Unilateral actions as signals of high damage costs: the case of international environmental problems

Urs Steiner Brandt and Niels Nannerup

Abstract

In multilateral negotiations between nations on problems of global pollution, associated national actions to control pollution can be seen as a complex international public good. Such actions are costly and incentives to pass the main burden of reduction to other countries therefore exist. We show that when governments possess private information about national damage costs, signalling through emission levels may occur, and a variety of credible actions that manipulates emissions before negotiations (or in-between different stages of negotiation) can be identified.

In particular, we identify that unilateral actions to reduce emissions can be explained by the desire to credibly signal high damage costs, and therefore gives an explanation for unilateral actions as strategic manipulation of emissions. These incentives arise whenever pre-agreement actions can influence the final outcome of the negotiations, through reduction demands of other countries.

The implication is that unilateral actions can be seen as a credible move, in situations with private information about damage costs, and therefore a rational strategy to get progress in e.g., the climate negotiations.

JEL Classification: Q28, H4, D8

Key words: International environmental problems, reduction levels as signals, private information about abatement costs and damage costs.

1. Introduction

The climate change negotiations are progressing very slowly despite mounting evidence that serious negative consequences are unavoidable in case of continued inaction (IPCC, 2007, Stern, 2006). Therefore, much is at stake, and according to Stavin (2011), the climate change issue is the ultimate commons problem in the twenty-first century.

Even though an overall objective of not accepting global mean temperature to rise more than 2 degrees Celsius over the next 100 years (UN, 2010), so far no credible policy to reach this target has been established (IEA, 2010). The Kyoto protocol, the main international agreement to control emissions and which control period ends 2012, has so far not found any successor. Moreover, in this protocol, none of the developing economics have any reduction target. The ineffectiveness of the international society to control greenhouse gas emissions can be seen from the fact that global emissions show no trend of being reduced and emission from coal usage in developing countries is unprecedented high (IEA, 2012).

Reasons for the struggling to progress are plentiful, and can be attributed both to economic, political and distributional/moral issues. Reasons are attributed to the free riding issues (Barrett, 2003), the North-South issue and environmental justice (Gupta, 2000), and issues about collective responsibility and inclusion of major developing countries (Walsh el al., 2011). Moreover, the climate change issue still is surrounded be lots of uncertainty, regarding the amount and timing of damages, and privately held information about damages and preferences for the climate change issue held by e.g., governments. Such information comprises of strategic national interests, lobby interests, and the perceived climate risk of the population. (Holland et al, 2011, Hulme, 2009). The point of departure of our analysis is that real policy situations are to a large extent also characterised by private information between decision-makers about abatement cost and damage cost from pollution, and that countries will exploit informational advantages if possible.

The main contribution of this paper is to show that depending on the private information we consider a country might have incentive to overinvest in national climate polices prior to an agreement. This is denoted a unilateral action, and in the literature it has been a puzzle why unilateral actions have been undertaken. Certain countries, and or regions, have undertaken reductions, and/ or in the Kyoto protocol, accepted high reduction targets (relative to other comparable countries). Such unilateral actions are not easy explained. Reasons for unilateral actions has been attributed to "setting a good example" Hoel (1992), or as in Lemione and Farrell (2009) to encourage future abatement by others, which could mean focusing on the promotion of technological innovation and diffusion and on providing policy models that others could adapt to their own contexts. Our model / incentives are such that unilateral actions might also attributed to strategic moves that have the objective of improving the bargaining position in expected further negotiations. Compared to Hoel result, there a unilateral action implies less reduction by the other countries, in our setting, a credible unilateral action is a signal of high damage costs, and therefore implies that the other countries reduce more.

In the context of signalling, the present paper analyses strategic spill-over effects among nations arising from observed national policy actions in a pre-negotiation phase on global pollution. The focus is on incentives by nations to distort national emission levels prior to negotiations in order to achieve a more favourable position in the final agreement. There have been some papers addressing the issue of signalling. (Brandt, 2002, 2004, Rose and Spiegel (2009) and Jakob and Lessmann (2011)), and our present paper extend Brandt (2002) by also considering private information about damage costs. Finally, Arredondo and Garcia (2011), analyse a signalling model where a county leads the negotiation in an international environmental agreement. This country can signal its non-compliance costs through committing to the agreement. We do not consider the issue of non-compliance but assume that countries comply with the final outcome of the negotiations.

Since we are mainly interested in the possibility of manipulating pre-agreement emissions level, we focus exclusively on the possibility of separating equilibrium. That is, in our two-type framework, a situation where one type has an incentive, by a costly signal, to reveal its true type. Moreover, all focus also on first period (pre-agreement) strategic incentive. For a given institutional setting, some countries will overinvest. This situation arises in cases where a country expects that when it reduces its emission, this will imply that the other reduce sufficiently much in return, believing that the signalling country has high damage costs.

Our result adds to an understanding of action prior to an agreement, and a possible explanation to why some countries seemingly overinvest while other seemingly underinvests in national reduction effort in stages before an international environmental agreement.[1] E.g. the EU proposal to reduce 30% CO_2 can be a signal of high willingness to pay for reduction (high damage costs) while the reluctance of the USA to engage in any target, might be a signal of high abatement costs (or e.g. high political costs).[2] In order to signal true costs, extreme positions are needed regarding the pre-agreement emissions level. Finally, a "pre-agreement" stage might also be a first round of a negotiation process, like the Kyoto agreement can be seen as a first step towards to more demanding second agreement. In this case, the achieved emission reduction can also be thought of as a signaling device (or investment) for better bargaining positions in the next round of negotiations.

A remarkable lack of analysis of effects of private information in relation to international environmental agreements can be observed and to our understanding, the implications of private information have not received the attention it deserves. There are, however, few exemptions. The impact of private information on global environmental problems and their solutions has been addressed by Bac (1996), who includes incomplete information about valuation of environmental damage and Brandt (2002) who includes private information about abatement costs. Both analyses show that private information leads to inefficiency relative to the case of perfect information. Finally, Jakob and Lessmann (2011) show that a in a two-stage game early (delayed) action can act as a signal to reveal private information on high (low) benefits. The cooperative solution with asymmetric information is Pareto-dominated by the outcome with perfect information. They also develop a signaling game model and analysis the strategic incentives are to hide private information about the magnitude of a country's damage. They do, however, not consider the existence of a negotiated treaty and how pre-treaty action affects the final bargaining outcome.

Few papers address this issues about strategic consideration about how pre-agreement performance translate into outcomes of the treaty is also, an exemption is Harstad (2011), who notes that without a climate treaty, countries tend to pollute too much and invest too little, partly to induce the others to pollute less and invest more in the future. The consequence, according to Harstad is that short-term agreements on emission levels can reduce welfare, since countries invest less when they

[1] Other papers have also considered incentives arising in such a setting. Harstad (2009) and Beccherle and Tirole (2011) derive incentives for countries to lower their investments in abatement technologies to improve their future bargaining position. It is also recognized that countries might act strategically in international environmental issues is also noted by Brandt (2002) and Rose and Spiegel (2009).

[2] Signaling abatement costs are described in Brandt (2002).

anticipate future negotiations" The paper by Beccherle and Tirole (2011) analyses the consequence of the "waiting game" and find several strategic incentives to manipulate national climate policy such as to affect a country's benefit in future international climate negotiations. Their analysis is founded in a full information framework, and our model extends their reasoning to a private information setting. Essentially the same idea underlies our model, but here it is through the expectation of the type (damage cost) that pre-agreement emissions can influence the own and the other countries emission targets in the negotiations. Buchholz and Peters (2005) note that in a two-stage setting, considerable disincentives are to be expected at stage 1 are to be for a broad class of cost-sharing arrangements which generally can be attributed to the creation of positive externalities at stage 2, which is exactly the kind of incentive structure that underlies our model. See also Heitzig et al (2011), who look at self-enforcing strategies to deter free riding in climate change negotiations.

This paper is organized as follows: The formal model is presented in section 2, and thereafter section 3 where we specify the negotiations process and type of agreement, we consider and the basic the basic incentives that this implies are discussed in section 4. In the second part of the paper
We turn to the signaling approach, first defining a sequential and separating equilibrium in our setting (section 5), finding of the separating equilibrium where high damage costs countries signal high damage costs by increasing emissions (section 6) the and equilibrium refinement is explained in Section 7, and using this we find an unique prediction while section 8 concludes our paper.

2. Model

First, a model of an abstract international environmental problem is presented. The set of countries affected by and/ contributing to this problem is given by $I = \{1,2, \cdots, N\}$. Each country, donoted $i \in I$ emits $e_i \in E_i \in R_+$ of the polluting substance. For simplicity, assume a uniformly mixed pollutant giving rise to a global emission problem, such that each country is affected by the total emission level $e = \sum_i e_i$.

We consider two periods, a pre-agreement period and a period, where the agreement is settled. For climate change, the total emission of GHG in such a period adds to the stock of GHG in the atmosphere. The added emission creates additional damage, measured by $D_i(e)$. (Since the problem of climate change is a stock pollutant, the damage will be the NPV of all future damage costs due to this added emission). As usual, we assume that $D_i'(e) > 0$ and $D_i''(e) > 0$. Moreover, let $D_i'(e_i) > 0$ and $D_i''(e_i) > 0$, and $\frac{\partial D_i'(e_i)}{\partial e_{-i}} = 0$.[3]

A country receives benefit from its emission, measured by $B_i = B_i(e_i)$. Without any environmental concerns, there exists a national optimum of emissions called e_i^N defined where $B_i'(e_i) = 0$. We also assume that $\frac{\partial B_i'(e_i)}{\partial e_{-i}} = 0$. We look at situation where an interior solution exists by requiring that $B_i'(e_i) > 0$ for $e_i < e_i^N$ and $B_i'(e_i) < 0$ for $e_i > e_i^N$.

The net benefit for a country from choosing emission level e_i is given by:

$$NB_i(e_i; e) = B_i(e_i) - D_i(e)$$

We compare a situation where countries do not expect any form of international environmental agreement, (and therefore strategic interactions on national emission levels are absent), with a situation in which expectations of an agreement among countries exist and each country responds optimally to this knowledge.

In the first-mentioned situation a country will choose a given emission level derived solely from its own damage costs and abatement costs (which again depends on consumption and production pattern and technologic level), its environmental concern in general (reflected in the populations preferences for the climate change issue) and trade relations and other relations with other countries. This emission level is denoted e_i^0, and any movement of emission levels away from e_i^0 therefore implies a cost in that period to the country (in terms of lost consumption and production opportunity).

Formally, Let e_i^o be the emission level that maximizes $NB_i(e_i; e)$, that is let $e_i^o = arg\{\frac{dB_i(e_i)}{de_i} = \frac{dD_i(e)}{de_i}\}$. That is, in this one-period consideration, any change in emission away from e_i^o implies

[3] The last assumption implies that we will not consider secondary effects coming that might arise: when country *i* changes its emission and this affects the other country emission, then this might again have an effect on the optimal change of emission of country *i*.

additional costs, and given the shape of the benefit and damage functions, will be increasingly in the distance from e_i^o.

Compared to this situation, a country that is faced with the prospect of a future negotiated agreement on reductions of emission, might consider acting strategically to optimize its bargaining position in the forefront of the negotiation process. In the analysis, we focus on the pre-agreement emission level and the information about the countries damage costs this emission level might carry.

Consider the possibility that the choice of emission level for country i is guided by strategic considerations of interactions among nation. Each country may now be willing to impose additional costs on itself in the current period by departing emission levels from e_i^o if this result in higher expected benefits in the future agreement period. For this purpose, we introduce a loss function:

$$L_i(e_i) = NB_i(e_i) - NB_i(e_i^o)$$

$L_i'(e_i) = 0$ for $e_i = e_i^o$, and given the assumption on NB_i, the loss is convex increasing in the distance $\lceil e_i - e_i^o \rceil$.

Private information is crucial for our analysis. We consider the circumstances that private information can be present regarding the damage function of a country. We assume that damages both can be either high or low.

Formally, define $\theta^D = \{L, H\}$ as a parameter measuring the damages to country i of moving emission away from e_i^o. $\theta^D = L$ implies low damages and $\theta^D = H$ implies high damages.

The net benefit function can, therefore, be written as:

$$NB_i\left(e_i, e; \theta^D\right) = B_i(e_i) - D_i\left(e; \theta^D\right)$$

To be precise, let the type be defined as follows. For any given (feasible) individual and total emission, let

$$D_i(e; H) > D_i(e; L)$$

We define $e_i^o = e_i^o(\theta^D)$ as the full information non-strategic emission level given its type is $\theta^D(\theta^C)$. From the above definitions of types, we have that

$$e_i^o(L) < e_i^o(H)$$

That is, a country with high damage costs will – without any strategic considerations – have a lover emission level than if it has low damage costs.

3. Specification of the negotiations process / type of agreement

Before the negotiations take place, each country obtains perfect information about its own damage costs, whereas it remains uninformed about the types of other countries. Moreover, any country i holds a common prior probability assessment about the value of θ_j^D for all $j \in I$.[4] Let common knowledge be assumed regarding the damage types, and we write the common prior as $p_j^D = prob(\theta_j^D = H)$, and $p_j^D < 1$. In a similar manner, we have $prob(\theta_j^D = L) = 1 - p_j^D$. After observing e_j^M, the other countries update their beliefs to common posterior beliefs, given by $p_j^D(e_j^M) = prob(\theta_j^D = H|e_j^M)$.

A negotiation on alleviating a major environmental problem is a highly complex and dynamic interaction, consisting of most of the world's nations, with highly varying economic performance, and emission level. Moreover, expected damages are not evenly distributed among nation. We will make the following assumptions, which we consider to represent important features of such negotiations

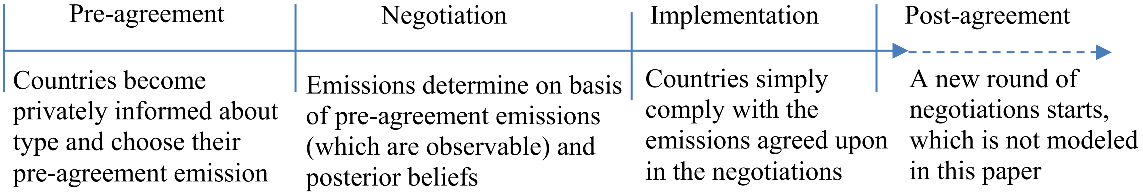
- The countries know in advance the "rules of the game", so we are not investigating into the design/architecture of an international environmental agreement (IEA).
- The solution of the IEA specifies for each particpant an emission target.
- The determination of emissions target is dependent on each countries pre-agreement emsiison level (e_i^M) and a commonly hold posterior belief about type of damage.
- All the participants comply fully with the requrements implies by the IEA.

As a consequence, all strategic behaviour takes place in the pre-agreement phase (period 1).

[4] We assume that types are not correlated between countries,, that is, knowing own type no information about other type. See Brandt (2004) for an analysis of consequences of correlation for the possibility of making unilateral actions.

Next, an emission target is agreed upon. Let a solution to an international environmental problem (that is, an agreement) specify an emission target for each participating country, and denote this solution by $e^S = \{e_1^S, e_2^S, \cdots, e_N^S\}$. For an individual country i, $e_i^S = e_i^S(\rho(e^M), e_i^M)$ such that individual emission targets in the agreement depends on the vector of posterior beliefs and its pre-agreement emission level. Figure 1 summerizes the timing of events.

FIGURE 1: *Timeline*



Pre-agreement

Countries become privately informed about type and choose their pre-agreement emission

Negotiation

Emissions determine on basis of pre-agreement emissions (which are observable) and posterior beliefs

Implementation

Countries simply comply with the emissions agreed upon in the negotiations

Post-agreement

A new round of negotiations starts, which is not modeled in this paper

A large set of agreements (S) exits that has such a feature. We will narrow down the class of soultion to solutions with the following property: We are interested in the class of solutions denoted by S^G and defined as:

$$S^G = \{S | \frac{\partial e_i^S}{\partial \rho_i^D} \leq 0, \frac{\partial e_j^S}{\partial \rho_i^D} < 0\}$$

The signs of the derivatives in this solution reflects natural responses on damage cost arguments in a process of negotiation on burden sharing: A country credibly claiming high damage costs, roughly speaking, increases the seriousity of the environmental problem among negotiaters, resulting in acceptance of higher reductions among all countries.[5] Note that e.g. the solution implementing the globally optimal emission levels, defined by $e_i^C : \frac{\partial B_i}{\partial e_i} = \sum_j \frac{\partial D_j}{\partial e_i}$ is in S^G, as well as the Nash bargaining solution or any uniform solution of the type, where $e_i^S = \alpha' \cdot e_i^o$, where $(1 - \alpha')$ is the common percentage reduction level.[6]

The two-period net benefit function is given by

$$NB_i^T(e_i^M, \rho(e_i^M); \theta^D) = B_i(e_i) - D_i(e_i^M, e_{-i}^M \theta^D) + \delta[B_i(e_i^S) - D_i(e_i^S, e_{-i}^S \theta^D)]$$

[5] The essence here is that we consider solutions with particular characteristics where signalling is possible. Other types of arrangements could be considered where signalling is not relevant, like a solution where each participant reduces a fixed level, independent of country characteristics. The class of solutions in focus is rather general and encompasses most relevant cases implying that the presented analysis is highly relevant for most cases.

Where $e_i^S = e_i^S(\rho(e^M), e_i^M)$ and δ is the discount factor and with emissions under the agreement (period 2 emissions) being determined from the posterior beliefs variable ρ^D as discussed above.

4. Basic incentives when damage costs are private information

As a clarification of the underlying incentives of countries in this set-up, it is useful to analyse how the net benefit to a country changes as a function of the common beliefs that other countries hold about this country. We do this by looking at the changes in posterior beliefs for given emission level of this country, and the effect on e_i^S. Before doing that, a useful result is stated below:

Lemma 1: (incentives to increase own emission in an agreement): For any $e^S \in S$, where $e_i^S < e_i^0, \forall i \in I$ country i prefers an increase in individual emission e.g., $\frac{\partial NB_i(e^S)}{\partial e_i^S} < 0$.

The argument is that for any solution, where $e_i^S < e_i^0, \forall i \in I$, it is optimal to increase emission unilateral. Given any set of emissions that is the result of an agreement, a country would gain individually from an increase in its own emission. (This because $e_i^S < e_i^0$). This result is valid as long as all other countries emissions are hold constant for a change in e_i^S. Our focus here is to derive how the net benefit changes with posterior beliefs.

Private information about damage cost

$$NB_i^T(e_i^M, \rho(e_i^M); \theta^D) = B_i(e_i) - D_i(e_i^M, e_{-i}^M \theta^D) + \delta[B_i(e_i^S) - D_i(e_i^S, e_{-i}^S \theta^D)]$$

Differentiating $NB_i^T(e_i^M, \rho(e_i^M); \theta^D)$ with respect to ρ_i^D yields:

$$\frac{dNB_i^T}{d\rho_i^D} = \delta[\underbrace{\frac{d[B_i(e_i^S)}{de_i^S} \cdot \frac{\partial e_i^S}{\partial \rho_i^D}}_{-} - \sum_j \underbrace{\frac{dD_j(e^S)}{de^S} \cdot \frac{\partial e^S}{\partial \rho_i^D}}_{+}$$

The sign is ambiguous and this gives us the following result:

Result 1: We have two situations (assuming "=" is unlikely)

1) $\frac{dNB_i^T}{d\rho_i^D} > 0$ Here a country gains from being perceived as having *high* damage costs.

2) $\frac{dNB_i^T}{d\rho_i^D} < 0$ Here a country gains from being perceived as having *low* damage costs.

The derivations tell that when beliefs that the country has high damage costs increase, then in bargaining situation all countries emissions will be smaller. For country this implies higher costs, given lemma 1, due to the decrease in own emission, but on the other hand it benefits from the reduced emission of other countries. Which effect is the dominating one is not to determine, unless a specific IEA and its bargaining process is specified. In this analysis we focus solely on the first situation, which is the unilateral action case. Situation two is the case where countries will undertake action to show having low damage costs.[7]

The strategic incentive in situation 1 is to signal high damage. In our setting, high damage cost is associated with a low first-period emission (relative to have low damage costs). More precisely, a low damage type would be tempted to invest too much in national climate policies to signal high damage costs and thereby get a "better" deal in the second-period agreement.

5. Sequential equilibrium

We now proceed with the formal signalling model. In the signalling game, we have a sender and a receiver. The sender is an individual country, sender a signal, the pre-emission level, e_i^M. The receiver is the "collective negotiation body". Consistent with our interpretation of the participants to the negotiations, that there exists a common understanding about the formation of posterior beliefs upon observation of the pre-emission levels.

A collection (of strategies and beliefs) forms a sequential equilibrium if the following conditions are met:

(i) Optimality for country i:

[7] The strategic incentive in situation 2, on the other hand, is to signal low damage. In our setting, low damage cost is associated with a high first-period emission. More precisely, a high damage would be tempted to invest too little in national climate policies to signal low damage costs and thereby get a "better" deal in the second-period agreement.

$$\hat{e}_i(\theta^D) \in argmax\{NB_i^T\left(e_i^M, e_S\left(\hat{\rho}_i(e_i^M)\right);\theta^D\right)\}$$

(ii) Consistency of beliefs:

If $\hat{e}_i(L) \neq \hat{e}_i(H)$ then $\hat{e}_i(L) = 0$ and $\hat{e}_i(H) = 1$
If $\hat{e}_i(L) = \hat{e}_i(H) = \hat{e}_i^p$ then $\hat{e}_i(L) = \hat{e}_i(H) = \rho^o$
If $e_i \notin \{\hat{e}_i(L), \hat{e}_i^p, \hat{e}_i(H)\}$ then any $\rho(e_i)$ are admissible.

In a separating equilibrium, the two types are separated and both are perfectly recognized by their true types, while in a pooling situation no new information is revealed and the beliefs are not revised. To fully describe the set of possible separating equilibrium outcomes, we assume that out-of-equilibrium signals are followed by the most unfavourable beliefs seen from the sender's point of view implying that $\rho(e_i) = 0$ if $e_i \notin \{\hat{e}_i(L), \hat{e}_i^p, \hat{e}_i(H)\}$.

Finally, as already noted and motivated in the introduction, we only look at separating equilibrium in this analysis.

6. Signalling high damage costs by increasing emissions

In this section we analyse the case of damage cost private information. We look at $\frac{dNB_i^T}{d\rho_i^D} > 0$, implying a second period gain for a low damage cost country in being perceived as having high damage costs with a positive probability. Under private information, the sender wants the receiver to *believe* abatement damage costs are high, and to be perceived as such a type, the country must increase its emission.

Note that with regards to $\rho(e^M)$, $NB_i^T(e_i^M, \rho_i(e_i^M); \theta^D)$ is maximized for $\rho_i(e_i^M) = 1$ and minimized for $\rho_i(e_i^M) = 1$.

Therefore, if a sender knows it cannot change beliefs, (or it is too costly to do so), and given the out-of-equilibrium beliefs specified in section 5, it as well can choose the emission level that maximizes its net benefit function given that it will be perceived as having low damage cost for

certain. This amounts for country i to maximize $NB_i^T(e_i, 0; \theta^D)$. For consistent notation, we denote this emission level for $e_i^M(0; \theta^D)$, and the net benefit as $NB_i^T(e_i^M(0; \theta^D), 0; \theta^D)$.
The benefit from increasing the beliefs about damage costs are received in the second period and we will call the second period net benefit as $NB_i^2\left(e_i^S\left(\rho_i(e_i^M)\right), e_{-i}^S\left(\rho_i(e_i^M)\right); \theta^D\right)$.

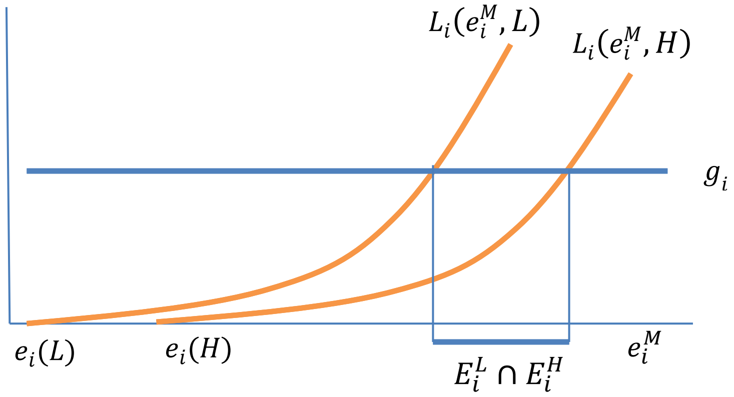
From this we can define the following two functions, which measures the second period gains from successful signalling and the corresponding first period cost (investment) needed to change beliefs in the second period:

Gain: $g_i(e_i^M, \rho_i(e_i^M); \theta^D) = NB_i^2\left(e_i^S\left(\rho_i(e_i^M)\right), e_{-i}^S\left(\rho_i(e_i^M)\right); \theta^D\right) - NB_i^2(e_i^S(0), e_{-i}^S(0); \theta^D)]$

Loss: $L_i(e_i^M, \theta^D) = NB_i^1(e_i^M, \theta^D) - NB_i^1(e_i^M(0; \theta^D), \theta^D)$.

We can best explain these functions by using Figure 2. First take the gain function. We are looking at separating equilibria where out of equilibrium are such that if a country cannot fully convince that its type is high, it will be perceived as having low damage costs. Therefore, the gain, if a country can convince that it has high damage costs will be the net benefit from changing beliefs from 0 to 1). For a given problem this is simply a constant value, represented by a straight line. For simplicity in the Figure we assume that this is equal for the two types.

Figure 2: *The possibility of a separating equilibrium outcome*



The loss function is convex increasing from $e_i^M(0; L)$ and $e_i^M(0; H)$ for the two types, respectively. In Figure 11, we show a situation, where separating equilibria exists. Existence depends on $e_i^M(0; L)$ compared to $e_i^M(0; H)$ and $\frac{\partial L_i(e_i^M, H)}{\partial e_i^M}$ compared to $\frac{\partial L_i(e_i^M, L)}{\partial e_i^M}$.

As seen from the Figure, existence is guaranteed if

$$e_i^M(0;L) > e_i^M(0;H)$$

$$\frac{\partial L_i(e_i^M,H)}{\partial e_i^M} \geq \frac{\partial L_i(e_i^M,L)}{\partial e_i^M}$$

Let us now derive this formally. For the high cost type to separate in a sequential equilibrium there must exist an $e_i^M > e_i^o$ such that:

(C1) $g_i(e_i^M,1;H) \geq L_i(e_i^M,H)$

(C2) $g_i(e_i^M,1;L) < L_i(e_i^M,L)$

Conditions C1 and C2 states that for a separating equilibrium to exists, there must exists emissions level where it is beneficial for the high type to be perceived as having high damage, while for the same emission it is not beneficial for the low cost type to be perceived as having high damage. We can define the following sets related to these conditions:

$$E_i^H = \{e_i^M \in E_i | g_i(e_i^M,1;H) \geq L_i(e_i^M,H)\}$$

$$E_i^L = \{e_i^M \in E_i | g_i(e_i^M,1;L) < L_i(e_i^M,L)\}$$

Finally, define

$$\bar{e}_i^H = arg\{g_i(e_i^M,1;H) = L_i(e_i^M,H)\}$$

$$\bar{e}_i^L = arg\{g_i(e_i^M,1;L) = L_i(e_i^M,L)\}$$

It is possible to establish the following result:[8]

Proposition 1: Sequential separating equilibria exist if $\bar{e}_i^H > \bar{e}_i^L$ and have the following structure:

$$\hat{e}_i(L) = e_i^M(0;L)$$
$$\hat{e}_i(H) \in E_i^L \cap E_i^H$$

Figure 2 shows a situation where an e_i exists that satisfies (C1) and (C2). The reason why a separating equilibrium is not guaranteed is that while it is more costly for the H-type to increase emissions than for the low cost type, the H-type has a larger emission in the non-cooperative

[8] We omit the proof, see Brandt (2004) for a thorough treatment of this case.

setting. Hence, only in cases where the deviation between non-cooperative emission levels is large or where the difference in costs is low, it is likely that the high costs type can separate.

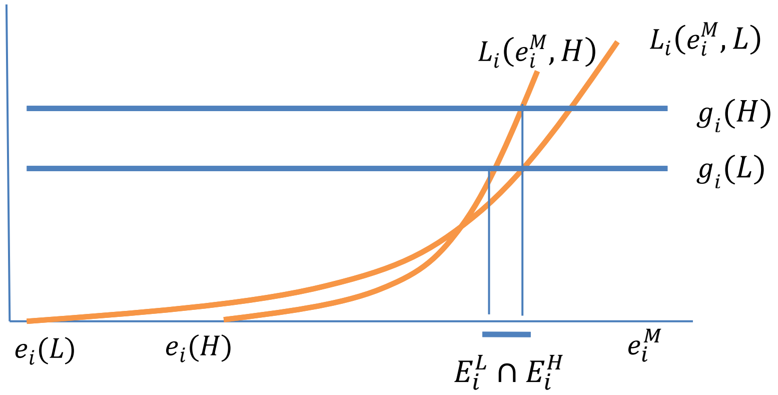
Proposition 2: A separating equilibrium is asserted given either of the two conditions are met (these are sufficient conditions):

1) $e_i^M(0;L) > e_i^M(0;H)$ and $\frac{\partial L_i(e_i^M,H)}{\partial e_i^M} = \frac{\partial L_i(e_i^M,L)}{\partial e_i^M}$

2) $e_i^M(0;L) = e_i^M(0;H)$ and $\frac{\partial L_i(e_i^M,H)}{\partial e_i^M} < \frac{\partial L_i(e_i^M,L)}{\partial e_i^M}$

As seen in Figure 2, we would expect that if separating equilibira exists, then there is a whole interval of emissions that satisfy condition C1 and C2. To make more precise prediciton, we turn to the technique of equilibirum refinements to eliminate all but one equlilibrium outcome. The conditions are minimum conditions to be met.

Figure 3: *Another possibility of a separating equilibrium outcome*



An important issue is wheter it is likely that condition as specified in proposition 2 are satisfied. One major concern is that the high damage cost type, since it has high enphasis on damage, also likely to have high costs of deviating from $e_i^M(0;H)$ compared to the low costs type. On the other hand, it could also be likely that the high damage type would value the second period gain higher than the low damage cost type. An example of this is shown in Figure 3, where we still have an separating equilibrium even though the conditions in propostion 2 are not met.

7. Equilibrium refinements

Since the set of separating outcomes is large, a selection among them is necessary in order to obtain a unique prediction of the signalling game. In the following this selection is done by use of equilibrium refinements.[9] Such refinements used for signalling games are based on the notion of forward induction, asserting that rational players in evaluating strategies would reason from the beginning of the game-tree by using introspection, i.e. by examining which players would have an incentive to send possible out-of-equilibrium messages, and rational players would then revise beliefs accordingly. Given it is common knowledge among players that everyone engages in this introspection process, an implicit communication emerges.

To see how refinements based on forward induction will work, imagine that a player picks a candidate equilibrium outcome and reviews the beliefs about out-of-equilibrium information sets sustaining this outcome. The player then applies a refinement criterion that describes what constitutes a reasonable belief. If, by taking into account the reasonableness of these beliefs and believing that the other players do so too, at least one player has an incentive to deviate, then this outcome is no longer an equilibrium in the refined game.

The requirement for formation of beliefs applied in the present analysis says it should be common knowledge among rational players that they never play a strategy profile a particular player has no incentive to play. We say that a strategy e_i^1 is weakly dominated by another strategy e_i^2 for type θ, if, no matter what beliefs the uninformed player may have after observing the move of the informed player, the expected payoff of playing e_i^2 always exceeds the expected maximum payoff of playing e_i^1 for the informed player. To minimize notation, Let the total net cost function of playing e.g., e_i^1 be $NC_i^T(e_i^1, \rho, \theta_i^l)$.

We present the definitions with respect to private information about abatement costs:

[9] For more on refinements of signaling games, see e.g., Fudenberg and Tirole (1993), Cho and Kreps (1987) and in this context, Brandt (2002).

Definition 1: Weakly dominated (WD) strategy: A strategy e_i^1 is WD by e_i^2 for type θ_i^C, if $\max_\rho NC_i^T(e_i^2,\rho_i,\theta_i^C) \leq \min_\rho NC_i^T(e_i^1,\rho,\theta_i^C) \Rightarrow NC_i^T(e_i^2,0,\theta_i^C) \leq NC_i^T(e_i^1,1,\theta_i^C)$.

It appears from Definition 1 that for e_i^1' to be weakly dominated by e_i^2, the even in the case where e_i^2 is followed by the worst possible circumstances from the point of view of the informed player, this reduction level is still preferred to e_i^1, even when e_i^1 is followed by the best possible circumstances. By invoking the following requirement, we reduce the set of separating equilibria in focus. If a strategy (signal, emission level) e_i is weakly dominated for one type, θ_i^j but not for the other type, then the uninformed players' belief should place zero probability that θ_i^j has sent e_i, i.e. e_i must be followed by posterior beliefs $\rho(\theta_i^j \mid e_i) = 0$.

Applying this equilibrium selection criteria results in a unique prediction concerning a separating equilibrium for private information on damage costs:

Proposition 3: Given $E_i^L \cap E_i^H$ is non-empty one undominated separating equilibrium exists:

$$\hat{e}_i(L) = e_i^M(0; L),$$
$$\hat{e}_i(H) = \bar{e}_i^L$$

Proof, first, for the L-type, via definition of E_i^L, all $e_i^M \in E_i^L$ are weakly dominated by $e_i^M(0; L)$. On the other hand, also via definition of E_i^H, none of $e_i^M \in E_i^H$ are weakly dominated by $e_i^M(0; H)$. Next fix any candidate equilibrium $\hat{e}_i(H) > \bar{e}_i^L$, if the receiver observes a $e_i^M = \hat{\mathrm{e}}_\mathrm{i}(\mathrm{H}) - \varepsilon$, posterior beliefs should be updated to $\rho_i(e_i^M) = 1$, and consequently, $\hat{e}_i(H)$ is no longer a sequential equilibrium. This can be done for all $\hat{e}_i(H) > \bar{e}_i^L$. The only non-dominated sequential equilibrium is $\hat{e}_i(H) = \bar{e}_i^L$.

In proposition 3, $\hat{e}_i^H$ is the lowest emission level in the set of separating equilibrium outcomes. The intuition is that the high damage type uses least costly actions in order to separate from the shadow of the low damage type.

8. Conclusion

The aim of this paper has been to investigate into incentives in order to try to predict how these incentives might distort the countries actions before entering an agreement.

However, the question remains whether governments, the players in our games, really behave like game theory suggests. This issue is also discussed in Barrett (2003), and as he notes that, fundamentally, we do not know and we will probably never know. On the other hand, as also noted by Barrett (2003) say that most agreements fail to alter the state government significantly, since incentives are not supportive for a self-enforcing agreement. Hence, the implicit claim here is that countries do act on economic incentives.

Therefore, our intention is to lay out incentives that are surrounding negotiations about the control of international environmental problems. Once such incentives are present, countries will either react on these incentives, or believe that others do, creating a situation with less trustworthiness. That is why we believe that our analysis is important, in order to restructure the incentives such that countries behaviour can be altered such that cooperation can be sustained, all the incentives must be identified.

Our analysis shades new light in the prospect of unilateral actions as a way forward in the impasse of the climate negotiations. Our results imply that given that the conditions specified in our analysis are met, the unilateral moves a rational way of improving the achievement in terms of overall emissions reduction of a given agreement. Secondly, it also points to that significantly effort might be necessary in order to act credibly. This could explain the EU proposal of a 30 % reduction as a credible signal that EU government and citizens have high (perceived) damage costs and therefore push other countries to reduce more themselves.

9. References

Arredondo, A.E and F. M. García (2011), ‘Free-riding in international environmental agreements: A signaling approach to non-enforceable treaties’, *Journal of Theoretical Politics*, **23**, 111-134.

Beccherle, J. and J. Tirole (2011), Regional initiatives and the cost of delaying binding climate change agreements’ , *Journal of Public Economics*, **95**, 1339–1348.

Barrett, S. (2003), Environment and statecraft, the strategy of environmental treaty-making. UK: Oxford University Press.

Brandt, U.S (2002), ‘Actions Prior to Entering an International Environmental Agreement’, *Journal of Institutional and Theoretical Economics,* **158**, 695-714.

Brandt, U.S. (2004) ‘Unilateral actions, the case of international environmental problems’, *Resource and Energy Economics*, **26**, 373-391.

Buchholz, W. and W. Peters (2005), ‘The Distortive Effect of Efficient Negotiation Procedures’, http://www.wiwi.europa-uni.de/de/lehrstuhl/fine/fiwi/team/peters/Buchholz-Peters-negotiation-final.pdf.

Cho, In-K. and D. M. Kreps (1987), ‘Strategic stability and Uniqueness in signalling games’, *Quarterly Journal of Economic Theory*, **102**, 179-221.

Fudenberg, D. and J. Tirole (1993), Game theory, MIT Press, Cambridge, Massachusetts.

Gupta, J. (2000), North-South aspects of the climate change issue: towards a negotiating theory and strategy for developing countries’, *International Journal of Sustainable Development*, **3**, 115-135.

Harstad, B. (2009), ‘The Dynamics of Climate Agreements’, Harvard Project on International Climate Agreements’, Discussion Paper 09-28, Harvard Kennedy School.

Harstad, B (2011), ‘The Dynamics of Climate Agreements’, mimeo, Northwestern University.

Heitzig, J, K. Lessmann and Y. Zou (2011),’ Self-enforcing Strategies to Deter Free Riding in the Climate Change Mitigation Game and Other Repeated Public Good Games’, *Proceedings of the National Academy of Sciwence of the United States of America (PNAS),* **108**, 15739-15744.

Hoel, M. (1992), ‘International environment conventions: The case of uniform reductions of emissions’, *Environmental and Resource Economics*, **2**, 141-159.

Holland, S., J. Hughes, C. Knittel, and N. Parker (2011), ‘Some inconvenient truths about climate change policy: The distributional impacts of transportation policies’, Working papers, Department of Economics, University of North Carolina.

Hulme, M. (2009): Why we disagree about climate change: Understanding controversy, inaction and opportunity, Cambridge, United Kingdom: Cambridge University Press.

IEA (2010). International Energy Agency (IEA): Responding to Climate Change: A Brief Comment on International Emissions Reduction Pledges. http://www.iea.org/journalists/docs/pledges.pdf

IEA (2012). World Energy Outlook 2012.

IPCC (2007) Summary for policymakers. Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC), eds. Solomon S, et al. (Cambridge University Press, Cambridge, UK).

Jakob, M. and K. Lessmann (2011), 'Signaling in International Environmental Agreements: The Case of Early and Delayed Action', Working paper, Potsdam Institute for Climate Impact Research.

Lemoine, D.M. and A.E. Farrell (2008), 'The Strategic Value of Unilateral Abatement in Games of Climate Change Policy', USAEE WP 08-016, University of California, Berkeley

Rose, A.K. and Spiegel, M.M. (2009): Noneconomic Engagement and International Exchange: The Case of Environmental Treaties, *Journal of Money, Credit and Banking*, **41**, 337-363.

Stavins, R. N. (2011), 'The Problem of the Commons: Still Unsettled after 100 Years', *American Economic Review,* **101**, 81-108.

Stern, N. (2006), Stern Review on the Economics of Climate Change, Cambridge University Press.

UN (2010): COP15/CMP5: Analysis of the Process, Outcomes and Implications. http://www.unep.org/ROA/amcen/docs/AMCEN_Events/climate-change/COP15_Analysis.pdf

Walsh, S., H. Tian, J. Whalley and M. Agarwal (2011), China and India's participation in global climate negotiations, *International Environmental Agreements*, **11**, 261–273.