible for leakage rates to be negative in trade-theoretic models (Copeland and Taylor, 2005) or with imperfect competition in product markets (Ritz, 2009). Conversely, it is possible for leakage rates to exceed 100% in some models, for example, if unregulated firms have significantly dirtier production technologies in settings with imperfect competition (Ritz, 2009) and if there is significant international relocation of production facilities (Babiker, 2005).

Lemma 2 *If country i's true preference $\theta_i > 0$, then its abatement level $dX_i^*/d\lambda_i > 0$ for any $\lambda_i \in [0,1]$.*

The reason for the result is simply that a more unselfish strategic preference, that is, higher λ_i, inflates the marginal return to abatement activity.

Strictly speaking, Lemma 2 does not apply to a country with a true preference which is entirely selfish, $\theta_i = 0$. To complete our preliminary discussion, we establish another lemma to show that such a country does not want to engage in such a strategic commitment in any case.

Lemma 3 *If country i's true preference $\theta_i = 0$, then its optimal commitment $\lambda_i^* = 0$.*

3 Small commitments

To build intuition, we begin our main analysis by considering the welfare impact of a "small commitment" by an individual country.

□ **Impact on true objective**. At Date 0, each country has been endowed with a true preference towards climate change, as indexed by its degree of unselfishness, $\theta_i \in [0,1]$. Initially, both countries act purely in their national self-interest, that is, $\lambda_i = 0$ (so $\Omega_i = \Pi_i$) for $i = 1,2$. What is the impact of a small commitment $d\lambda_i > 0$ by country i (at Date 1) towards incorporating its true preference S_i in decision-making on abatement at Date 2?

Proposition 1 *The impact of a small unilateral commitment $d\lambda_i > 0$ by country i on its equilibrium true objective*

$$\left.\frac{dS_i^*}{d\lambda_i}\right|_{\lambda_1=\lambda_2=0} = \left[\left(\theta_i B_j' - B_i' L_i\right)\frac{dX_i^*}{d\lambda_i}\right]_{\lambda_1=\lambda_2=0}$$

is (a) ambiguous in general, (b) negative for a ratio of marginal benefits B_i'/B_j' sufficiently large, (c) negative for a true preference θ_i sufficiently small, (d) positive if the ratio of marginal benefits satisfies $B_i' \leq B_j'$ and the true preference exceeds the leakage rate $\theta_i \geq L_i$.

Part (a) of the result shows us that it is not clear whether a country would in fact wish to go ahead with a small commitment. Moreover, parts (b) and (c) show that, if either it has a relatively large marginal abatement benefit, or its true preference towards climate change shows only a small amount of unselfishness, then it is never a good idea for the country to make such a commitment. By part (d), however, two simple conditions which are *jointly* sufficient for $dS_i^*/d\lambda_i \geq 0$ are that the country has a (weakly) lower marginal benefit as well as a true preference that exceeds the rate of carbon leakage.

These results can be understood as follows. With a slight abuse of notation, let $dX_i^* > 0$ denote the increase in country 1's abatement effort due to its small unilateral commitment

$d\lambda_i > 0$. (More formally, $dX_i^* = \left[(dX_i^*/d\lambda_i)_{\lambda_1=\lambda_2=0}\right] d\lambda_i > 0$ by Lemma 2.) Due to the free-riding effect which leads to carbon leakage by Lemma 1, country j adjusts its abatement effort by $dX_j^* = (-L_i)\, dX_i^* < 0$ in response.[12]

By the envelope theorem, the *direct* effect of a small change in country i's abatement effort on its *own* net benefit Π_i^* is zero, for $i = 1, 2$. The reason is that both countries were initially choosing their respective abatement efforts selfishly to maximize their own net benefit, so any (small) change their own abatement only has a second-order effect.

However, the unilateral commitment by country i also has two *strategic* effects, one positive and one negative. First, the increase in country i's abatement effort yields an increase in the abatement benefits enjoyed by the *other* country j of $B_j' dX_i^* > 0$. Second, the induced reduction in country j's abatement effort means that country i's abatement benefit changes by $B_i' dX_j^* = (-B_i' L_i)\, dX_i^* < 0$.

Now recall that country i's true objective $S_i = \Pi_i + \theta_i \Pi_j$, so it places weight $\theta_i \in [0, 1]$ on the first (positive) strategic effect and full weight on the second (negative) strategic effect. The weighted sum of these two strategic effects, $(\theta_i B_j' - B_i' L_i) dX_i^*$, thus determines the impact of a small unilateral commitment by country i on its own true objective function and behaves according to Proposition 1.

Intuitively, a unilateral commitment by country i increases the net benefit Π_j^* enjoyed by country j but acting unselfishly hurts its own net benefit Π_i^*. The commitment thus enhances its own true objective if (and only if) the former effect outweighs the latter. The positive effect will be large if country j's marginal benefit from abatement is large, and receives large weight according to country i's degree of unselfishness, $\theta_i \in [0, 1]$. The negative effect will be small if there is little free-riding, resulting in carbon leakage, by country j, and if country i's own marginal benefit from abatement is small. So a small commitment in direction of its true preference is beneficial for country i where θ_i is sufficiently high, B_i'/B_j' is sufficiently small, and L_i is also sufficiently small.

Proposition 1 already makes clear that whether a small commitment towards incorporating its unselfishness in decision-making is beneficial for a country depends crucially on the details of the environment. Perhaps surprisingly, under a fairly wide range of conditions, such a commitment backfires in that it actually *reduces* S_i^*.

□ **Impact on global welfare**. A closely related question is whether, and under which conditions, a small commitment by country i improves *global* welfare.

Proposition 2 *The impact of a small unilateral commitment $d\lambda_i$ by country i on equilibrium global welfare*

$$\left.\frac{dW^*}{d\lambda_i}\right|_{\lambda_1=\lambda_2=0} = \left[\left(B_j' - B_i' L_i\right) \frac{dX_i^*}{d\lambda_i}\right]_{\lambda_1=\lambda_2=0}$$

is (a) ambiguous in general, (b) negative for a ratio of marginal benefits B_i'/B_j' sufficiently large, and (c) positive if the ratio of marginal benefits satisfies $B_i' \leq B_j'$.

[12] Formally, $dX_j^* = \left[(dX_j^*/d\lambda_i)_{\lambda_1=\lambda_2=0}\right] d\lambda_i = \left[R_j'(X_i^*)\, (dX_i^*/d\lambda_i)\big|_{\lambda_1=\lambda_2=0}\right] d\lambda_i = (-L_i) dX_i^*$.

The logic underlying Proposition 2 follows that of Proposition 1. Once again, the direct effects on each country's net benefit are both zero by the envelope theorem. The only difference arises because, from a global-welfare perspective, the combined effect of the two strategic effects depends on their *unweighted* sum. So the increase in the abatement benefits enjoyed by the *other* country j of $B_j' dX_i^* > 0$ plus the induced reduction country i's abatement benefit of $B_i' \left(-L_i\right) dX_i^* < 0$ yield an overall welfare impact $dW^* = (B_j' - B_i' L_i) dX_i^*$. Previous arguments make clear that the sign of this expression, too, is ambiguous. A reoccurring theme in our analysis is that an additional commitment by a country is more likely to be welfare-enhancing if the national benefit it derives from additional abatement is *small* relative to the other country.[13]

4 Optimal commitments

We now analyze the general version of our model in which (i) both countries may have a true preference to be unselfish to some degree, as indexed by $\theta_i \in [0, 1]$, and (ii) each country chooses the extent to which it wishes to act unselfishly, as indexed by $\lambda_i \in [0, 1]$, so as to maximize its true objective $S_i = \Pi_i + \theta_i \Pi_j$.

□ **Generalized formula**. This analysis is more complicated because our previous argument, based on the envelope theorem, that the two direct effects of commitment are zero no longer applies. The reason is that countries are no longer necessarily acting entirely selfishly initially (that is, $\lambda_1 \neq 0$ and/or $\lambda_2 \neq 0$).

However, as before, a small increase $d\lambda_i > 0$ in country i's strategic preference (not necessarily starting from $\lambda_i = 0$), leads to an increase in its own abatement effort of $dX_i^* > 0$. (More formally, $dX_i^* = [(dX_i^*/d\lambda_i)]\, d\lambda_i > 0$ by Lemma 2.) By Lemma 1, country j adjusts its abatement effort by $dX_j^* = \left(-L_i\right) dX_i^* < 0$ in response.

The two strategic effects of an additional commitment are also as before. First, the increase in country i's abatement effort yields an increase in the abatement benefits enjoyed by the *other* country of $B_j' dX_i^* > 0$. Second, the induced reduction in country j's abatement effort means that country i's abatement benefit changes by $B_i' \left(-L_i\right) dX_i^* < 0$.

The direct effect of a small change dX_i^* in country i's abatement effort on its own net benefit Π_i, in general, is equal to $(B_i' - C_i') dX_i^*$. Recalling country i's first-order condition for its abatement level, $(B_i' - C_i') + \lambda_i \theta_i B_j' = 0$, the generalized direct effect equals $(-\lambda_i \theta_i B_j') dX_i^* \leq 0$, which is non-zero whenever country i initially was not entirely unselfish. Similarly, the direct effect of small (induced) change dX_j^* on country j's net benefit Π_j, in general, is equal to $(B_j' - C_j') dX_j^*$. By its first-order condition for abatement, $(B_j' - C_j') + \lambda_j \theta_j B_i' = 0$, the generalized

[13] An important difference between Propositions 1 and 2 lies in the case where the countries share similar benefits from global abatement activity. In particular, if the two countries have identical benefit *functions*, $B_1(X_1 + X_2) = B_2(X_1 + X_2)$, then the impact of a small commitment by an individual country on equilibrium global welfare W^* is always positive (Proposition 2, since $L_i < 1$). By contrast, the impact on its own true objective S_i^* may still be negative (Proposition 1, for $\theta_i \leq L_i$).

Similarly, if the two countries have identical marginal abatement costs in the initial equilibrium $C_1'(X_2^*) = C_1'(X_2^*)$, then a small unilateral commitments always improves global welfare W^* but has an ambiguous impact on its true objective S_i^*. (Note that $B_i' = C_i'$ in the initial equilibrium since both countries are entirely selfish.)

direct effect equals $(-\lambda_j \theta_j B_i')\, dX_j^*$, which we can also write as $(\lambda_j \theta_j B_i' L_i)\, dX_i^* \geq 0$ (since the induced change $dX_j^* = (-L_i)\, dX_i^* < 0$).

The overall equilibrium impact of a small additional commitment by country i on its true objective $S_i = \Pi_i + \theta_i \Pi_j$ takes into account all of these effects, with appropriate weightings:

$$\begin{aligned} dS_i^* \quad = \quad & \underbrace{(-\lambda_i \theta_i B_j') dX_i^*}_{\substack{\text{direct effect} \\ \text{on country } i\ (\ \leq 0)}} \quad + \quad \underbrace{\left(-B_i' L_i\right) dX_i^*}_{\substack{\text{strategic effect} \\ \text{on country } i\ (\ < 0)}} \\ + \quad & \underbrace{\theta_i}_{\substack{\text{true unselfishness} \\ \text{of country } i\ (\ \in [0,1])}} \times \Big[\ \underbrace{\left(\lambda_j \theta_j B_i' L_i\right) dX_i^*}_{\substack{\text{direct effect} \\ \text{on country } j\ (\ \geq 0)}} \quad + \quad \underbrace{\left(B_j'\right) dX_i^*}_{\substack{\text{strategic effect} \\ \text{on country } j\ (\ > 0)}} \quad \Big]. \end{aligned}$$

Writing this expression more compactly gives us the following result.[14,15]

Lemma 4 *The generalized impact of a small unilateral commitment $d\lambda_i$ by country i on its equilibrium true preference*

$$\frac{dS_i^*}{d\lambda_i} = \left[(1-\lambda_i)\theta_i B_j' - (1-\lambda_j \theta_i \theta_j) B_i' L_i\right] \frac{dX_i^*}{d\lambda_i}.$$

□ **Full commitment.** We begin by exploring the implications of this result for one of the limiting cases: When is a full commitment $\lambda_i = 1$ optimal?

Proposition 3 *(a) If the countries' true preferences are entirely unselfish $\theta_1 = \theta_2 = 1$, then their optimal commitments $\lambda_1^* = \lambda_2^* = 1$ achieve first-best abatement levels;*
(b) If at least one country has partially selfish true preferences $\theta_1 < 1$ or $\theta_2 < 1$, then countries' optimal commitments $\lambda_1^* < 1$ and $\lambda_2^* < 1$ and both abatement levels fall short of first-best.

Part (a) of the result shows that the first-best solution is sustainable in our model as long as *both* countries want to be entirely unselfish. The intuition is that if both countries care about global welfare then neither as has incentive to unilaterally deviate from full commitments $\lambda_1^* = \lambda_2^* = 1$, since any such deviation, by construction, must cause global welfare to fall below its first-best level. So, given the optimal strategic objective chosen at Date 1, each country chooses its abatement effort at Date 2 to $\max_{X_i} W$.

Part (b), however, shows that this optimistic conclusion applies *only* if both countries are entirely unselfish. Whenever at least one country places greater weight on national welfare in its true objective function, the optimal commitments of *both* countries fall short of a full

[14] This decomposition shows that, in general, the equilibrium impact of more unselfish action by a country on its own net benefit is strictly negative ($d\Pi_i^* < 0$), while its impact on the other country's net benefit is strictly positive ($d\Pi_j^* < 0$).

[15] Note that the formulae in Proposition 1 ($\lambda_1 = \lambda_2 = 0$) and Proposition 2 ($\lambda_1 = \lambda_2 = 0$ and $\theta_i = 1$) can be obtained as special cases of Lemma 4.

commitment, $\lambda_1^* < 1$ and $\lambda_2^* < 1$. In such cases, given the optimal strategic objective chosen at Date 1, country i chooses its abatement effort at Date 2 to $\max_{X_i} \Omega_i = (1 - \lambda_i^*)\Pi_i + \lambda_i^* S_i = \Pi_i + \lambda_i^* \theta_i \Pi_j$ (with $\lambda_i^* \theta_i < 1$).

This result can be understood by thinking about the impact of the "last step" towards a full commitment with $\lambda_i = 1$. Observe that, in this case, the negative direct effect on country i is sufficiently negative to entirely offset the *weighted* positive strategic effect on country j. Therefore, the impact of the last step is determined by the two remaining effects, the strategic effect on country i plus the weighted direct effect on country j. This equals $[-(1 - \lambda_j \theta_i \theta_j) B_i' L_i] \, dX_i^* < 0$, and is negative since $\theta_i < 1$ or $\theta_j < 1$ by assumption (and also $\lambda_j \leq 1$). Therefore, the last step actually reduces the equilibrium value of country i's true objective S_i^*. Since the same reasoning applies symmetrically to the other country, it follows that, in equilibrium, $\lambda_1^* < 1$ and $\lambda_2^* < 1$. In terms of strategic delegation, it is optimal for a countries' citizens to delegate decision-making on abatement efforts to politicians whose preferences are closer to national self-interest.

Perhaps the most striking statement of this latter result is obtained by considering a situation in which country 1 is entirely unselfish, so $\theta_1 = 1$, while country 2 is unselfish only to some degree, so $\theta_2 < 1$. Then part (b) says that the *optimal* commitment by country 1 satisfies $\lambda_1^* < 1$, so a full commitment is dominated by a weaker policy. So the optimal way for country 1 to maximize global welfare W is to maximize a strategic objective $\Omega_1 = (1 - \lambda_1^*)\Pi_1 + \lambda_1^* W$ that is partially skewed towards its own national welfare. In other words, *a country that genuinely wants to maximize global welfare actually does best by being at least somewhat selfish.* This result is a manifestation of the general insight due to Schelling (1960) that, in strategic settings, the maximum A may be achieved by maximizing B.

Intuitively, why can country 1 do better than playing according to its true, entirely unselfish preference? The key is that a small decrease in its own level of abatement only leads to a second-order loss in global welfare. However, the resulting induced *increase* in the other country's abatement level leads to a first-order gain in global welfare (whenever the other country is not already choosing the first-best abatement). The reason why a full commitment is almost never optimal is essentially "*reverse leakage*"—a weaker commitment reduces free-riding by the other country.

Therefore, in a world in which not all countries have entirely selfish preferences towards climate change, it is not optimal for *any* country to act according to its true degree of unselfishness. It is worth emphasizing that this insight is quite general as it does not depend on any particular assumptions on the functional forms of countries' benefit and cost functions (beyond positive and decreasing marginal abatement benefits, so leakage rates are strictly positive).

□ **Zero commitment.** We now turn to the opposite limiting case. Our next result gives two characterizations for when the optimal commitment by a single country or by both countries is a *zero* commitment.

Proposition 4 *(a) If the ratio of marginal benefits B_i'/B_j' is sufficiently large (and $\theta_i < 1$ or $\theta_j < 1$), then country i's optimal commitment $\lambda_i^* = 0$;*

(b) If the countries' true preferences θ_1 and θ_2 are positive but sufficiently small, then their optimal commitments, $\lambda_1^* = \lambda_2^* = 0$.

Part (a) of the result essentially gives a non-local version of our earlier finding that a small commitment by an individual country may not raise S_i^* or indeed W^*. Note that, in extreme cases, it may even be *optimal* for an entirely unselfish country (that is, when $\theta_i = 1$ but $\theta_j < 1$) to choose its abatement level in its own strict national interest ($\lambda_i^* = 0$).

An implication is that a policy of zero commitment may welfare-dominate one of full commitment. Suppose that country 1 is has a completely altruistic true preference while country 2 is entirely self-interested, $(\theta_1, \theta_2) = (1, 0)$. By Lemma 3, we know that $\lambda_2^* = 0$ irrespective of country 1's climate policy. But also, if B_1'/B_2' is sufficiently large, then equilibrium global welfare W^* is higher with zero commitments $(\lambda_1, \lambda_2) = (0, 0)$ than with $(\lambda_1, \lambda_2) = (\ell, 0)$ for *any* $\ell \leq 1$ (since then $dW^*/d\lambda_i \leq 0$ for all $\lambda_i \in [0, \ell]$). In this example, a global-welfare oriented country does better by acting selfishly than by pursuing its true global welfare objective.

The reason for part (b) is that, if a country has only a small degree of unselfish preferences, then it places relatively little weight on the positive direct and strategic effects that accrue to the other country's net benefits, so that these effects have too little weight in the calculus to be able to overcome the negative impact any commitment has on the country's own national net benefit. Applying this logic to both countries, optimal commitments are both zero.

Our analysis of the limiting cases thus shows that the first-best outcome is generally unattainable, except in a single "knife-edge" case. Furthermore, countries' *optimal* commitments may, in a fairly wide range of cases, be low or even zero—despite importantly unselfish underlying true preferences.

□ **Interior commitments.** Given our analysis of the limiting cases, we can now provide a characterization of countries' optimal commitments in a subgame-perfect Nash equilibrium where the solution is interior.

Proposition 5 *In an interior equilibrium with $(\lambda_1^*, \lambda_2^*) \in (0,1)^2$, country i's optimal commitment λ_i^* satisfies*

$$\lambda_i^* = \frac{\left[\theta_i(1 - L_i L_j) - (1 - \theta_i\theta_j)\,\dfrac{B_i'(X_1^* + X_2^*)}{B_j'(X_1^* + X_2^*)} L_i\right]}{\theta_i\,(1 - \theta_i\theta_j L_i L_j)} \in (0,1),$$

where the equilibrium rates of carbon leakage

$$L_i = \frac{\left[1 + \lambda_j^*\theta_j \dfrac{B_i''(X_1^* + X_2^*)}{B_j''(X_1^* + X_2^*)}\right]}{\left[1 + \dfrac{C_j''(X_j^*)}{\left|B_j''(X_1^* + X_2^*)\right|} + \lambda_j^*\theta_j \dfrac{B_i''(X_1^* + X_2^*)}{B_j''(X_1^* + X_2^*)}\right]} \in (0,1),$$

and country i's abatement level X_i^* satisfies

$$X_i^* = C_i'^{-1}\left[B_i'(X_1^* + X_2^*) + \lambda_i^* \theta_i B_j'(X_1^* + X_2^*)\right] > 0.$$

Proposition 6 implicitly describes the two countries' optimal interior commitments $(\lambda_1^*, \lambda_2^*)$, given their respective abatement benefit and cost functions as well as their true preferences (θ_1, θ_2) towards unselfishness.[16] The solution involves six equations and six unknowns: Two equations for optimal commitments $\lambda_1^*(L_1, L_2, X_1^*, X_2^*; \theta_1, \theta_2)$ and $\lambda_2^*(L_1, L_2, X_1^*, X_2^*; \theta_1, \theta_2)$ as functions of leakage and abatement, two equations for leakage rates $L_1(\lambda_1^*, \lambda_2^*, X_1^*, X_2^*; \theta_1, \theta_2)$ and $L_2(\lambda_1^*, \lambda_2^*, X_1^*, X_2^*; \theta_1, \theta_2)$ as functions of commitments and abatement, and two equations for countries' equilibrium abatement levels $X_1^*(\lambda_1^*, \lambda_2^*; \theta_1, \theta_2)$ and $X_2^*(\lambda_1^*, \lambda_2^*; \theta_1, \theta_2)$ as functions of their optimal commitments.

In principle, a numerical solution for the six unknowns can be obtained by making specific assumptions on the functional forms of the abatement benefit and cost functions, $B_i(\cdot)$ and $C_i(\cdot)$. It turns out that obtaining these numerical solutions is rather cumbersome and messy; they involve higher-order polynomials even in the simplest cases, and appear to offer little additional economic insight. We therefore now instead discuss the underlying informational requirements in more detail, and, relatedly, how optimal commitments might be estimated in practice.

In terms of model primitives, the informational requirement to determine optimal commitments is as follows. First, the ratio of countries' marginal benefits, B_i'/B_j', and the ratio of countries' slopes of marginal benefits, B_i''/B_j'', both evaluated at equilibrium. Second, each country's ratio of the slope of marginal cost to the slope of marginal benefits, $C_i''/\left|B_i''\right|$, again evaluated at equilibrium. Finally, each country's underlying true preference for unselfishness θ_i, which our model has been taking as exogenously given.

Instead of relying on full-fledged estimates of countries' cost and benefit functions, it is possible to obtain significant simplification under some standard assumptions. First, suppose that country i's benefit function $B_i(X_1 + X_2) = \mu_i B(X_1 + X_2)$, where $\mu_i > 0$ represents the weight it places on a *global* benefit function $B(\cdot)$.[17] This formulation has the advantage that the ratio terms $B_i'/B_j' = B_i''/B_j'' = \mu_i/\mu_j$ become constants, and are thus invariant to the details of countries' abatement efforts. Second, consider a setting where country i's marginal costs and benefits are both affine functions of abatement, $B_i'(X_1 + X_2) = [\alpha_i - \beta_i(X_1 + X_2)]$ and $C_i'(X_i) = \delta_i X_i$. In this case, the ratio of the slope of marginal cost to slope of marginal benefits, $C_i''/\left|B_i''\right| = \delta_i/\beta_i$ is also a constant. This latter assumption is essentially equivalent to the classic analysis of Weitzman (1974) on whether price- or quantity-based regulation is socially preferable. It can be seen as a second-order approximation to the unknown shapes of the underlying cost and benefit functions (see also Barrett, 1994).

Taken together, these two additional assumptions simplify considerably the analytics of leakage rates and optimal commitments. In particular, these can now be calculated using Proposition 5 without reference to the associated equilibrium levels of abatement. In other words, optimal

[16] The more cumbersome notation emphasizes that terms involving marginal abatement benefits and costs may depend on equilibrium abatement levels, which in turn depend on optimal commitments.

[17] So the proportion of global abatement benefits that accrues to country i simply equals $\mu_i/(\mu_1 + \mu_2)$.

commitments can be determined as the solution to a system of four equations and four unknowns $(\lambda_1^*, \lambda_2^*, L_1, L_2) \in (0,1)^4$, for given underlying true preferences (θ_1, θ_2).

□ **Inferring unselfishness.** Our results also imply that *caution is required in inferring whether or not a country is "being selfish" from its observed behaviour.* Recall that a country chooses its abatement level at Date 2 to maximize its strategic objective, $\max_{X_i} \Omega_i = \Pi_i + \lambda_i^* \theta_i \Pi_j$. Suppose it is observed or otherwise estimated that a country's abatement efforts appear to be entirely selfish, $\lambda_i^* \theta_i = 0$. It does *not* follow that this country's underlying true preference is to be completely selfish. Little or no additional abatement can be consistent even with fully altruistic true preferences—simply because it may arise from $\lambda_i^* = 0$ rather than $\theta_i = 0$. So the question "How unselfish is country A?" is not that easy to answer based on its observed abatement behaviour.

Our model also reveals an unusual asymmetry:

Proposition 6 *For countries' true preferences $0 < \theta_j < \theta_i$ (where $\theta_j < 1$), optimal commitments may satisfy $\lambda_j^* \theta_j > \lambda_i^* \theta_i$.*

Surprisingly, therefore, there may be a non-obvious relationship between countries' true preferences for unselfishness (θ_is) and their strategic actions ($\lambda_i^* \theta_i$s). A country with a higher true preference for unselfishness (higher θ_i) may, in equilibrium, be the country whose abatement actions are closer to the standard self-interested solution (lower $\lambda_i^* \theta_i$) because it places less weight on global welfare. In particular, a country with fully altruistic preferences ($\theta_i = 1$) may end up acting more selfishly than others. So the question "Is country A more unselfish than country B?" also has no straightforward answer based on countries' observed behaviour.

5 Further results and extensions

Our analysis shows that it is almost always optimal for countries to pursue emissions abatement according to a strategic objective that falls short of their true preferences for unselfish action. A natural question, therefore, is how countries might be able to alleviate the underlying free-riding problems. We here discuss a further result on carbon leakage effects as well as three related extensions of our basic modelling approach.

A. Carbon leakage effects

We can derive a further analytical insight regarding the impact of altruistic behaviour on carbon leakage in settings with particular classes of benefit and cost functions:

Proposition 7 *(a) Suppose that $B_j''' \leq 0$ and $C_j''' \leq 0$. Then the rate of carbon leakage L_i associated with country i's commitment, $d\lambda_i^* > 0$, is higher when country j also has an unselfish commitment, $\lambda_j^* > 0$, than when country j acts entirely selfishly, $\lambda_j^* = 0$.*
(b) Suppose that $B_j''' = 0$, $C_j''' = 0$ and B_i''/B_j'' is constant. Then the rate of carbon leakage L_i associated with country i's commitment, $d\lambda_i^* > 0$, increases in country j's commitment, $\lambda_j^* > 0$.

Proposition 7 gives simple *sufficient* conditions under which more unselfish policies are associated with higher rates of leakage. For part (a), the condition $B_j''' \leq 0$ is satisfied, for example, by marginal benefits of the form $B_j'(X_1+X_2) = \left[\alpha_j - \beta_j\,(X_1+X_2)^{\gamma_j}\right]$ with $\gamma_j \in (0,1]$, while the condition $C_j''' \leq 0$ is met by any cost function of the form $C_j(X_j) = (\delta_j/\sigma_j)X_i^{\sigma_j}$ for $\sigma_j \in (0,2]$. For part (b), the conditions that $B_j''' = 0$, $C_j''' = 0$ and B_i''/B_j'' is constant essentially come down to the affine formulation $B_j'(X_1+X_2) = \left[\alpha_j - \beta_j\,(X_1+X_2)\right]$ and $C_j'(X_j) = \delta_j X_j$.

To understand the result, recall country j's strategic objective $\Omega_j = \Pi_j + \lambda_j\theta_j\Pi_i$ which leads to the first-order condition for abatement, $\partial\Omega_j/\partial X_j = (\partial\Pi_j/\partial X_j) + \lambda_j\theta_j\,(\partial\Pi_i/\partial X_j) = 0$. The overall rate of leakage can hence be thought of in two parts: firstly, a part for the selfish component $\partial\Pi_j/\partial X_j$, and, secondly, a part for the unselfish component $\partial\Pi_i/\partial X_j$ (which, in equilibrium, receives weight $\lambda_j^*\theta_j$). The key point is that the part connected to the unselfish component has a leakage rate of 100%.[18] Greater weight on the unselfish part therefore certainly tends to increase the overall leakage rate as long as the selfish part does not decline as a result. The conditions given in Proposition 7 are, respectively, (a) sufficient for the selfish part to not decline, and (b) necessary and sufficient for it to stay constant.

Proposition 7 is somewhat unexpected as it seems natural to think that less selfish behaviour by countries will tend to mitigate carbon leakage. This intuition turns out to be potentially misleading. Although global welfare may (but need not) be higher when countries are acting unselfishly, the associated leakage rates can be higher than with self-interested behaviour. Put differently, unselfish behaviour here actually *worsens* the free-riding problem at the margin.

The more general point is that rates of carbon leakage—though useful and important—are not always reliable welfare indicators.

B. Sequential commitments

Our model has focused on the subgame-perfect Nash equilibrium in which both countries simultaneously choose their strategic preference towards climate change at Date 1, and then simultaneously choose their levels of emissions abatement at Date 2. What is the impact of a *sequential* move order on the equilibrium outcome? In particular, what if one country acts as a first-mover in choosing its strategic preference λ_i^*?

Lemma 5 *Suppose that B_i'/B_j' is constant. In an interior equilibrium with $(\lambda_1^*, \lambda_2^*) \in (0,1)^2$, country i's optimal commitment varies with country j's commitment according to*

$$\operatorname{sign}\left(\frac{d\lambda_i^*}{d\lambda_j}\right) = \operatorname{sign}\left(\theta_i\theta_j(1+\eta_i) - \frac{\eta_i}{\lambda_j}\right)$$

where $\eta_i \equiv (d\log L_i/d\log\lambda_j)$ is the elasticity of country i's leakage rate with respect to country j's commitment.

Lemma 5 shows that it is ambiguous, in general, whether countries' commitments are strategic complements ($d\lambda_i^*/d\lambda_j > 0$) or strategic substitutes ($d\lambda_i^*/d\lambda_j < 0$). Proposition 7 and

[18] To see why, observe that holding $\partial\Pi_i/\partial X_j = B_i'(X_i+X_j)$ fixed in response to a small increase in country i's abatement effort $dX_i > 0$ requires a decrease in country j's abatement $dX_j = -dX_i < 0$ that is exactly offsetting.

its discussion suggest that, in a fairly broad range of cases, we can expect that $dL_i/d\lambda_j > 0$, so the leakage-commitment elasticity is positive, $\eta_i > 0$. This effect tends to push towards strategic substitutability; indeed commitments are always strategic substitutes if $\eta_i > 0$ and country j's commitment λ_j is sufficiently small. By contrast, strategic complementarity can obtain if countries' "joint unselfishness", as measured by $\lambda_j\theta_i\theta_j$, is sufficiently pronounced and the leakage-commitment elasticity η_i is sufficiently small. Loosely speaking, this suggests that commitment are strategic complements when they are already quite strong, and strategic substitutes when they are relatively weak.

The case with strategic complements is particularly interesting from a free-riding perspective. It is the reverse of the basic feature of our model that countries' abatement efforts are strategic substitutes—which leads to carbon leakage (Lemma 1). If commitments are complements, then a first-moving country that increases its commitment induces the follower to do the same, so abatement efforts in both countries increase and global emissions damages are reduced.

Lemma 5, however, does not provide a full analysis of optimal sequential commitments. Firstly, it only deals with interior equilibria for countries' commitments. It is entirely possible that one country's (or both countries') optimal commitment in our benchmark model is zero (Proposition 4). In such cases, a change from simultaneous to sequential moves may make no difference if a country remains "stuck in a corner" at $\lambda_i^* = 0$. Secondly, while global abatement will rise if commitments are strategic complements, our preceding analysis highlights that this need not yield a welfare improvement. For example, the additional abatement may "overshoot" what is welfare-optimal from a national or global viewpoint. Thirdly, we have unfortunately not been able to characterize the conditions under which a country would in fact want to become a first-mover in the first place.

Nonetheless, based on these arguments, we conjecture that it is ambiguous (i) if sequential commitment makes a difference, and (ii) if any such difference raises or reduces welfare.

C. Global cost-effectiveness

The benchmark model does not make explicit which instruments are used by countries to achieve their desired abatement levels. While we can interpret each individual country's cost of abatement function as satisfying *domestic* cost-effectiveness, there is no mechanism in the model that delivers cost-effectiveness at an international level. We believe this is a reasonable way to model the kinds of unilateral commitments discussed in the introduction—given the absence of a global carbon price, and the resulting divergence of marginal abatement costs across countries.

It is nonetheless interesting to explore the role of cost-effectiveness in some more detail. Observe that the expression from Lemma 4 can be rewritten as follows:

$$\begin{aligned}
\left.\frac{dS_i^*}{d\lambda_i}\right|_{\lambda_i=1} &= -\left[(1-\lambda_j\theta_i\theta_j)B_i'L_i\right]\frac{dX_i^*}{d\lambda_i} \\
&= -[B_i' + \theta_i(B_j' - C_j')]L_i\frac{dX_i^*}{d\lambda_i} \text{ (since } (B_j' - C_j') + \lambda_j\theta_j B_i' = 0) \\
&= -(C_i' - \theta_i C_j')L_i\frac{dX_i^*}{d\lambda_i} \text{ (since } (B_i' - C_i') + \theta_i B_j' = 0 \text{ with } \lambda_i = 1),
\end{aligned}$$

from which we obtain the following result.

Lemma 6 *If country i's true preference $\theta_i = 1$ and global cost-effectiveness $C_i' = C_j'$ obtains, then country i's optimal commitment $\lambda_i^* = 1$.*

If a country has entirely altruistic preferences *and* global cost-effectiveness obtains, then a full commitment is optimal and it performs the first-best level of abatement (X_i^* satisfies $(B_1' + B_2') = C_i'$). This outcome is possible in our benchmark model only if the other country also has entirely altruistic preferences (so $\theta_1 = \theta_2 = 1$). Then, by part (a) of Proposition 3, we have that $\lambda_1^* = \lambda_2^* = 1$ and first-best is achieved (X_i^* satisfies $(B_1' + B_2') = C_i'$ for $i = 1, 2$). The key point is that, in the benchmark model, global cost-effectiveness is guaranteed only if the two countries are perfectly symmetric—that is, have identical benefit and cost functions, as well as the same degree of unselfishness.

A direct—yet admittedly also rather ad-hoc—way of modelling the impact of, say, a global cap-and-trade scheme, is to super-impose cost-effectiveness on the equilibrium conditions of the benchmark model by setting $C_i' = C_j'$. Lemma 6 then shows that a country optimally engages in a full commitment irrespective of the other country's degree of unselfishness. This suggests that *global cost-effectiveness can act as a substitute for other countries' unselfishness* in terms of enabling strong unilateral commitment. This in turn provides a potential justification for a global cap-and-trade scheme even if a subset of countries in the scheme does not engage in any significant abatement. The purpose of the scheme is to encourage cost-effectiveness (as usual) but also to enable unselfish countries to follow through on their preferences.

As noted, we have derived this insight using an ad-hoc formalization of cost-effectiveness. A full model would include an explicit treatment of the underlying instrument, such as an emissions trading scheme or emissions tax, that delivers cost-effectiveness across countries. Such a model would likely also raise important distributional issues, for example, concerning the allocation of emissions permits across countries and the usage of revenue generated by the tax. Moreover, a country's degree of unselfishness regarding the benefits and costs of emissions abatement may well differ from its preferences regarding international rent distribution more generally.

D. Other objectives

In our benchmark model, a country's true objective function S_i can represent a continuum of preferences, ranging from pure self-interest ($\theta_i = 0$) to a concern for global welfare ($\theta_i = 1$). This seems a very natural modelling choice to us, but there are, of course, other possible objective functions to incorporate unselfishness and other behavioural biases. We here discuss how our main insights can carry over to other objectives.

Suppose that the citizens of country i's have a true objective function is given by $S_i = (1 - \widehat{\theta}_i)\Pi_i + \widehat{\theta}_i\Psi_i$, where Ψ_i is an arbitrary objective that represents preferences that depart from national self-interest. It will be convenient to also define $\Phi_i \equiv (\Psi_i - \Pi_i)$, so that we can write $S_i = \Pi_i + \widehat{\theta}_i\Phi_i$.[19] As in the benchmark model, decision-making on abatement policy is

[19] In the benchmark model, $\Psi_i = W$ (for $i = 1, 2$) and so $\Phi_i = \Pi_j$.

delegated to politicians by way of a strategic objective $\Omega_i = (1-\widehat{\lambda}_i)\Pi_i + \widehat{\lambda}_i S_i$, where the strategic preference $\widehat{\lambda}_i \in [0,1]$ plays an analogous role to λ_i above.

This generalized model nests as special cases two other potentially relevant kinds of altruism. First, it has been argued that providing a contribution to a public good yields a "warm glow" (Andreoni, 1990). Translated into our setting, warm-glow models are essentially equivalent to an objective $\Phi_i = X_i$, such that a country derives direct benefits from its amount of emissions abatement. Second, it has been suggested to us that some countries may effectively be willing to "ignore" some of the abatement costs they incur in unilateral action. This can be cast into an objective of the form $\Phi_i = C_i(X_i)$, such that a country acts as though its abatement costs are lower than they actually are. Such other objectives, however, do have the disadvantage that they do not nest the first-best outcome in the simple way our benchmark model does.

Assume that the properties of Ψ_i (respectively, Φ_i) are such that the first-order conditions

$$\frac{\partial \Omega_i}{\partial X_i} = \frac{\partial \Pi_i}{\partial X_i} + \widehat{\lambda}_i \widehat{\theta}_i \frac{\partial \Phi_i}{\partial X_i} = 0$$

define an interior equilibrium for abatement efforts, and that the associated second-order conditions $\partial^2 \Omega_i / \partial X_i^2 < 0$ are satisfied. Further, assume a stable equilibrium which features positive leakage rates $L_i \in (0,1)$ (corresponding to Lemma 1). Using the same arguments as in the proof of Lemma 2, the impact of a stronger commitment on equilibrium abatement levels obeys $\text{sign}\Big(dX_i^*/d\widehat{\lambda}_i\Big) = \text{sign}\Big(\widehat{\theta}_i \times \partial\Phi_i/\partial X_i\Big)$. So as long as $\partial \Phi_i/\partial X_i > 0$, that is, Ψ_i has higher marginal returns than the country's net benefit Π_i, we have that $dX_i^*/d\widehat{\lambda}_i > 0$, akin to Lemma 2. Moreover, if the country's true preference is pure selfishness, $\widehat{\theta}_i = 0$, then its optimal commitment $\widehat{\lambda}_i^* = 0$ as in Lemma 3.

To examine the welfare impact of a stronger policy commitment, observe that

$$\begin{aligned}
\frac{dS_i^*}{d\widehat{\lambda}_i} &= \left(\frac{dS_i^*}{dX_i} - \frac{dS_i^*}{dX_j} L_i\right) \frac{dX_i^*}{d\widehat{\lambda}_i} \\
&= \left(\frac{\partial \Pi_i^*}{\partial X_i} + \widehat{\theta}_i \frac{\partial \Phi_i^*}{\partial X_i} - \left[\frac{\partial \Pi_i^*}{\partial X_j} + \widehat{\theta}_i \frac{\partial \Phi_i^*}{\partial X_j}\right] L_i\right) \frac{dX_i^*}{d\widehat{\lambda}_i} \text{ (since } S_i = \Pi_i + \widehat{\theta}_i \Phi_i) \\
&= \left((1-\widehat{\lambda}_i)\widehat{\theta}_i \frac{\partial \Phi_i^*}{\partial X_i} - \left[\frac{\partial \Pi_i^*}{\partial X_j} + \widehat{\theta}_i \frac{\partial \Phi_i^*}{\partial X_j}\right] L_i\right) \frac{dX_i^*}{d\widehat{\lambda}_i} \text{ (since } \partial\Omega_i/\partial X_i = 0),
\end{aligned}$$

which is the analog to Lemma 4 above. The general point again is that the this expression depends on the details of the environment—its sign is ambiguous.

For the case of small commitment (where both countries are initially entirely selfish, $\lambda_1 = \lambda_2 = 0$), we have that $\partial \Pi_i/\partial X_i = 0$ and so this expression simplifies to

$$\left.\frac{dS_i^*}{d\widehat{\lambda}_i}\right|_{\widehat{\lambda}_1=\widehat{\lambda}_2=0} = \left[\left(-B_i' L_i + \widehat{\theta}_i \left[\frac{\partial \Phi_i^*}{\partial X_i} - \frac{\partial \Phi_i^*}{\partial X_j} L_i\right]\right) \frac{dX_i^*}{d\widehat{\lambda}_i}\right]_{\widehat{\lambda}_1=\widehat{\lambda}_2=0},$$

using that $\partial \Pi_i/\partial X_j = B_i' > 0$. From this, it is immediate that a small commitment may lead to a reduction in country i's true objective, $(dS_i^*/d\widehat{\lambda}_i)_{\widehat{\lambda}_1=\widehat{\lambda}_2=0} < 0$. For example, this is always the case for $\widehat{\theta}_i$ positive but sufficiently small—due to carbon leakage, just as in Proposition 1 in the benchmark model. The fact that a zero commitment may be optimal locally immediately opens up the possibility that this is also true globally (Proposition 4).

To show that a full commitment is again almost always sub-optimal, observe that

$$\begin{aligned}
\left.\frac{dS_i^*}{d\widehat{\lambda}_i}\right|_{\widehat{\lambda}_i=1} &= -\left[\left(\left[B_i' + \widehat{\theta}_i \frac{\partial \Phi_i^*}{\partial X_j}\right] L_i\right) \frac{dX_i^*}{d\widehat{\lambda}_i}\right]_{\widehat{\lambda}_i=1} \quad (\text{since } \partial \Pi_i / \partial X_j = B_i') \\
&= -\left[\left(\left[(1-\widehat{\theta}_i)B_i' + \widehat{\theta}_i \frac{\partial \Psi_i^*}{\partial X_j}\right] L_i\right) \frac{dX_i^*}{d\widehat{\lambda}_i}\right]_{\widehat{\lambda}_i=1} \quad (\text{since } \Phi_i \equiv (\Psi_i - \Pi_i)).
\end{aligned}$$

Note that $(dS_i^*/d\widehat{\lambda}_i)_{\widehat{\lambda}_i=1} < 0$ is certainly satisfied whenever $\widehat{\theta}_i < 1$ and $\partial \Psi_i^* / \partial X_j \geq 0$, in which case we can conclude that the optimal commitment $\widehat{\lambda}_i^* < 1$ (see Proposition 3 above).

We therefore believe that the main insights from our analysis carry over to a range of other plausible objectives reflecting altruism, including "warm glow" and "ignoring costs".

6 Conclusion

We have studied the impact of unselfish preferences in a model of climate policy characterized by free-riding and carbon leakage. In short, our analysis shows that a country's *optimal* unselfishness: (i) is almost always less than its *willingness* to pursue unselfish action; (ii) is *much* less than its willingness in a surprisingly broad range of cases; and (iii) may be less than that of a country with a *lower* willingness to be unselfish. Although we have, for simplicity, developed these results in the context of a two-country model, we believe that they also apply in a general setting with $N > 2$ countries.

Our results can be related to some of the policy initiatives discussed in the introduction. By incorporating countries' social preferences we can, in principle, explain any observed outcome between the standard self-interested equilibrium and first-best. So the unilateral actions observed at the local, national, and regional levels *might* indeed be driven by altruistic preferences. At the same time, however, our analysis shows that in a world in which not all countries have entirely selfish preferences towards climate change, it is not optimal for *any* country to act according to its true degree of unselfishness. Amongst other things, this suggests that it may not be good idea for an individual country to unilaterally commit to taking the full "social cost of carbon" into account in national decisions on abatement activity.

In our model, unselfish preferences are necessary—but not sufficient—for countries to deviate from their self-interested levels of abatement. Small degrees of true unselfishness are, in equilibrium, negated by carbon leakage, so optimal commitments are zero. In some cases, it may be optimal even for a country with completely altruistic preferences to act purely in its own national self-interest. It is therefore *not* possible to simply infer that countries that have to date not engaged in unilateral action are, in fact, completely selfish.

We should also mention a number of caveats to our analysis:

First, in our model, a country has perfect information both on its own and the other country's benefit and cost functions. This has, amongst other things, guaranteed the commitment value of strategic delegation to politicians with different preferences regarding abatement policies. Though such assumptions are common in the literature on international environmental agreements (Barrett, 1994), it would be interesting to have them relaxed.

Second, other mechanisms absent from our benchmark model may be able to help sustain more favourable outcomes. For example, we have examined a static setting in which countries take decisions on commitments and abatement efforts at a single point in time. It is well-known that environmental cooperation may be sustainable in repeated games (Barrett, 2003), or also if arbitrary side payments between countries are feasible.

Third, and relatedly, we have ignored the intertemporal features of climate-change policy, including the problem of discounting. Many abatement costs are incurred in the near term while abatement benefits may only accrue much later on. Incorporating such intertemporal features would make the model more complex but also more closely aligned with other policy issues.

Fourth, it is possible that a country's unselfish commitment has other beneficial "knock-on" effects. For example, a stronger commitment today may induce additional low-carbon innovation that shifts tomorrow's abatement cost function downwards. Also, in other models, joining the "club" of countries that are committed to acting unselfishly may induce other countries to do the same (Heal and Kunreuther, 2010).

Such issues might usefully be studied in future research. Also, while our discussion has focused on climate policy, it seems likely that similar insights could also be applied to other problems of public good provision such as foreign aid, defense spending, and disaster relief.

7 Appendix: Proofs

□ **Proof of Lemma 2**. Observe that $dX_i^*/d\lambda_i = \partial X_i^*/\partial\lambda_i + R_i'(X_j^*)[dX_j^*/d\lambda_i]$ and $dX_j^*/d\lambda_i = R_j'(X_i^*)[dX_i^*/d\lambda_i]$, so that

$$\frac{dX_i^*}{d\lambda_i} = \frac{\partial X_i^*/\partial\lambda_i}{\left[1 - R_i'(X_j^*)R_j'(X_i^*)\right]}.$$

The denominator of this expression is positive by Lemma 1. Differentiating country i's first-order condition yields that the numerator

$$\frac{\partial X_i^*}{\partial\lambda_i} = \frac{\theta_i B_j'}{\left(-B_i'' + C_i'' - \lambda_i\theta_i B_j''\right)},$$

from which the result is immediate.

□ **Proof of Lemma 3**. Let $X_i^*(0)$ denote the Nash equilibrium level of abatement that solves the first-order condition $\partial\Pi_i/\partial X_i = 0$ at Date 2. Committing to deviate from this abatement level at Date 1 affects the country's equilibrium payoff according to

$$\frac{d\Pi_i^*}{dX_i} = \frac{\partial\Pi_i^*}{\partial X_i} + \frac{\partial\Pi_i^*}{\partial X_j}R_j'.$$

The first term, $\partial\Pi_i^*/\partial X_i$, is non-positive: by definition, it equals zero at $X_i^*(0)$ and it is negative by the concavity of the payoff function Π_i for any $X_i > X_i^*(0)$. The second term, $(\partial\Pi_i^*/\partial X_j)\, R_j'$, is negative since $\partial\Pi_i^*/\partial X_j = B_i' > 0$ and $R_j' < 0$ by Lemma 1. This shows that the Nash equilibrium level of abatement $X_i^*(0)$ is optimal, which is seen to be equivalent to a zero commitment being optimal, $\lambda_i^* = 0$.

□ **Proof of Proposition 3**. For part (a), setting $\theta_1 = \theta_2 = 1$ in the formula from Lemma 4 shows that

$$\left.\frac{dS_i^*}{d\lambda_i}\right|_{\theta_1=\theta_2=1} = \frac{dW^*}{d\lambda_i} = \left[\left((1-\lambda_i)B_j' - (1-\lambda_j)B_i'L_i\right)\frac{dX_i^*}{d\lambda_i}\right]_{\theta_1=\theta_2=1}.$$

So if country j is playing $\lambda_j = 1$, then $(dW^*/d\lambda_i)|_{\lambda_j=1} = [(1-\lambda_i)B_j']\,(dX_i^*/d\lambda_i) \geq 0$ for all $\lambda_i \in [0,1]$. So country i's best response is to also play $\lambda_i = 1$, and so optimal commitments $\lambda_1^* = \lambda_2^* = 1$. For part (b), observe, again using Lemma 4, that

$$\left.\frac{dS_i^*}{d\lambda_i}\right|_{\lambda_i=1} = -\left[(1-\lambda_j\theta_i\theta_j)B_i'L_i\frac{dX_i^*}{d\lambda_i}\right]_{\lambda_i=1} < 0,$$

since $(1-\lambda_j\theta_i\theta_j) > 0$ if $\theta_1 < 1$ or $\theta_2 < 1$ (since also $\lambda_j \leq 1$), and so optimal commitments $\lambda_1^* < 1$ and $\lambda_2^* < 1$ as claimed.

□ **Proof of Proposition 4**. For part (a), use the formula from Lemma 4 to obtain

$$\frac{dS_i^*}{d\lambda_i} = B_j'\left((1-\lambda_i)\theta_i - (1-\lambda_j\theta_i\theta_j)\frac{B_i'}{B_j'}L_i\right)\frac{dX_i^*}{d\lambda_i}.$$

If $\theta_i < 1$ or $\theta_j < 1$, then $dS_i^*/d\lambda_i < 0$ for all $\lambda_i \in [0,1]$ for B_i'/B_j' is sufficiently large, so that the optimal commitment $\lambda_i^* = 0$. For part (b), observe similarly that $dS_i^*/d\lambda_i < 0$ for all $\lambda_i \in [0,1]$ if θ_i is sufficiently small. So if θ_1 and θ_2 are sufficiently small, then optimal commitments $\lambda_1^* = \lambda_2^* = 0$ as claimed.

□ **Proof of Proposition 5**. In an interior equilibrium, country i's strategic choice of preference $\lambda_i^*(\lambda_j)$ is determined by the first-order condition $dS_i^*/d\lambda_i = 0$. Using the formula from Lemma 4, this condition can be written as

$$\lambda_i^*\theta_i = \theta_i - (1-\lambda_j\theta_i\theta_j)\frac{B_i'}{B_j'}L_i.$$

Now using this together with the analogous expression for country j's first-order condition for $\lambda_j^*(\lambda_i)$ yields

$$\lambda_i^*\theta_i = \theta_i - \left[(1-\theta_i\theta_j)\frac{B_i'}{B_j'}L_i + \theta_i(1-\lambda_i^*\theta_i\theta_j)L_iL_j\right],$$

and solving this for λ_i^* gives

$$\lambda_i^* = \frac{\left[\theta_i(1-L_iL_j) - (1-\theta_i\theta_j)\,\dfrac{B_i'}{B_j'}L_i\right]}{\theta_i\,(1-\theta_i\theta_jL_iL_j)}$$

as claimed. The expression for the rates of carbon leakage L_i is obtained from Lemma 1 together with some rearranging of (6). The expression for country i's abatement level $X_i^* > 0$ is obtained by rewriting its first-order condition from (5) and noting that the inverse $C_i'^{-1}(\cdot)$ is well-defined under the maintained assumptions $C_i'(\cdot) > 0$, $C_i''(\cdot) > 0$, and $C_i(0) = C_i'(0) = 0$.

□ **Proof of Proposition 6**. Since $\theta_j < 1$, it follows that $\lambda_i^* = 0$ by Proposition 4(a) for sufficiently large B_i'/B_j', and so $\lambda_i^*\theta_i = 0$. Using the result from Lemma 4, $dS_j^*/d\lambda_j \geq 0$ for almost all $\lambda_j \in [0,1]$ if B_i'/B_j' is sufficiently large, so $\lambda_j^* \lessapprox 1$, and so $\lambda_j^*\theta_j \lessapprox \theta_j$. But since $0 < \theta_j < \theta_i$, optimal commitments in this example satisfy $\lambda_j^*\theta_j > \lambda_i^*\theta_i$ as claimed.

□ **Proof of Proposition 7**. For part (a), it follows from Lemma 3 that $\lambda_j^* > 0$ must imply that $\theta_j > 0$, and so also $\lambda_j^*\theta_j > 0$. By contrast, $\lambda_j^* = 0$ obviously also implies that $\lambda_j^*\theta_j = 0$. By Lemma 1 and (6), it follows that the carbon leakage when $\lambda_j^* = 0$ equals

$$L_i|_{\lambda_j^*=0} = \frac{1}{\left(1 + \left[C_j''/(-B_j'')\right]_{\lambda_j^*=0}\right)},$$

while carbon leakage with $\lambda_j^* > 0$ is given by

$$L_i|_{\lambda_j^*>0} = \frac{\left(1 + \lambda_j^*\theta_j\left[B_i''/B_j''\right]_{\lambda_j^*>0}\right)}{\left(1 + \left[C_j''/(-B_j'')\right]_{\lambda_j^*>0} + \lambda_j^*\theta_j\left[B_i''/B_j''\right]_{\lambda_j^*>0}\right)}.$$

Observe that $\left[C_j''/(-B_j'')\right]_{\lambda_j^*>0} \leq \left[C_j''/(-B_j'')\right]_{\lambda_j^*=0}$ is a sufficient condition for $L_i|_{\lambda_j^*>0} > L_i|_{\lambda_j^*=0}$. Furthermore, note that

$$\text{sign}\left(\frac{d}{d\lambda_j^*}\left[\frac{C_j''(X_j^*)}{-B_j''(X_i^*+X_j^*)}\right]\right) = \text{sign}\left(\left[C_j'''(-B_j'') + B_j'''(1-L_i)C_j''\right]\frac{dX_j^*}{d\lambda_j^*}\right),$$

where the right-hand side is certainly non-positive if $B_j''' \leq 0$ and $C_j''' \leq 0$ (since $L_i \in (0,1)$ by Lemma 1 and $dX_j^*/d\lambda_j^* > 0$ by Lemma 2), from which the claim follows. For part (b), straightforward differentiation of the leakage rate L_i shows that it is increasing in λ_j^* if $B_i''' = 0$, $C_i''' = 0$ and B_i''/B_j'' is constant, as claimed.

□ **Proof of Lemma 5.** In an interior equilibrium, country i's strategic choice of preference $\lambda_i^*(\lambda_j)$ is determined by the first-order condition $dS_i^*/d\lambda_i = 0$. Again using the formula from

Lemma 4, this condition can be written as

$$\lambda_i^* \theta_i = \theta_i - (1 - \lambda_j \theta_i \theta_j) \frac{B_i'}{B_j'} L_i.$$

Differentiating this expression, and using the assumption that B_i'/B_j' is constant, yields

$$\frac{d\lambda_i^*}{d\lambda_j} \theta_i = \frac{B_i'}{B_j'} \left[\theta_i \theta_j L_i - (1 - \lambda_j \theta_i \theta_j) \frac{dL_i}{d\lambda_j} \right],$$

which can be rearranged as $d\lambda_i^*/d\lambda_j = (B_i'/B_j')\,(L_i/\theta_i)\,[\theta_i\theta_j(1+\eta_i) - \eta_i/\lambda_j]$, where the leakage-commitment elasticity $\eta_i \equiv (d \log L_i / d \log \lambda_j^*)$, from which the result follows immediately.

References

Andreoni, James (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *Economic Journal* 100, 464–477.

Babiker, Mustafa H. (2005). Climate Change Policy, Market Structure and Carbon Leakage. *Journal of International Economics* 65, 421–445.

Barrett, Scott (1994). Self-Enforcing International Environmental Agreements. *Oxford Economic Papers* 46, 878–894.

Barrett, Scott (2003). *Environment and Statecraft.* Oxford, UK: Oxford University Press.

Copeland, Brian R. and M. Scott Taylor (2005). Free Trade and Global Warming: A Trade Theory View of the Kyoto Protocol. *Journal of Environmental Economics and Management* 49, 205–234.

DECC (2009). *Climate Change Act 2008: Impact Assessment.* Department of Energy and Climate Change (United Kingdom), March 2009.

Edgeworth, Francis Y. (1881). *Mathematical Physics: An Essay on the Application of Mathematics to the Moral Sciences.* London, UK: Kegan Paul.

Greenstone, Michael, Elizabeth Kopits and Ann Wolverton (2011). Estimating the Social Cost of Carbon for Use in U.S. Federal Rulemakings: A Summary and Interpretation. NBER Working Paper 16913.

Heal, Geoffrey and Howard Kunreuther (2010). Social Reinforcement: Cascades, Entrapment, and Tipping. *American Economic Journal: Microeconomics* 2, 86–99.

Hoel, Michael (1991). Global Environmental Problems: The Effects of Unilateral Actions Taken by One Country. *Journal of Environmental Economics and Management* 20, 55–70.

Kolstad, Charles D. (2012). Bridging Reality and the Theory of International Environmental Agreements. In: Robert W. Hahn and Alistair Ulph, *Climate Change and Common Sense: Essays in Honour of Tom Schelling*, Oxford, UK: Oxford University Press.

IPCC (2007). *Contribution of Working Group III to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Chapter 11: Mitigation from a Cross-Sectoral Perspective. Cambridge, UK: Cambridge University Press.

Nordhaus, William D. and Joseph Boyer (2000). *Warming the World: Economic Models of Global Warming.* Cambridge, MA: MIT Press.

Ritz, Robert A. (2009). Carbon Leakage under Incomplete Environmental Regulation: An Industry-Level Approach. Working Paper at Oxford Institute for Energy Studies, November 2009.

Roelfsema, Hein (2007). Strategic Delegation of Environmental Policy Making. *Journal of Environmental Economics and Management* 53, 270–275.

Schelling, Thomas C. (1960). *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.

Tol, Richard S. J. (2011). A Cost-Benefit Analysis of the EU 20/20/2020 Package. ESRI Working Paper No. 365, January.

Vickers, John (1985). Delegation and the Theory of the Firm. *Economic Journal* 95, 138–147.

Watkiss, Paul and Chris Hope (2011). Using the Social Cost of Carbon in Regulatory Deliberations. *WIREs Climate Change* 2:6, 886–901.

Weitzman, Martin L. (1974). Prices vs Quantities. *Review of Economic Studies* 41, 477–491.