“Unilateral actions as signals of high damage costs: distorting pre-negotiations emissions in international environmental problems”

AUTHORS	Urs Steiner Brandt Niels Nannerup

ARTICLE INFO	Urs Steiner Brandt and Niels Nannerup (2013). Unilateral actions as signals of high damage costs: distorting pre-negotiations emissions in international environmental problems. *Environmental Economics*, *4*(2)
RELEASED ON	Tuesday, 02 July 2013
JOURNAL	"Environmental Economics"
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES
0

NUMBER OF FIGURES
0

NUMBER OF TABLES
0

Urs Steiner Brandt (Denmark), Niels Nannerup (Denmark)

Unilateral actions as signals of high damage costs: distorting pre-negotiations emissions in international environmental problems

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. The authors 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, the paper identifies 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.

Keywords: international environmental problems, reduction levels as signals, private information about damage costs.
JEL Classification: Q28, H40, D80.

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 characterized by private information between decision-makers about 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 a country might have incentive to overinvest in national climate polices prior to an agreement. This denotes 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 (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's result, where 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. Note that if signalling succeeds, the analysis in this paper shows that the total emission level will decrease, and therefore our result is in contrast to the findings of Hoel (1992).

In the context of signalling, the present paper analyzes 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 favorable 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, 2012), and our present paper extend Brandt (2002) by also considering private information about damage costs. Finally, Arredondo and Garcia (2011), analyze 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 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). In order to signal true damage 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 signalling 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 (2012) show that 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 signalling 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 the issue about strategic consideration about how pre-agreement performance translates into outcomes of the treaty. 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 anticipate future negotiations. The paper by Beccherle and Tirole (2011) analyzes 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 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.

[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).

This paper is organized as follows. The formal model is presented in section 1. In section 2 we specify the negotiations process and type of agreement. The basic incentives that this implies are discussed in section 3. Section 4 first defines a sequential and separating equilibrium in our setting and hereafter presents the finding of the separating equilibrium where high damage costs countries signal high damage costs by decresing emissions. The equilibrium refinement is explained in section 5. The final section concludes the paper.

1. 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 $I = \{1, 2, \ldots N\}$. Each country, denoted $i \in I$ emits $e_i > 0$ 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 $e = \Sigma_i e_i$.

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 emission creates 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, $\frac{\partial D_i'(e_i)}{\partial e_{-i}} = 0$, where e_{-i} denotes the emission of all other countries than country i[1].

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^0 be the emission level that maximizes $NB_i(e_i; e)$, that is let $e_i^0 = \arg\left\{\frac{dB_i(e_i)}{de_i} = \frac{dD_i(e)}{de_i}\right\}$. That is, in this one-period consideration, any change in emission away from e_i^0 implies additional costs, and given the shape of the benefit and damage functions, will be increasingly in the distance from e_i^0.

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^0 if this result in higher expected benefits in the future agreement period.

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 $\theta = \{L, H\}$ as a parameter affecting the damage from emission. $\theta = L$ implies low damages and $\theta = H$ implies high damages. To be precise, let the type be defined as follows. For any given (feasible) individual and total emission, let $D_i(e; H) > D_i(e; L)$.

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

$$NB_i(e; \theta) = B_i(e_i) - DB_i(e; \theta).$$

We define $e_i^0 = e_i^0(\theta^D)$ 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 $e_i^0(L) > e_i^0(H)$.

[1] 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.

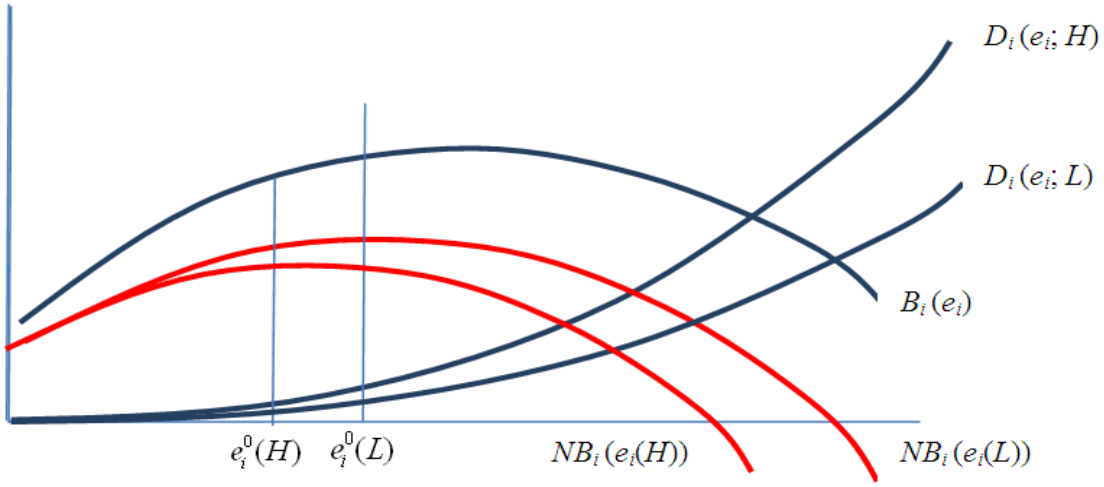

Fig. 1. Full information, non-strategic situation

That is, acountry with high damage costs will – without any strategic considerations – have a lower emission level than if it has low damage costs. See Figure 1 for a graphic illustration of this property.

2. 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[1]. Moreover, any country i holds a common prior probability assessment about the value of θ_i for all $i \in I$. Let common knowledge be assumed regarding the damage types, and we write the common prior as $p_i = prob\ (\theta_i = H)$, and $p_i < 1$. In a similar manner, we have $prob\ (\theta_i = L) = 1 - p_i$. After observing e_i^M, the other countries update their beliefs to common posterior beliefs, given by $p_i\left(e_i^M\right) prob\left(\theta_i = H \mid e_i^M\right)$.

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 the design/ architecture of an international environmental agreement (IEA).
- The solution of the IEA specifies for each participant an emission target.
- The determination of emissions target in the agreement is dependent on each country's pre-agreement emission.
- Emission level $\left(e_i^M\right)$ are realized and commonly hold beliefs about type of damages are assigned.
- All the participants comply fully with the requirements implies by the IEA.

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

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 $e^S = \left\{e_1^S, e_2^S, ..., e_N^S\right\}$. For an individual country i, $e_i^S = e_i^S\left(\rho_i\left(e_i^M\right) e_i^M\right)$ such that individual emission targets in the agreement depends on the vector of posterior beliefs and its pre-agreement emission level. Figure 2 summarizes the timing of events.

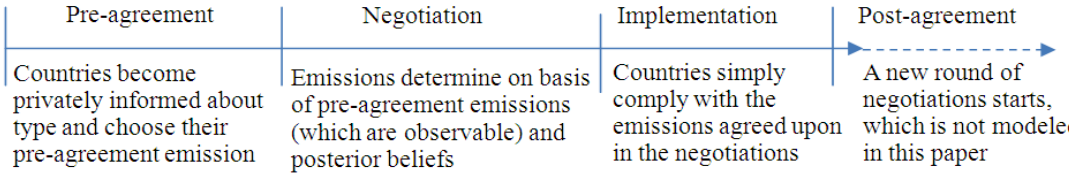

Fig. 2. Timeline

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

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

$$S^G = \left\{ S \mid \frac{\partial e_i^S}{\partial \rho_i} \le 0, \frac{\partial e_j^S}{\partial \rho_i} < 0 \right\}.$$

The sign 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 seriousness of the environmental problem among negotiators, resulting in acceptance of higher reductions among all countries[1]. 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 $e_i^S = \alpha' \cdot e_i^0$, where $(1-\alpha')$ is the common percentage reduction level.

The two-period net benefit function is given by

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

where $e_i^S = e_i^S\left(\rho_i\left(e_i^M\right) e_i^M\right)$ and δ is the discount factor.

3. 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 analyze 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 (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. Differentiating $NB_i^T\left(e_i^M, \rho_i\left(e_i^M\right); \theta\right)$ with respect to $\rho_i = \rho_i\left(e_i^M\right)$ yields:

$$\frac{dNB_i^T}{d\rho_i} = \delta\left[\frac{dB_i\left(e_i^S\right)}{de_i^S} \cdot \frac{\partial e_i^S}{\partial \rho_i} - \sum_j \frac{dD_j\left(e^S\right)}{de^S} \cdot \frac{\partial e^S}{\partial \rho_i} \right] = sign\{?\}$$
$$+ \qquad - \qquad + \qquad -$$

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

1. $\frac{dNB_i^T}{d\rho_i} > 0$. Here a country gains from being perceived as having high damage costs.
2. $\frac{dNB_i^T}{d\rho_i} \le 0$. Here a country does not gain from being perceived as having high 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 (2) is the case where countries will undertake action to show having low damage costs[2].

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.

[1] 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.

[2] 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.

In this paper we analyze the case of damage cost private information. We look at $\frac{dNB_i^T}{d\rho_i^D} > 0$ which implies that the sender wants the receiver to *believe* abatement damage costs are high, and under private information, to be perceived as such a type, the country must increase its emission.

Note that with regards to $\rho_i\left(e_i^M\right) = NB_i^T\left(e_i^M, \rho_i\left(e_i^M\right);\theta\right)$ is maximized for $\rho_i\left(e_i^M\right) = 1$ and minimized for $\rho_i\left(e_i^M\right) = 0$.

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 4, 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 $NB_i^T\left(e_i, 0;\theta\right)$. For consistent notation, we denote this emission level for $\left(e_i, 0;\theta\right)$, and the net benefit as $NB_i^T\left(e_i(0;\theta), 0;\theta\right)$.

The two-period net benefit function is given by

$$NB_i^T\left(e_i^M, \rho_i\left(e_i^M\right);\theta\right) = NB_i^1(e_i^M(\theta),\theta) +$$
$$+\delta\left[NB_i^2\left(e_i^S\left(\rho_i\left(e_i^M\right)\right), e_{-i}^S\left(\rho_i\left(e_i^M\right)\right);\theta\right)\right].$$

Ultimately, given the structure of the sequential separating equilibrium and given the specification of the out of equilibrium beliefs, the country in question has to make a choice between two situations. It can either accept that it will be recognized as having low damage costs with certainty, and optimize given that situation. Here, the country will choose $e_i(0, \theta)$ and its total net benefit will be:

$$NB_i^T\left(e_i(0,\theta), 0;\theta\right) = NB_i^1(e_i(0,\theta),\theta) +$$
$$+\partial[NB\partial_i^2(e_i^S(0), e_{(-i)}^S(0);\partial)].$$

It can however also reduce first-period emission level sufficiently such as being recognized as having high damage costs. Denoting the choice of emissions as $\left(e_i^M\right)$, the net benefit is given by:

$$NB_i^T\left(e_i^M, 1;\theta\right) = NB_i^1(e_i^M(\theta),\theta) + \delta\left[NB_i^2\left(e_i^S(1), e_{-i}^S(1);\theta\right)\right].$$

Finally, focus in this analysis is on damage cost uncertainty. Therefore, to focus exclusively on the this, we will employ the following assumption throughout the paper.

Assumption A1:

$$B_i\left(e_i^M;H\right) - B_i\left(e(H,0);H\right) = B_i\left(e_i^M;L\right) - B_i\left(e(L,0);L\right).$$

Assumption A1 essentially states that the change in benefit from e (0, θ) to e_i^M is identical for both types. This is not a nesseary condition, in the sence that our results can still hold if A1 is not satisfied, but it makes the results more easy assesable and highlights the focus of our analysis on the damage costs side.

4. 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 separating sequential equilibrium if the following conditions are met:

1. Optimality for country *i*:
   $\hat{e}_i(\theta) \in argmax\left\{NB_i^T\left(e_i^M, \hat{p}_i\left(e_i^M\right);\theta\right)\right\}$.
2. Consistency of beliefs:

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

As already noted and motivated in the introduction, we only look at separating equilibrium in this analysis. In a separating equilibrium, the two types are separated and both are perfectly recognized by their true types. 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 $p\left(e_i\right) = 0$ if $\hat{e}_i \notin \left\{\hat{e}_i(L), \hat{e}_i(H)\right\}$.

Given these beliefs, a sufficient condition for a strategy pair to form a separating equilibrium for $e_i^M < e(0,H)$ is that:

P1: $NB_i^T\left(e_i^M, 1; H\right) \geq NB_i^T\left(e_i(0,H), 0; H\right)$.
P2: $NB_i^T\left(e_i^M, 1; L\right) < NB_i^T\left(e_i(0,L), 0; L\right)$.

P1 states that an e_i^M exists that will make a high damage cost type better off by choosing this emission level and being perceived as having high damage cost than accepting not to being recognized as having high damage costs and maximizing net benefit under that situation. P2 then implies the same is not true for the low damage cost type: Even if receiving the highest possible belief, it will not be optimal for the low cost type to choose e_i^M.

To describe the set of emissions that are equilibrium strategies, it is convenient to define the following two sets:

$$P_i^H = \left\{ e_i^M < e(0,H) \mid NB_i^T\left(e_i^M,1;H\right) \geq NB_i^T\left(e_i\left(0,H\right),0;H\right) \right\},$$

$$P_i^L = \left\{ e_i^M < e(0,H) \mid NB_i^T\left(e_i^M,1;L\right) < NB_i^T\left(e_i\left(0,L\right),0;L\right) \right\}.$$

Finally, to guarantee that $P_i^L \cap P_i^H$ is non-empty, define the following version of a "single crossing property" (SCP), which is explained in details below:

$$\delta\left[D_i\left(e_i^S(0),e_{-i}^S(0);H\right) - D_i\left(e_i^S(1),e_{-i}^S(1);H\right)\right] -$$
$$- \delta\left[D_i\left(e_i^S(0),e_{-i}^S(0);L\right) - D_i\left(e_i^S(1),e_{-i}^S(1);L\right)\right] >$$
$$> \left[D_i\left(e_i^M),e_{-i}^M;H\right) - D_i\left(e\left(0,H\right),e_{-i}^M;H\right)\right] -$$
$$- \left[D_i\left(e_i^M,e_{-i}^M;L\right) - D_i\left(e\left(0,L\right),e_{-i}^M;L\right)\right].$$

We can now state the conditions under which such equilibria exists.

Proposition 1: Separating equilibrium satisfying P1 and P2 exists with $\hat{e}_i(H) \in P^L \cap P^H$ and $\hat{e}_i\left(L\right) = e(L,0)$, given that SCP is satisfied (see Appendix for proof). The SCP, which is a version of the generic single crossing property (SCP) adjusted to this context, states that the total change in damage must be higher for the low damage cost type than for the high damage costs type. The part $D_i\left(e_i^S(0),e_{-i}^S(0);H\right) - D_i\left(e_i^S(1),e_{-i}^S(1);H\right)$ measures the benefit in terms of lower damage in the second period from being recognized as a high damage cost. Therefore, the left hand side of the equation measures the additional gain that the high cost country receives compared to the low damage cost type. Due to the slope of the damage-functions, the left hand side is positive. $D_i\left(e_i^M,e_{-i}^M;H\right) - D_i\left(e\left(H,0\right),e_{-i}^M;H\right)$ measures the first period additional damage for country *i* of the higher emission $e\left(H,0\right) > e_i^M$. Both brackets on the RHS are negative, such the sign of the right hand side is ambiguous. Therefore, this condition is needed to ensure the existence of the separating equilibrium.

To provide a graphical explanation of the properties of the set of separating equilibrium outcomes, refer to Figure 2 and Figure 3.

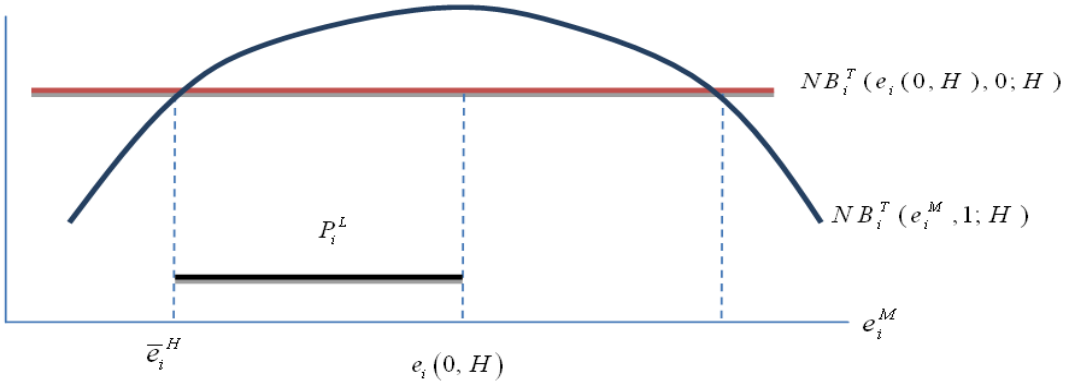

Fig. 2. The set of outcomes that satisfies condition P1

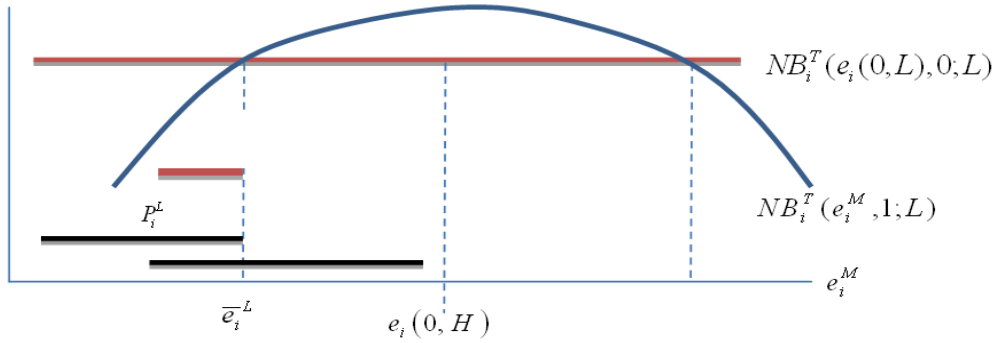

Fig. 3. The set of separating equilibrium outcomes

Not that $NB_i^T\left(e_i\left(0,\theta\right),0;\theta\right)$ has a fixed positive value. Note furhter that

$$NB_i^T\left(e_i\left(0,\theta\right),1;\theta\right) > NB_i^T\left(e_i\left(0,\theta\right),0;\theta\right),$$

which essentially tells that P^H is non-empty. As e_i^M is decreased, $NB_i^T\left(e_i^M,1;L\right)$ decresses, until it gets lower than $NB_i^T\left(e_i\left(0,\theta\right),0;\theta\right)$.

Now define the two emission level that lower and upper boundary of the set P_i^H and P_i^L, respectively:

$$\overline{e}_i^H = \arg\left\{ e_i^M < e(0,H) \mid NB_i^T\left(e_i^M,1;H\right) = NB_i^T\left(e_i\left(0,H\right),0;H\right) \right\},$$

$$\overline{e}_i^L = \arg\left\{ e_i^M < e(0,H) \mid NB_i^T\left(e_i^M,1;L\right) = NB_i^T\left(e_i\left(0,L\right),0;L\right) \right\}.$$

A separating equilibrium is therefore guaranteed, as long as $\overline{e}_i^H < \overline{e}_i^L$, as shown in Figure 3. And the condition for this is the SCP.

The reason why it is worthwhile for a high damage cost type to separate, is that a reduction of emission yields a higher net-benefit. This comes around because the high damage type values reductions in emissions more than the low damage costs type. Therefore, the high damage costs type is willing to reduce emission more than the low damage cost type, and therefore the high damage costs type will be able to reduce emission to a level (still inside P_i^H), where it is revealed that this emission can only be played by an high damage costs type, because such emission level are outside P_i^L: Even if the low damage cost type will get the best possible beliefs, it is not worthwhile for a low damage costs type to play.

To summarize, the reason why the high damage costs wants to differentiate is that without differentiation, it will be recognized as a low damage costs type, due to the assumption on out of equilibrium beliefs. The SCP is the condition that makes it less costly for the high damage cost type to reduce emissions than for the low damage cost type. Moreover, for the same emissions level, getting a higher p (e) always increases net benefit for this type, therefore, as shown in Figure 1 and 2, a set of emissions exists, where it is profitable to being recognized as having high damage cost. Taking these observations together, separating equilibrium outcomes exit where the high damage cost type reduces its emission below e (0, H) while the low damage cost type does not reduce its emission.

6. 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[1]. 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. We present the definitions with respect to private information about damage costs.

Definition of a weakly dominated (WD) strategy: A strategy e_i^1 is WD by another strategy e_i^2 for type θ, if $\min_\rho NB_i^T\left(e_i^2, \rho_i; \theta\right)) \geq \max_\rho NB^T(e_i^1, \rho, \theta)$.

It appears from the above definition that for e_i^1, to be weakly dominated by e_i^2, 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. Given the out of equilibrium beliefs, it follows that in our setting e_i^1 is weakly dominated by strategy e_i^2 if $NB^T\left(e_i^2, 0; \theta\right)) \geq NB^T(e_i^1, 1, \theta)$.

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, say type θ^l but not for the other type, then the uninformed players' belief should place zero probability that θ^l has sent e_i, i.e. e_i must be followed by posterior beliefs p ($\theta^l | e_i$) = 0.

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

Proposition 2: Given $P_i^L \cap P_i^H$ is non-empty one undominated separating equilibrium exists:

$\hat{e}_i(L) = e_i(0, L)$,

$\hat{e}_i(H) = \overline{e}_i^L$.

Proof see in Appendix.

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

Proposition 2 implies that a rational optimizing country with high damage should use the minimum resources necessary to distinguish itself from the shadow of the low damage cost type. Refer to Figure 3, where the unique non-dominated separating equilibrium is shown as the highest emission level in the set of separating equilibrium outcomes.

Unilateral action in such model is needed for the high damage cost type to reveal its type. Such a type is the one that needs to make a costly action to reveal its type. Therefore, unilateral action here is not motivated by good-hearted behavior and the desire to "set a good example" but a costly attempt for the one type to being recognized as a particular type. Whether or not this is possible are described by the conditions in proposition 1. Moreover, if these conditions are meet, a rational country would avoid overinvestinging in costly signalling, and proposition 2 shows the least costly credible signal. Finally, note that if signalling succeeds, the total emission level will decrease, and therefore our result is in contrast to the findings of Hoel (1992).

Conclusion

The story of this paper has been that countries meet to agree on a new treaty on reducing an international environmental problem. The treaty is such that the emissions targets depend on the damage costs of countries. Higher damage costs imply that all will accept more stringent reduction targets. Countries have, however, private information about damage costs, and therefore need to make a credible move to convince the other countries that damages are high. In the model, reducing pre-agreement emissions acts as such a signal. Such signals are interpreted as unilateral action to signal high damage costs. The analysis shows that under realistic conditions, and given the institutional setting and information structure specified in the paper, unilateral action is a credible strategy to signal high damage costs.

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.

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 behavior can be altered such that cooperation can be sustained, all the incentives must be identified.

References

1. Arredondo, A.E and F.M. García (2011). 'Free-riding in international environmental agreements: A signalling approach to non-enforceable treaties', *Journal of Theoretical Politics*, 23, pp. 111-134.
2. Bac, M. (1996). 'Incomplete information and incentives to freeride on international environmental resources', *Journal of Environmental Economics and Management*, 30, pp. 301-315.
3. Beccherle, J. and J. Tirole (2011). 'Regional initiatives and the cost of delaying binding climate change agreements', *Journal of Public Economics*, 95, pp. 1339-1348.
4. Barrett, S. (2003). *Environment and statecraft, the strategy of environmental treaty-making*, UK: Oxford University Press.
5. Brandt, U.S. (2002). 'Actions prior to entering an international environmental agreement', *Journal of Institutional and Theoretical Economics*, 158, pp. 695-714.
6. Brandt, U.S. (2004). 'Unilateral actions, the case of international environmental problems', *Resource and Energy Economics*, 26, pp. 373-391.
7. 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.
8. Cho, In-K. and D.M. Kreps (1987). 'Strategic stability and uniqueness in signalling games', *Quarterly Journal of Economic Theory*, 102, pp. 179-221.
9. Fudenberg, D. and J. Tirole (1993). *Game theory*, MIT Press, Cambridge, Massachusetts.

10. 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, pp. 115-135.
11. Harstad, B. (2009). 'The dynamics of climate agreements', Harvard Project on International Climate Agreements', Discussion Paper 09-28, Harvard Kennedy School.
12. Harstad, B (2011). 'The dynamics of climate agreements', Mimeo, Northwestern University.
13. Heitzig, J., K. Lessmannand 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 Science of the United States of America (PNAS)*,108, pp. 15739-15744.
14. Hoel, M. (1992). 'International environment conventions: The case of uniform reductions of emissions', *Environmental and Resource Economics*, 2, pp. 141-159.
15. 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.
16. Hulme, M. (2009). *Why we disagree about climate change: Understanding controversy, inaction and opportunity*, Cambridge, United Kingdom: Cambridge University Press.
17. 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.
18. IEA (2012). World Energy Outlook 2012.
19. 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.
20. Jakob, M. and K. Lessmann (2012). 'Signalling in international environmental agreements: The case of early and delayed action', *International Environmental Agreements: Politics, Law and Economics*, 12, pp. 309-325.
21. 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,
22. 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, pp. 337-363.
23. Stavins, R.N. (2011). 'The problem of the commons: Still unsettled after 100 years', *American Economic Review*, 101, pp. 81-108.
24. Stern, N. (2006). *Stern review on the economics of climate change*, Cambridge University Press.
25. 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.
26. Walsh, S., H. Tian, J. Whalley and M. Agarwal (2011). China and India's participation in global climate negotiations, *International Environmental Agreements*, 11, pp. 261-273.

Appendix

Proof of proposition 1. We need to proof that there exists an $e_i^M < e_i(0,H)$, such that

P1: $NB_i^T\left(e_i^M,1;H\right) \geq NB_i^T\left(e_i(0,H),0;H\right)$,

P2: $NB_i^T\left(e_i^M,1;L\right) < NB_i^T\left(e_i(0,L),0;L\right)$.

Rewrite P1 and P2:

P1 rewritten:

$$B_i\left(e_i^M;H\right)-B_i\left(e(0,H);H\right)+\delta\left[B_i\left(e_i^S(1);H\right)-B_i\left(e_i^S(0);H\right)\right]\geq D_i\left(e_i^M,e_{-i}^M;H\right)-D_i\left(e(0,H),e_{-i}^M;H\right)+$$
$$+\delta\left[D_i\left(e_i^S(1),e_{-i}^S(1);H\right)-D_i\left(e_i^S(0),e_{-i}^S(0);H\right)\right]. \quad \text{(P3)}$$

P2 rewritten:

$$B_i\left(e_i^M;L\right)-B_i\left(e(0,L);L\right)+\delta\left[B_i\left(e_i^S(1);L\right)-B_i\left(e_i^S(0);L\right)\right]<D_i\left(e_i^M,e_{-i}^M;L\right)-D_i\left(e(0,L),e_{-i}^M;L\right)+$$
$$+\delta\left[D_i\left(e_i^S(1),e_{-i}^S(1);L\right)-D_i\left(e_i^S(0),e_{-i}^S(0);L\right)\right]. \quad \text{(P4)}$$

Using that benefits are equal for the high damage and the low damage type for equal emissions level and given Assumption A1, this implies that the left hand side of P3 and P4 are identical, and can be combined as:

$$D_i\left(e_i^M, e_{-i}^M; L\right) - D_i\left(e(0,L), e_{-i}^M; L\right) + \delta\left[D_i\left(e_i^S(1), e_{-i}^S(1); L\right) - D_i\left(e_i^S(0), e_{-i}^S(0); L\right)\right] > D_i\left(e_i^M, e_{-i}^M; H\right) - D_i\left(e(0,H), e_{-i}^M; H\right) +$$
$$+ \delta\left[D_i\left(e_i^S(1), e_{-i}^S(1); H\right) - D_i\left(e_i^S(0), e_{-i}^S(0); H\right)\right].$$

Rewritten:

$$\delta\left[D_i\left(e_i^S(0), e_{-i}^S(0); H\right) - D_i\left(e_i^S(1), e_{-i}^S(1); H\right)\right] - \delta\left[D_i\left(e_i^S(0), e_{-i}^S(0); L\right) - D_i\left(e_i^S(1), e_{-i}^S(1); L\right)\right] >$$
$$> \left[D_i\left(e_i^M), e_{-i}^M; H\right) - D_i\left(e(0,H), e_{-i}^M; H\right)\right] - \left[D_i\left(e_i^M, e_{-i}^M; L\right) - D_i\left(e(0,L), e_{-i}^M; L\right)\right].$$

This is a version of the generic single crossing property (SCP), adjusted to this context. Therefore, the SCP is a sufficient condition for the existence of a separating equilibrium.

Proof of proposition 2. First, for the L-type, via definition of E_i^L, all $e_i^M \notin E_i^L$ are weakly dominated by e_i (0, L). On the other hand, also via definition of E_i^H, none of $e_i^M \notin E_i^H$ are weakly dominated by e_i (0, H). Next fix any candidate equilibrium $\hat{e}_i(H) < e_i^L$, if the receiver observes a $e_i^M = \hat{e}_i(H) + \varepsilon$, posterior beliefs should be updated to $\rho_i\left(e_i^M\right) = 1$, and consequently, $\hat{e}_i(H)$ is no longer a sequential equilibrium. This can be done for all $\hat{e}_i(H) < \overline{e}_i^L$. The only non-dominated sequential equilibrium is $\hat{e}_i(H) < \overline{e}_i^L$.