Signaling in International Environmental Agreements: The Case of Early and Delayed Action

Michael Jakob[1], Kai Lessmann

October 2011

Potsdam Institute for Climate Impact Research (PIK)

Abstract

This paper presents a stylized IEA game with two regions differing in their preference for environmental quality. If side payments are allowed, cooperation can increase the payoffs accruing to both regions. However, cooperation can be impeded by asymmetric information about the regions' types and only become feasible once a region has credibly revealed its type. We show how in a two-stage game, as well as in continuous time, early (delayed) action can act as a credible signal to reveal private information on high (low) benefits. Yet, the cooperative solution with asymmetric information is Pareto-dominated by the outcome with perfect information.

Keywords: International Environmental Agreements, Climate Policy, Asymmetric Information, Signaling Game

[1] Corresponding author: jakob@pik-potsdam.de, Tel. ++49-331 288 2650, P.O.Box 601203, 14412 Potsdam, Germany, www.pik-potsdam.de

1. Introduction

Several recent studies based on integrated assessment modeling find that early action can significantly reduce the costs of mitigating climate chance (Edmonds et al. 2008, van Vliet et al. 2009, Clarke et al. 2009). Furthermore, it has been pointed out that delaying the inception of a global climate agreement can influence strategic behavior, as countries face incentives to lower their investments in abatement technologies to improve their future bargaining position (Harstad 2009, Beccherle and Tirole 2010). In spite of these arguments in favor of early action, negotiations on a climate agreement to replace the Kyoto protocol after its expiry in 2012 suffered a serious set-back at COP-15 in Copenhagen. Nonetheless, the EU has maintained its target to unilaterally reduce GHG emissions by 20% to 30% below 1990 levels by 2020.

It has been argued that this pattern of action by only a small number of countries but inaction by the large majority can be explained by free-riding behavior, which allows for only a low number of active participants in international environmental agreements (IEAs) (e.g. Cararro and Siniscalco 1993, Barrett 1994). However, several recent contributions have identified schemes that can bring the level of climate protection close to the global optimum by matching regions that exhibit the highest willingness to pay for abatement of GHG emissions with regions that feature the lowest mitigation costs (Carbone et al. 2009, Weikard et al. 2006, Carraro et al. 2006). Yet, despite the possibility of employing optimal transfer schemes in practice, currently there seems little prospect for a global agreement to emerge in the near future. It has repeatedly been argued that uncertainty arising from asymmetric information can provide a plausible explanation of the observed outcome (Afionis 2011). This paper demonstrates how mutually desirable cooperation can be impeded by imperfect information and identifies constellation in which early or delayed abatement of emissions can be employed as a signal that credibly reveals private information to render cooperation possible.

In a setting in which parameters relevant for policy formulation (such as citizens' concern for the environment and future generations, which affect the benefits of climate change mitigation) cannot be verified by outsiders, informational asymmetries can crucially shape strategic interactions. While the insights provided by the economics of information have revolutionized numerous branches of economics (Stiglitz 2000), the issue of informational asymmetries does not feature prominently in the context of IEAs. The large majority of contributions in this area focus on optimal contracts designed to ensure that truthful revelation of one's type by means of self-selection is a dominant strategy. For instance, Matsueda (2004) shows that if a pollutee has imperfect information with regard to a polluter's environmental preference, incentive conflicts can arise that make it impossible to achieve an agreement to mitigate trans-boundary pollution, and Batabyal (2000) demonstrates that if an IEA is unable to observe firms' private information, it can be hampered by collusion between national governments and firms. Several optimal second-best contracts have been proposed to deal with problems related to asymmetric information. Helm and Wirl (2009) identify optimal transfer payment schemes subject to bargaining in North-South climate negotiations in the face of private information. Laffont and Martimont (2005) focus on pricing strategies in games with multiple hierarchies in which a principal pays an agent for the provision of a transnational public good, and Mason (2010) proposes a solution to reduce CO_2 emissions from deforestation that minimizes its budgetary impact by means of a menu of two-part contracts (consisting of lump-sum plus variable payments) from which forest owners can choose.

Likewise, while the signaling function of international agreements has been widely recognized, for instance in the context of disarmament (Fearon 1998), regional integration (Schneider and Cederman 1994), and human rights (Hollyer and Rosendorff 2011)[2], it has received only limited attention in the literature on IEAs. The models by Harstad and Eskeland (2010), in which firms over-purchase emission permits to signal high mitigation costs to the regulator (and receive higher allocation of free permits in the next period), and Denicolò (2008), in which firms engage in environmental over-compliance in order to induce stricter regulation and raise their rivals' costs by more than their own could very likely also be applied to the interaction between national governments and an IEA's supra-national authority. All these contributions assume an already existing environmental agreement, while we

[2] See Walsh (2007) for a review of signaling applied to the theory of international relations.

analyze how signaling can contribute to the conclusion of such an agreement. Bac (1996) presents a model in this spirit, in which the underlying abatement game can either constitute a Prisoners' dilemma, or a Chicken game. With asymmetric information, a so-called 'war of attrition' can arise, in which countries with a high preference for environmental quality forego abatement in order to conceal their type. By contrast, we study constellations in which the underlying game can either be a Prisoners' dilemma, or a game of cooperation. In the latter case, countries have an incentive to truthfully reveal their type instead of concealing it. We show how early or delayed abatement, respectively, can then act as a credible signal of a player's true type. Brandt (2004) develops a model in which abatement costs are uncertain and positively correlated across countries. A country that becomes privately informed about its particular cost structure can then engage in unilateral early action to reveal a low overall level of abatement costs. Brandt's model significantly differs from ours, as in the former early action is desirable in order to reveal information to other countries regarding *their own* mitigation costs (and trigger additional abatement), while in our model the motivation to engage in early or delayed action, respectively, is to reveal private information regarding *the sender's* type. Caparrós et al. (2004) present a model of North-South negotiations in which North can enact transfer payments for the provision of a global public good by South, whose cost-benefit ratio is private information. One of this paper's central results is that refusing to sign an IEA can act as a signal of South's true type, such that delay may arise before an IEA emerges. Our model departs from this contribution in two important aspects: firstly, instead of exclusively focusing on uncertainty regarding South's costs or benefits, we also show how early action by the North can be employed as a signal. Secondly, we introduce the possibility that countries 'cheat' by entering an IEA without fulfilling their commitments. Hence, in our model, actual abatement instead of signing an agreement serves as the signal.

Finally, Espinola-Arredondo and Munoz-Garcia (2011) discuss how the (non-binding) signature of an IEA can act as a signal of private information on benefits if it entails political costs (i.e. reputation effects). In our model, by contrast, the timing of actual abatement acts as a signal (instead of the mere signature of an IEA), without any need to introduce ad-hoc political costs. As it provides a plausible explanation why countries have strategic incentives for early or delayed action on climate change mitigation, our paper fills an important gap in the previous literature, which is particularly relevant in the light of the current political discussion.

In the following, this research article presents a stylized model in which asymmetric information can preclude the successful conclusion of an IEA that would be stable under perfect information. We propose a game structure (in a two-stage game as well as in continuous time) appropriate to demonstrate how unilateral early (delayed) action can be employed to credibly signal high (low) benefits from mitigating climate change. We also derive conditions necessary for the existence of a separating equilibrium - in which truthful revelation of private information constitutes a dominant strategy - and show that in terms of welfare the cooperative outcome under asymmetric information is strictly inferior to the outcome obtained with perfect information.

This paper proceeds as follows: Section 2 presents a simple model of provision of a global public good with complete information, Section 3 shows how asymmetric information can inhibit cooperation, Section 4 discusses in which way the timing of action can be employed as a signal to reveal private information. Section 6 discusses the results in the light of the current policy debate and draws conclusions for further research.

2. Cooperation in International Environmental Agreements with Complete Information

In order to highlight the economic mechanism through which asymmetric information enters a game of global public good provision, we adopt a simple model which divides the world in two regions, 'North' and 'South'[3]. The two regions can contribute to the global public good at different local costs and benefits and will do so to some extent even when the regions act independently. We will normalize the payoffs for

[3] Formally this model is similar to Caparrós et al. (2004), where informational advantages can improve a region's bargaining position. In our paper, however, it is desirable to truthfully reveal private information which can only be achieved through a signal, leaving both parties worse off than in the case with complete information.

this case to zero for both regions. The regions may be able improve upon this situation by agreeing to have one region finance an additional contribution by the other region. Let the costs of such additional contributions by North or South be c_N and c_S, respectively, and the corresponding benefits b_N and b_S[4]. While admittedly stylized, the assumption that the provision of the public good predominantly takes place in one region seems plausible, e.g. for the case of climate change. As carbon-intensive infrastructure is 'locked in' over several decades, abatement of carbon emission below projected business-as-usual levels might – at least in the short to mid-term – to a large part take place in developing countries. Avoiding emissions by building up low-carbon energy infrastructures in developing countries (where future energy demand can be expected to grow more rapidly) may be more cost-efficient than scrapping existing installations in developed countries. Another example concerns reducing carbon emissions from deforestation and forest degradation (REDD), which offers ample low-cost mitigation options concentrated in developing countries[5].

The case of interest is the one in which each region's costs are higher than its individual benefits (i.e. $c_N > b_N$ and $c_S > b_S$) but in which mutual cooperation can bring about a Pareto-superior outcome (i.e. $b_N + b_S > c_S$ or $b_N + b_S > c_N$, respectively). As is common in public good provision, this setting causes underprovision of the good. However, in this situation side-payments can be used to alter the game's pay-off structure such that cooperation becomes desirable for both regions.

Let both players simultaneously choose their (pure) strategies[6] in a one-shot game from the set 'cooperation' or 'non-cooperation'. In this context, cooperation for the North means effecting a transfer payment T to the South, while for the South cooperation is understood as providing the public good. If both players choose not to cooperate, they receive zero payoffs. If any player chooses non-cooperation but the other cooperates (let's call this behavior 'cheating') either (a) South supplies the

[4] The same pay-off structure arises from a model with linear cost and benefit functions (as e.g. in Barrett 1994 or Finus and Rübbelke 2008), which results in a binary choice whether abatement is performed or not (i.e. a 'bang-bang' solution).

[5] For instance, Kindermann et al. (2008) point out that more than 2.000 $MtCO_2$ can be avoided annually at a carbon price not exceeding US$ 20.

[6] While mixed strategies are usually employed in repeated games, we argue that the relevant case of an international climate agreement has the character of a one-shot game. Hence, we restrict our analysis to pure strategies.

public good without receiving the promised payment, or (b) North pays the transfer without South engaging in the provision of the public good. Assume that a system of measurement, reporting, and verification (MRV), whose quality is inversely related to the parameter γ ($\gamma < 1$), exists. For instance, γ can be understood as the probability that cheating can occur without being noticed. By means of MRV cheating by any player can be detected before the end of the game. In this case, both players continue to behave non-cooperatively until the end of the game and the costs and benefits that occur are only a fraction γ of what they would be if cooperation were upheld over the entire game[7] (Figure 1).

<< Figure 1 about here >>

Whether the cooperative outcome constitutes a Nash-equilibrium depends on the relative payoffs associated with cooperation and cheating: if North chooses to cooperate, South's best response is cooperation if $b_S - c_S + T > \gamma T$, i.e. if North offers a transfer exceeding the minimum transfer $T^{\min}$

$$T > \frac{c_S - b_S}{(1-\gamma)} \equiv T^{\min}. \tag{1}$$

Likewise, if South cooperates, North prefers cooperation over non-cooperation if $b_N - T > \gamma b_N$, i.e. if the transfers due are not overly expensive.

$$T < (1-\gamma) b_N \equiv T^{\max}. \tag{2}$$

Let us at as a shorthand for later use introduce a (exogenously given) surplus sharing scheme characterized by a parameter α ($0 < \alpha < 1$) to determine which value of the transfer between the minimal amount that South is willing to accept and the maximum value North is prepared to provide will actually be realized, provided that cooperation is feasible (i.e. Eq. 1 and 2 both hold):

[7] Hence, γ is a conversion factor between those costs/benefits that occur over the entire game and those that are terminated after non-cooperation is detected

$$T = \alpha T^{\min} + (1-\alpha)T^{\max} . \tag{3}$$

Combining the expressions for $T^{\min}$ and $T^{\max}$ yields the condition for the existence of transfers that sustain a cooperative equilibrium:

$$b_N > \frac{c_S - b_S}{(1-\gamma)^2} . \tag{4}$$

The above inequality is satisfied for (i) low net costs $c_S - b_S$ of providing the public good in the South, (ii) high benefits b_N in the North, and (iii) a low value of γ, which mitigates incentives for both players to free-ride on the other player's cooperative behavior. If cheating is more likely to be detected (i.e. γ is low), South's rewards to free-riding decline, and she will be prepared to settle for a lower transfer (Eq. 1). A lower γ will also increase the transfer payment North is willing to provide, as the benefits of enjoying abatement by South while cheating are reduced (Eq. 2). Both these effects, i.e. a lower $T^{\min}$ and a higher $T^{\max}$, increase the scope for cooperation.

If condition 4 is fulfilled, the game is no longer a Prisoners' Dilemma (in which non-cooperation is a dominant strategy for each player) but becomes a game of coordination (in which mutual cooperation constitutes a second Nash-equilibrium that Pareto-dominates the non-cooperative outcome; see, for example, Schelling 1960)[8]. For the purpose of this paper, we posit that whenever mutual cooperation constitutes a Nash-equilibrium, the cooperative strategies are the "obvious way" to play the game and hence we will consider cooperation emerges as the "focal equilibrium" of the coordination game. This in fact does not require any kind of commitment device; it is sufficient that the players can communicate without incurring costs - once they agree to coordinate their actions, no player has an incentive to unilaterally switch to non-cooperation.

3. Asymmetric Information as an Obstacle to Cooperation

[8] We do not consider the situation of a Chicken game in our analysis (cf. Pittel and Rübbelke 2011)

Having discussed the conditions for cooperation with perfect information, we now turn to the case with asymmetric information about the other country's benefits. Own benefits are known to each region with certainty, but there can be uncertainty with regard to the benefits of the other region.[9] Asymmetric information on the country level are commonly assumed in the respective literature (e.g. Batabyal 2000, Matsueda 2004, Caparrós et al., 2004). This assumption is far from trivial, as it implies that the valuations and decision procedures of one country's government are only known with uncertainty to others. As this paper focuses on the question of how asymmetric information can impede cooperation and how signaling can be employed to achieve Pareto-improvements, we take asymmetric information as a given. Yet, the precise mechanism how these informational can asymmetries arise is an issue that deserves further attention in future research.

To keep the analysis tractable, we limit the discussion to settings in which the benefits of one region are known with certainty to both regions, while the other region's benefits can take on one of two possible values, either *high* or *low*. A region whose true benefits are high is referred to as being of the *h-type*, conversely *l-type* regions have low benefits. From the perspective of South, if the type of North is unknown, it can take on either of the two discrete values b_N^h or b_N^l with probabilities p_N and $1-p_N$, respectively.[10] Likewise, North believes South to be an h-type with benefits b_S^h with a probability of p_S, or an l-type with benefits b_S^l with probability $(1-p_S)$ if her true benefits are uncertain.

<< Figure 2 about here >>

Figure 2 formalizes the case in which North's true benefits are unknown to South as a one-stage game in extended form. In stage 0, Nature moves, pinning down whether North is an h-type or l-type. Then, North and South move simultaneously in stage 1. For this game, we consider a Bayesian Nash Equilibrium where, in essence,

[9] Benefits arguably constitute the most relevant source of asymmetric information, as they represent not only physical and economic damages, but also a willingness to pay for climate change mitigation. This also depends on normative parameters, such as intergenerational justice and concern for the environment (Gardiner, 2004).

[10] While only the players know their true type, we assume that all probabilities are common knowledge

equilibrium strategies maximize each player's *expected* utility given the other players' strategies. Expectations are formed over the *belief* about the other players' types, i.e. the probability distribution over the types. When benefits of North are uncertain to South, the existence condition for a cooperative solution from the previous section (Eq. 4) translates into a threshold where cooperation (non-cooperation) is North's best response to cooperation by South if b_N is above (below) the threshold. Otherwise, the realization of the (uncertain) b_N does not matter for North's choice of strategy and the problem becomes formally equivalent to the one with full information discussed in Section 2. Hence, we suppose that whether North's benefits are high or low makes a difference for his optimal strategy, i.e.

$$b_N^l < \frac{c_S - b_S}{(1-\gamma)^2} < b_N^h \,. \qquad (4')$$

Notice that the threshold depends on the benefit parameter of the North and therefore on its type. Under these conditions, what is the impact of uncertainty on the strategy of South?

<u>Lemma 1</u>: If b_N is uncertain and North's action depends on his type (i.e. if the condition of Eq. (4') holds), the cooperative outcome is only obtained if North is an h-type and South expects with a sufficiently high probability that North is an h-type.

<u>Proof</u>: From the previous section and the definition of high and low benefits we know the equilibria of the second stage: if North expects South to cooperate, he too cooperates if he is an h-type, but plays non-cooperatively if he is an l-type. If, however, North expects non-cooperation by South, he chooses non-cooperation regardless of his type. p_N being public knowledge allows North to anticipate South's action. He expects cooperation by South (and, if he is an h-type, North cooperates, too) if the latter's expected payoff from cooperation exceeds her expected payoff from non-cooperation[11]. Hence, South's expected payoff from cooperation is:

[11] That is, mutual cooperation constitutes a Nash equilibrium that is Pareto-superior to the non-cooperative outcome.

$$\pi^e_{S,c} = p_N(b_S - c_S + T) + (1 - p_N)\gamma(b_S - c_S), \tag{5}$$

while the expected payoff of playing non-cooperatively is

$$\pi^e_{S,nc} = p_N \gamma T. \tag{6}$$

South chooses the pure strategy that yields the highest expected payoff. Solving for p_N then yields the following condition for cooperation (i.e. $\pi^e_{S,c} > \pi^e_{S,nc}$):

$$p_N > \frac{\gamma(c_S - b_S)}{(1-\gamma)(b_S - c_S + T)}. \tag{7}$$

Hence, South's expected benefit only warrants cooperation if the probability of North being an h-type is sufficiently high. This means that South needs to be adequately optimistic that playing cooperatively will be met by cooperation by North (and hence pay off). However, if North is an h-type but South assigns too low a probability to this state, no cooperation (which would be beneficial for both regions) can occur, as South has no means of verifying North's true type. □

Now consider the effect of incomplete information from the perspective of North: if North was inclined to cooperate, he would offer a transfer payment *T*. For low or high benefits in South, the volumes of the transfer payment are high or low, respectively (Eq. 1-3)[12]. This leaves North with three choices: either offering transfers T^- or T^+ corresponding to South being an h- or l-type, respectively ($T^- < T^+$), or no transfers (i.e. playing non-cooperatively) (*n*). Figure 3 shows the game structure and payoffs for this situation. If North agrees to pay the (higher) transfer T^+, South's incentive compatibility condition Eq.(1) is fulfilled for h- as well as l-types and cooperation occurs, regardless of South's true type. If, on the other hand, T^- (the transfer that corresponds to an h-type) is sufficient to induce cooperation even if South is an l-type (i.e. if $T^- > \frac{c_S - b_S^l}{(1-\gamma)}$), North obviously has no incentive to offer a

[12] That is, the transfer North has to offer to make cooperation worthwhile is the higher the lower South's benefits b_S.

higher transfer. In this case, the cooperative equilibrium is achieved with North offering T^-, provided that North's benefits are sufficiently high such that mutual cooperation yields a higher payoff than cheating (cf. Eq. 4). Otherwise, in the case in which North never cooperates (i.e. Eq. 4 holds for neither realization of b_N), South anticipates this strategy and the game trivially results in the non-cooperative equilibrium. The case of interest is hence the one in which South's reaction to North's transfer offer depends on her type. Restating Eq.(1), this can be expressed as:

$$\frac{c_S - b_S^l}{(1-\gamma)} > T^- > \frac{c_S - b_S^h}{(1-\gamma)} \qquad (1')$$

Again, if Eq.(1') does not hold, South's reaction to North's transfer offer does not depend on its type and the game can be expressed in the simple form discussed in Section 2.

<< Figure 3 about here >>

<u>Lemma 2</u>: *If b_S is uncertain and South's reaction to the transfer offered by North depends on her type (i.e. Eq.(1') holds), the cooperative outcome is only obtained if (a) either South is an h-type, or if (b) North expects with a sufficiently high probability that South is an l-type.*

<u>Proof</u>: We only have to consider constellations in which North prefers mutual cooperation over free-riding, i.e. b_N^h is high enough (cf. Eq. 4). With perfect information, North would then agree to pay a transfer T that is a function of South's benefits b_S. Under incomplete information, however, this transfer depends on South's type since T is a function of South's benefits[13]. South then has an incentive to pass as an l-type in order to extract a higher transfer payment, even if she is an h-type. North in turn offers the transfer payment which results in the highest expected payoff. For the transfer corresponding to South being an l-type (i.e. T^+), cooperation always occurs, yielding the certain payoff:

[13] Note that a transfer is always offered by North, as the (higher) transfer T^+ guarantees cooperation, which is Pareto-superior to free-riding for the cases considered here.

$$\pi^{+}_{N,c} = b_N - T^{+} . \tag{8}$$

Given the conditions stated in the Lemma, the expected payoff for offering T^{-} is given by:

$$\pi^{-,e}_{N,c} = p_S (b_N - T^{-}) + (1 - p_S)(-\gamma T^{-}) , \tag{9}$$

That is, North derives a net payoff of $b_N - T^{-}$ if South is an h-type (in which case mutual cooperation occurs), but $-\gamma T^{-}$ if South is an l-type (and hence does not cooperate). North then opts for T^{+} if $\pi^{+}_{N,c} > \pi^{-,e}_{N,c}$, which is satisfied if

$$p_S < \frac{b_N - T^{+} + \gamma T^{-}}{b_N - (1-\gamma)T^{-}} . \square \tag{10}$$

Assuming that Eq.(1') holds, a transfer T^{-} is sufficient to ensure mutual cooperation if South is an h-type; for an l-type, however, non-cooperation would yield a higher payoff[14]. Hence, if South is an l-type but North assigns a sufficiently high probability p_S to her being an h-type, the game results in a situation in which North is willing to cooperate but only offers T^{-}, such that South defects (even though cooperation with the higher transfer payment T^{+} would constitute a Pareto-superior outcome).

In summary, cooperation can be impeded by uncertainty if the actual realization of the uncertain parameters would mandate cooperation but the player being confronted with uncertainty expects that such a realization is too unlikely to mandate cooperation. For a brief illustration, we use a simple numerical example with plausible parameter values, assuming that the net present value of the future damages from unmitigated climate change range from 4% to 8% of global GDP (Stern 2007 reports estimates between 5% and 20%), distributed at a ratio of 3:1 between North and South and a surplus-sharing parameter of $\alpha = 2/3$ (Table 1). Let us assume that North's true

[14] Due to the game's information structure, South is able to anticipate North's action without uncertainty.

benefits are b_N^h, and South's are b_S^l. With perfect information, the minimum transfer for which South is prepared to cooperate is $T_{\min} = \frac{c_S - b_S}{(1-\gamma)} = 2.5\%$, and the maximum North is ready to pay amounts to $T_{\max} = (1-\gamma)b_N = 4.5\%$, such that the cooperative equilibrium is achieved with a transfer payment of:

$$T = \alpha T_{\min} + (1-\alpha)T_{\max} = \frac{2}{3} \cdot 2.5\% + \frac{1}{3} \cdot 4.5\% \approx 3.27\% .$$

With asymmetric information, however, for South the expected pay-off from cooperation only exceeds the payoff from non-cooperation if she expects North to be a high-benefit type with a probability of at least:

$$p_N > \frac{\gamma(c_S - b_S)}{(1-\gamma)(b_S - c_S + T)} = \frac{0.2 \cdot 2\%}{0.8 \cdot (3.27\% - 2\%)} \approx 39\% .$$

Likewise, for North it is only worthwhile to offer the transfer T^+ if it is sufficiently certain that South is a low-benefit type:

$$p_S < \frac{b_N - T^+ + \gamma T^-}{b_N - (1-\gamma)T^+} = \frac{6\% - 3.27\% + 0.2 \cdot 2.43\%}{6\% - 0.8 \cdot 2.43\%} \approx 79\% .$$

In this case T^+ is indeed necessary to bring about cooperation. Otherwise, North offers the lower transfer T^-, for which South's best response is non-cooperation.

<< Table 1 about here >>

4. Signaling as a Mechanism to Overcome the Information Problem

The preceding section has demonstrated that in certain settings where cooperation would be mutually advantageous (and an equilibrium outcome under perfect information) asymmetric information can thwart cooperation. As discussed in Lemmas 1 and 2, this outcome prevails if region A's best response to cooperation by B depends on A's type, but B is too pessimistic that A is of the cooperative type. This means that either (i) North is of the high-benefit type, but South thinks that his benefits are likely to be low, or (ii) South is a low-benefit type, but North assigns a high probability to her having high benefits. In both constellations, the player with private information would benefit from truthfully revealing his actual type. In this section we (a) demonstrate that early or delayed action can act as a signal[15] of a player's true type, and (b) identify conditions under which separating equilibria (in which the truthful revelation of one's type is a dominant strategy) are feasible.

For this reason, we extend our setting to a signaling game, in which one player's type is known with certainty, while the other one's is uncertain. As before, the game begins by Nature resolving this uncertainty to the corresponding player. This player may then take action that may or may not act as a signal of her type, before the game continues with the final stage as before. Benefits and expectations are as described above, such that cooperation would arise with perfect information but cannot occur with incomplete information (see previous section). However, a player's action in the signaling stage can reveal his type and make cooperation in the second stage possible. This signal – below we show that for North it consists of early unilateral abatement, and of refusing a transfer payment and delaying abatement for South – has to be incentive compatible, such that choosing the action that yields the highest payoff reveals a player's true type. In this case the outcome is characterized as a 'separating equilibrium'. Otherwise a 'pooling equilibrium', in which a player's type cannot be determined from their action in the signaling stage, results. Therefore, if North has credibly revealed in the first stage that he is of high-benefit type, South chooses to cooperate in the second stage; likewise, if South has credibly established that she is an l-type in the first stage, North is prepared to provide the appropriate transfer and the cooperative outcome is obtained in the second stage. Yet, as cooperation from the first

[15] In this context a signal that credibly conveys private information not directly observable to the counterparty is an action that is worthwhile pursuing for one type of player but not for the other one (Spence 1973).

stage on would have yielded a Pareto-superior result, the signaling outcome is not optimal from a social welfare perspective[16].

Solutions of this game are described by the concept of a Perfect Bayesian Equilibrium (PBE), which is most commonly used equilibrium for signaling games (Hovi and Areklett 2004). Essentially, a PBE is an equilibrium in strategies and beliefs, such that strategies maximize expected utility at the information sets along this strategy, given (a) the belief about the other players' types, and (b) the other players' strategies. Beliefs are given by probability distributions over all possible types, determined according to Bayes's rule.[17]

To be able to compare costs between periods, we further need to introduce a conversion factor $\delta_{N,S}$ to take into account the effects of discounting the future and the possibility that the game's two stages are of different length (in this case costs and benefits should be regarded as flow variables and $\delta_{N,S}$ can have values lower or higher than one).

<u>Proposition 1</u>: *If cooperation is impeded by asymmetric information, early action by North is a credible signal of high benefits if (a) North's h-type benefits are large, (b) North's costs as well as T are low, or (c) if the discount factor is sufficiently large.*

<u>Proof</u>: Payoffs for both players and the beliefs of South are shown in Figure 4. If North has credibly established that it is of the high-benefit type, updating South's belief about North's type entails $q_E = 1$, i.e. South is certain about North's type and hence will act cooperatively in the following stage.[18] North's credibility hinges on the existence of a separating strategy, hence we need to establish whether and when a separating equilibrium exists. Otherwise North's signal remains vacuous, and South's belief would remain unchanged, i.e. $q_E = p_N$. By our assumptions, cooperation would then fail.

---

[16] That is, acquiring a credible signal entails social costs (that would be avoided under perfect information) and the first best outcome cannot be attained.

[17] The Perfect Bayesian Equilibrium was introduced by Fudenberg and Tirole (1991). For an accessible introduction, see Gibbons (1992).

[18] According to Bayes's rule and given the prior $\mu(\text{"North is h-type"}) = p_N$, we have $q_E = \mu(\text{"North is h-type"} \mid \text{"early action"}) = p_N / p_N = 1$

Taking early action, North incurs negative net benefits $b_N - c_N$ in the first period, which are rewarded by positive ones corresponding to a present value of $\delta_N(b_N - T)$ if cooperation occurs, but $\delta_N \gamma b_N$ if he cheats in the second period, free-riding on South's mitigation effort. A separating equilibrium then exists if for an h-type the benefits of cooperation in the second stage exceed the costs borne in the first stage, but for an l-type the rewards of free-riding do not.

First, the incentive compatibility condition ensuring that it does not pay off for North to pretend to be an h-type by taking early action if in reality he is an l-type reads:

$$(b_N^l - c_N) + \delta_N \gamma b_N^l < 0, \tag{11}$$

see nodes #1 and #2 in Figure 4.

That is, the rewards of free-riding in the second stage are not sufficient to compensate for the costs of sending the signal in the first stage[19]. This condition of Eq.(11) can be rewritten as:

$$b_N^l < \frac{c_N}{1 + \gamma \delta_N}. \tag{11'}$$

Taking early action in order to be able to cheat in the second stage becomes less attractive for a Northern l-type, (a) the lower b_N^l (i.e. the benefits of enjoying the public good in the first period and of free-riding in the second one), (b) the larger c_N (i.e. the costs of sending the signal in the first stage), and (c) the lower γ and δ_N (which scale the expected net-benefits from free-riding).

[19] Without early action by North, mutual non-cooperation results in both stages of the game, yielding a payoff of zero.

Second, if North is an h-type, early action is only worthwhile if the sum of his net costs in the first stage and the benefits of cooperation in the second stage exceed the pay-off from non-cooperation during both periods, i.e.:

$$(b_N^h - c_N) + \delta_N (b_N^h - T) > 0\,, \qquad (12)$$

see nodes #3 and #4 in Figure 4.

Eq.(12) is equivalent to

$$b_N^h > \frac{c_N + \delta_N T}{1 + \delta_N}\,. \qquad (12')$$

A Northern h-type is thus more likely to successfully signal his type if (a) b_N^h is large (which results in lower net costs in the first stage as well as higher net benefits in the second one), if (b) c_N and T are low (lower costs to provide the signal in the first stage and to pay South for its provision of the public good in the second one), and if (c) δ_N is large (higher valuation of the benefits occurring in the second stage compared to first stage costs). □

Hence, a separating equilibrium for North exists if both conditions (11) and (12) are fulfilled, i.e. if b_N^l is sufficiently small and b_N^h sufficiently high. If one of them is violated, it either pays off for l-types to take early action but cheat in the second period, or it is not worthwhile for h-types to incur the extra costs of the signal in the first stage. In this case a pooling equilibrium, in which types cannot be distinguished by their first stage behavior (i.e both would choose identical actions) results. For instance, using the parameters employed in the numeric example (cf. Table 1), and assuming that $b_S = b_S^h = 2\%$ and $c_N = 8\%$, it is easy to verify that Eq.(11) is fulfilled if $\delta_N < 8.33$, and (Eq. 12) if $\delta_N > 0.73$, and a separating equilibrium exists if δ_N falls in the range $0.73 < \delta_N < 8.33$.

<u>Proposition 2</u>: If cooperation is impeded by asymmetric information, delay of action by South is a credible signal of low benefits if (a) South's l-type benefits are small, or (b) South's h-type benefits are large.

<u>Proof</u>: As we have shown in Lemma 2, cooperation can be impeded by asymmetric information if South is an l-type who only cooperates if offered the transfer T^+ (but not if offered T^-) but North assigns a high probability to South being an h-type. With this false expectation, North offers the transfer T^- in the first period. Using backward induction (Figure 5), we see that if South can credibly use non-cooperation in the first period to signal that she is a low-benefit type, North offers the transfer T^+ in the second period and a Pareto-improvement results from cooperation. Otherwise (i.e. if South cooperates in the first period), North's second period offer will remain T^-. Analoguous to Proposition 1, South's ability to affect North's belief depends on the existence of a separating equilibrium. Again, if a separating equilibrium exists, North will be certain about South's type according to Bayes's rule. Otherwise North's belief remains the prior probability distribution.

The second condition for a separating equilibrium is that low-benefit types must prefer playing non-cooperatively in the first period and cooperate in the second one with the transfer T^+ over cooperation in both periods with the lower transfer T^-:

$$(b_S^l - c_S + T^-) + \delta_S (b_S^l - c_S + T^-) < \gamma T^- + \delta_S (b_S^l - c_S + T^+), \tag{13}$$

see nodes #1 and #2 in Figure 5.

Eq.(13) can also be written as:

$$b_s^l - c_S + (1-\gamma)T^- < \delta_S (T^+ - T^-). \tag{13'}$$

Non-cooperation in the first period hence becomes the more attractive for l-types (a) the higher T^+, the lower T^- and the higher δ_S, as all three are directly related to the present value of the additional transfer ($T^+ - T^-$) in the second stage, (b) the lower

b_S^l, which lowers the opportunity costs of non-cooperation (i.e. foregone abatement) in the first stage, and (c) the higher γ and c_S, for which higher values increase the incentive to behave non-cooperatively in the first stage. The first condition for a separating equilibrium requires that if South is a high-benefit type, she prefers to cooperate and to accept the transfer T^- in both periods over the alternative of non-cooperation in the first stage and cooperation with transfer T^+ in the second stage:

$$(b_S^h - c_S + T^-) + \delta_S (b_S^h - c_S + T^-) > \gamma T^- + \delta_S (b_S^h - c_S + T^+), \qquad (14)$$

see nodes #3 and #4 in Figure 5.

Eq.(14) can also be expressed as:

$$b_s^h - c_S + (1-\gamma)T^- > \delta_S (T^+ - T^-). \qquad (14')$$

Cooperating in the first period (thus truthfully revealing her type) becomes the more attractive for an h-type (a) the larger T^-, the lower T^+, and the lower δ_S, which jointly determine the additional pay-off from the higher transfer ($T^+ - T^-$) in the second stage, (b) the larger b_S^h, which increases the opportunity costs of non-cooperation (i.e. foregone abatement) in the first stage, and (c) the lower c_S, as a lower values of this parameter decrease the incentive to behave non-cooperatively in the first stage[20]. □

Again, a separating equilibrium only exists if both conditions (13) and (14) are fulfilled, requiring b_S^l to be sufficiently small and b_S^h sufficiently large. Otherwise, a pooling equilibrium emerges. With the parameters of our numerical example, Eq.(13)

[20] Perhaps surprisingly, the impact of the parameter γ, which denotes the probability that cheating is detected, on the existence of a separating equilibrium is ambiguous for the case in which North's benefits are uncertain (Eq. 12') as well as the one in which there is uncertainty on South's costs (Eq. 13' and 14'). The reason for this finding is that a higher γ decreases the minimum transfer South is willing to accept, but increases the maximum transfer North is willing to provide. Hence, the effect on the transfer T (Eq. 3) depends on the specific parameters, and this ambiguity carries over to all incentive compatibility conditions that depend on T.

implies $\delta_S < 1.12$ and Eq.(14) $\delta_S > -0.067$. As δ_S is by definition non-negative, the condition for the existence of a separating equilibrium is $\delta_S < 1.12$.

<< Figure 4 about here >>

<< Figure 5 about here >>

5. Conclusions

This paper has argued that uncertainty concerning other regions' benefits of mitigating climate change, which can be considered private information, might play an important role in the current stalemate to achieve a global climate agreement. We have shown that there are indeed constellations in which signaling – i.e. truthfully revealing private information – can be welfare improving for both players. Sections 4 and 5 identify situations in which cooperation is mutually desirable but can only arise after a period of signaling activity and highlight that for North early action, and delaying action for South, can act as signals for high, respectively low, benefits. Therefore, it is conceivable that international cooperation on climate change might arise in the future once all players' benefits have been credibly established. As it is the case for any credible signal, early action by some regions and delayed action by others involves social costs. For climate change, these costs can be expected to be substantial, as delaying global action renders the most ambitious climate targets impossible to achieve and severely increases the costs of meeting intermediate stabilization targets.

The stylized model presented in this article suggests three conclusions that are directly relevant for policy: first, expectations about other regions' benefits from mitigating climate change are crucial for cooperation. Therefore, performing further research on regional climate change damages as well as achieving a shared understanding of these seems clearly mandated. Second, by setting up a system of monitoring and verification on a regular basis in short intervals, free-rider incentives can be reduced and cooperation be rendered more likely[21]. Third, applying appropriate incentive mechanisms derived from contract theory in international climate negotiations might offer an opportunity to circumvent some of the most serious problems related to informational asymmetries. These arguments underline that even without a 'world government' that enables countries to enter binding arrangements, appropriately designed institutions can play a crucial role to achieve cooperation by creating regimes that provide information and influence expectations (cf. Keohane and Martin 1995)

The stylized model presented in this contribution could be extended in several directions: perhaps most importantly, a multilateral framework could be considered, by introducing a setting in which South actually consists of two groups of countries with potentially different costs, as outlined in Caparrós et al. (2004). In such a framework, the outcome would crucially depend on the degree of complementarity or substitutability of abatement by South. In the former case (complementarity), we expect a result very similar to the one presented in this paper. As each group is aware that its cooperation is required to provide the public good, l-types would have an incentive to misrepresent their type in the first stage in order to extract higher transfer payments in the second stage. In the latter case (substitutability), however, a more complex outcome can be expected to emerge, in which the action of one group of Southern countries depends on its belief about the costs of the other group. An l-type country would then face a trade-off between extracting a higher second period transfer by misrepresenting its type, and the risk of not receiving a pay-off at all if the other group already accepts North's offer in the first period.

Further worthwhile extensions include examining the case with more than two countries could provide valuable insights on more complex strategic interactions (e.g. incentives to free-ride on other regions' provision of a signal), the inclusions of additional signaling devices, such as R&D, adaptation measures, or endogenous choice of abatement efforts and transfer payments. Another fruitful line of research might be the analysis of games in which all players are simultaneously confronted with uncertainty. Finally, we are convinced that examining the interplay of signaling

[21] Note that in the framework of our model, is the more likely to hold the smaller γ, i.e. the fraction of the payoff that can be appropriated with free-riding (cf. Eq.(4)).

motives with strategies to secure favorable bargaining positions in future negotiations à la Harstad (2009) and Beccherle and Tirole (2010) would make a significant contribution to the field.

References

Afionis, S. (2011): The European Union as a negotiator in the international climate change regime, International Environmental Agreements: Politics, Law and Economics, vol. 11(4), pp. 341-360

Bac, M. (1996): Incomplete Information and Incentives to Free Ride on International Environmental Resources, Journal of Environmental Economics and Management, vol. 30(3), pp. 301-315

Barrett, S. (1994): Self-Enforcing International Environmental Agreements, Oxford Economic Papers, New Series, vol. 46, Special Issue on Environmental Economics, pp. 878-894

Batabyal, A.A. (2000): On the Design of International Environmental Agreements for Identical and Heterogeneous Developing Countries, Oxford Economic Papers, vol. 52(3), pp. 560-83

Beccherle, J. and Tirole, J. (2010): Regional Initiatives and the Cost of Delaying Binding Climate Change Agreements, mimeo

Bliss, C., Nalebuff, B. (1984): Dragon-slaying and ballroom dancing: The private supply of a public good, Journal of Public Economics, vol. 25(1-2), pp. 1-12

Brandt, U.S. (2004): Unilateral actions, the case of international environmental problems, Resource and Energy Economics, vol. 26(4), pp. 373-391

Caparrós A., Péreau J.-C. and Tazdaït, T. (2004): North-South Climate Change Negotiations: A Sequential Game with Asymmetric Information, Public Choice, vol. 121(3), pp. 455-480

Carbone, J., Helm, C., and Rutherford, T.F. (2009): The case for international emission trade in the absence of cooperative climate policy, Journal of Environmental Economics and Management, vol. 58(3), pp. 266-280

Carraro, C., Eyckmans, J. and Finus, M. (2006): Optimal transfers and participation decisions in international environmental agreements, The Review of International Organizations, vol. 1(4), pp. 379-396

Carraro C. and D. Siniscalco (1993): Strategies for the International Protection of the Environment, Journal of Public Economics, vol. 52(3), pp. 309-328

Clarke, L., Edmonds, J., Krey, V., Richels, R., Rose, S., Tavoni, M. (2009): International climate policy architectures: Overview of the EMF 22 International Scenarios. Energy Economics, vol. 31, Supplement 2, pp. S64-S81.

Denicolò, V. (2008): A signaling model of environmental overcompliance, Journal of Economic Behavior & Organization, 68(1), pp. 293-303

EC (2010): Analysis of options to move beyond 20% greenhouse gas emission reductions and assessing the risk of carbon leakage, COM(2010) 265 final

Edmonds, J.; Clarke, L.; Lurz, J.; Wise, M. (2008): Stabilizing CO_2 concentrations with incomplete international cooperation, Climate Policy, vol. 8, pp. 355-376

Espinola-Arredondo, A. and Munoz-Garcia, F. (2011): Keeping Negotiations in the Dark: Environmental Agreements under Incomplete Information, Washington State University Working Paper

Fearon, J.D (1998): Bargaining, Enforcement, and International Cooperation, International Organization, vol. 52(2), pp. 269-305

Finus, M. and Rübbelke, D.T.G. (2008): Coalition Formation and the Ancillary Benefits of Climate Policy, Stirling Economics Discussion Papers 2008-13, University of Stirling, Department of Economics.

Fudenberg, D., Tirole, J. (1991): Perfect Bayesian Equilibrium and Sequential Equilibrium, Journal of Economic Theory, vol. 53, pp. 236-260.

Gardiner S. M. (2004): Ethics and Global Climate Change. Ethics 114 (3), pp. 555-598

Gibbons, R. (1992): Game theory for applied economists, Princton University Press

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

Harstad, B. and Eskeland, G.S. (2010): Trading for the Future: Signaling in Permit Markets, Journal of Public Economics, forthcoming

Helm, C. and Wirl, F. (2010): International Environmental Agreements: Incentive Contracts with Multilateral Externalities, mimeo

Hollyer, J.R. and Rosendorff, B.P. (2011) Why Do Authoritarian Regimes Sign the Convention Against Torture? Signaling, Domestic Politics and Non-Compliance, Quarterly Journal of Political Science, vol. 6(3–4), pp. 275-327.

Hovi, J. and I. Areklett (2004): Enforcing the Climate Change Regime: Game Theory and the Marrakesh Accords, International Environmental Agreements: Politics Law and Economics, vol 4, pp. 1-26.

IPCC (2007): Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson, Eds., Cambridge: Cambridge University Press, UK

Keohane, Robert O. and Lisa L. Martin (1995): The Promise of Institutionalist Theory, International Security, vol. 20(1), pp. 39-51.

Kindermann, G., Obersteiner, M., Sohngen, B., Sathaye, J., Andrasko, K., Rametsteiner, E., Schlamadinger, B., Wunder, S. and Beach, R. (2008): Global cost estimates of reducing carbon emissions through avoided deforestation. Proceedings of the National Academy of Sciences of the United States of America

Kreps, D.M. (1990): A Course in Microeconomic Theory, Princeton Univ Press.

Kydd, A. (2000): Trust, Reassurance, and Cooperation. International Organization, vol 54(2): 325-357

Laffont, J.J. and Martimort, D. (2005): The design of transnational public good mechanisms for developing countries, Journal of Public Economics, vol. 89(2-3), pp. 159-196

Mason, C. (2010): Carbon Sequestration Policy When Government Budgets Matter: Contracting Under Asymmetric Information, mimeo

Matsueda, N. (2004): Potential Failure of an International Environmental Agreement under Asymmetric Information, Economics Bulletin, vol. 17(4), pp. 1-8.

Nakada, M. (2006): Distributional Conflicts and the Timing of Environmental Policy, International Environmental Agreements: Politics, Law and Economics, vol. 6(1), pp. 29-38.

Pittel, K., and Rübbelke, D.T.G. (2011): Transitions in the negotiations on climate change: from prisoner's dilemma to chicken and beyond, International Environmental Agreements: Politics, Law and Economics, forthcoming

Schelling, T. (1960): The Strategy of Conflict, Harvard University Press

Schneider, G. and Cederman, L.-E. (1994). The change of tide in political cooperation: a limited information model of European integration, International Organization, vol 48, pp. 633-662

Spence, M.A (1973): Job Market Signaling, The Quarterly Journal of Economics, vol. 87(3), pp. 355-74

Stern, N. (2007): The Economics of Climate Change. Cambridge University Press

Stiglitz, J.E. (2000): The Contributions of the Economics of Information to Twentieth Century Economics, The Quarterly Journal of Economics, vol. 115 (4), pp. 1441-1478

van Vliet J., den Elzen M.G.J., van Vuuren D.P. (2009): Meeting radiative forcing targets under delayed participation. Energy Economics, vol. 31, pp. 152-162

Walsh, J.I. (2007): Do States Play Signaling Games?, Cooperation and Conflict, vol 42(2), pp. 441-459

Weikard, H.P. Finus, M., and Altamirano-Cabrera, J.C. (2006): The impact of surplus sharing on the stability of international climate agreements, Oxford Economic Papers, vol. 58, pp. 209-232

Figures

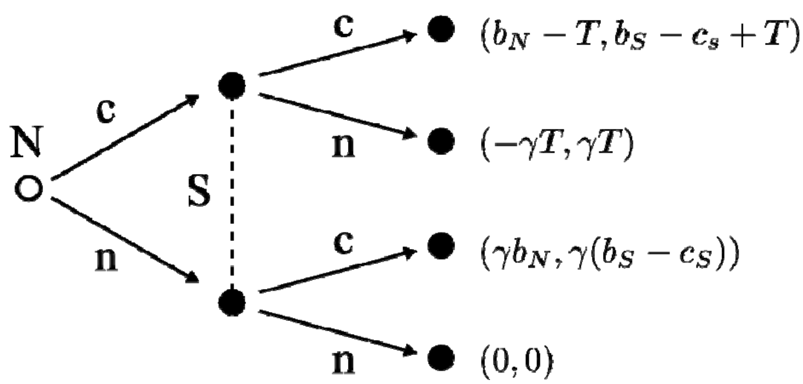

Figure 1: The IEA game with complete information in extended form[22]

[22] The graphical elements in our extended form games are borrowed from Kreps (1990), in particular dashed lines connect decision nodes that belong to the same information set.

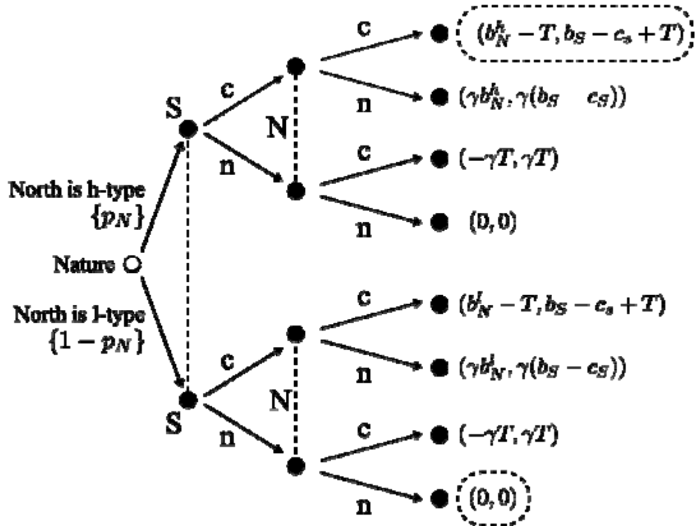

Figure 2: The IEA Game with Uncertainty about North's Benefits in Extended Form. Dashed ellipses denote the Bayesian Nash equilibria of the respective sub-game for the case that North's action depends on his type (i.e. Eq.(4') holds)

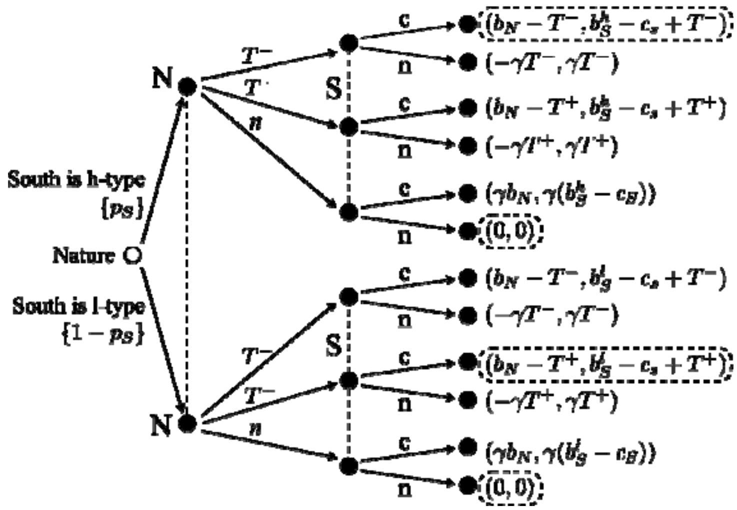

Figure 3: The IEA game with uncertainty about South's benefits. Dashed ellipses denote the Perfect Bayesian equilibria of the respective sub-game for the case that South's reaction to the transfer offered by North depends on her type (i.e. Eq.(1') holds)

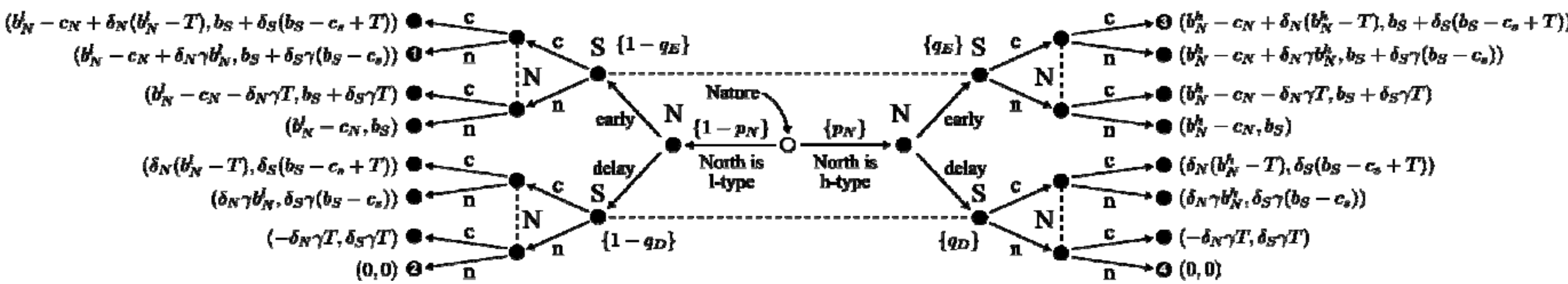

Figure 4: The two-stage signaling game if North's type is uncertain

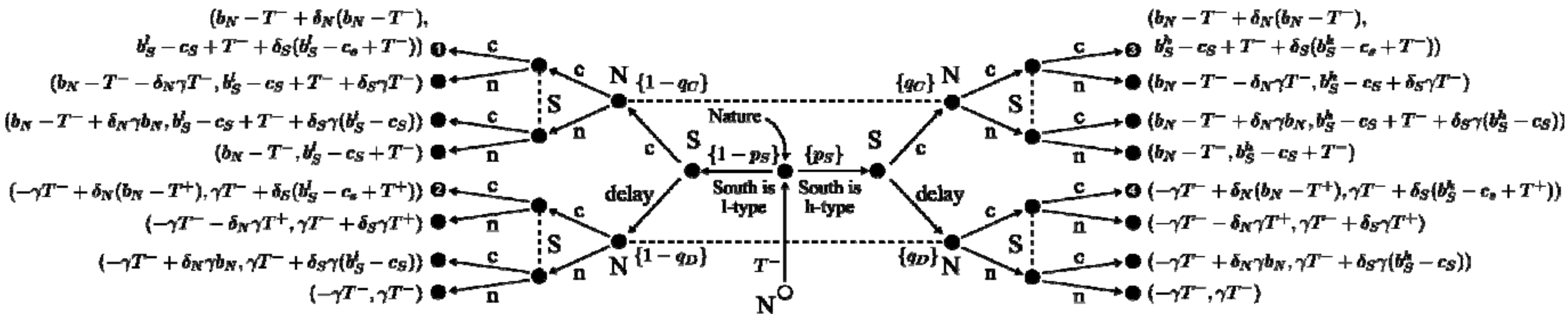

Figure 5: The two-stage signaling game if South's type is uncertain

Tables

Parameters			*Transfers*	*Probabilities*
$b_N^h = 6\%$	$b_S^h = 2\%$	$\alpha = 2/3$	$T^+ = 3.27\%$	$p_N > 0.39$
$b_N^l = 3\%$	$b_S^l = 1\%$	$\gamma = 0.2$	$T^- = 2.43\%$	$p_S < 0.79$
$c_N > 6\%$	$c_S = 3\%$			

Table 1: Parameters and results of the numerical example.

Costs and benefits are % of global GDP, p_N and p_S are probabilities required for cooperation