DISCUSSION PAPER SERIES

Discussion paper No. 22

Potential Failure of an International Environmental Agreement under Asymmetric Information

Norimichi Matsueda
Kwansei Gakuin University

April 2004

SCHOOL OF ECONOMICS

KWANSEI GAKUIN UNIVERSITY

1-155 Uegahara Ichiban-cho
Nishinomiya 662-8501, Japan

Potential Failure of an International Environmental Agreement under Asymmetric Information

Norimichi Matsueda

School of Economics, Kwansei Gakuin University, 1-1-155 Uegahara, Nishinomiya-shi, Hyogo 662-8501, Japan

ABSTRACT. The free-riding issue is generally considered to be the biggest obstacle in the success of an international environmental agreement. Even without free-riding incentives, however, asymmetric information can pose a potentially significant threat in establishing a cooperative relationship. In this study, we examine perfect Bayesian equilibria of a simple signaling game between a polluter country and a victim country over an agreement to mitigate unidirectional transboudary pollution. We show that the stalemate in addressing an international environmental issue can be partly explained by the incentive conflict due to the asymmetric information on the environmental preference of a polluter. (JEL Q20, D82)

Key words: asymmetric information, unidirectional transboundary pollution, Victim pays principle, signaling game, perfect Bayesian equilibrium.

I. INTRODUCTION

Recently, the developed world has finally seemed to start paying greater attention to the status of their environment, having attained certain standards of living domestically. In the examples of the Kyoto protocol against global warming and the Montreal protocol against the ozone depletion, certain developed nations appear quite willing to devote more financial and technological resources to protecting the global environment. In order to satisfy the basic economic needs of their citizens, on the other hand, developing nations still lean towards rapid economic growth, which sometimes leads to the dramatic environmental deterioration in many areas of the developing world. Furthermore, there is an increasing number of transboundary pollution issues in which developing countries are going to play more significant roles as pollution sources. Global warming, the ozone depletion and transboundary acid rain problems are just a few such examples.

Among transboundary pollution, there are several instances of “unidirectional”

pollution simply due to their climatic and geographical conditions. An example would be the transboundary acid rain problem from the U.K. to the Scandinavian countries where the westerly wind is predominant. In addition, with the developed world making more serious efforts to improve the global as well as its own environment, increasing instances of transboundary pollution are assuming the virtual characteristic of unidirectional pollution originating from developing countries. That is, even though physical conditions are reciprocal as in the case of the global atmosphere, a pollution issue can be treated essentially as unidirectional provided that some counties have contained its pollutant emissions quite significantly relative to the others.[1] The regional acid rain problem between China and Japan could fall into this category. In global pollution issues, such as global warming and the ozone depletion, the marginal abatement costs of the pollutants are considered to be significantly higher and getting even higher in developed countries than those in developing countries due to severe domestic restrictions on the pollutant emissions in the former nations. From an economic perspective, accordingly, such global pollution may be approximated in a simpler framework of unidirectional pollution where all the economically-relevant abatement opportunities are available in developing nations. Hence, the importance of investigating a case of unidirectional pollution is not as limited as it appears. In addition, whether it is physical or economic reasons that lead to a unidirectional pollutant flow, it often seems the case that a polluter nation is also suffering from its own pollutant emissions. This observation would be quite appropriate in the examples of the acid rain problem in China and the effects of global warming in the developing world.

As a general principle to solve international environmental conflicts, many nations have agreed to the so-called Polluter Pays Principle where a polluter should take full or, at least, partial responsibility for the environmental damages that it inflicted upon other nations. Without a proper institution to enforce this principle, however, any international agreement has to be established on a voluntary basis. Hence, economists have advocated the application of the Victim Pays Principle as a more successful and pragmatic approach to alleviate transboundary pollution under the current international circumstances (e.g., Baumol and Oates 1988; Missfeldt 1999). In case of unidirectional pollution, this usually requires some compensation from a victim country for promoting pollution abatement efforts in a polluter country. Given the fact that principal sources are recently shifting from wealthier developed countries to developing countries in an increasing number of transboundary pollution problems, a still prevalent economic disparity would render such

[1] In contrast, the study by Kaitala and Pohjola (1995) portrays the global warming issue as a unidirectional problem where one country is vulnerable to global warming but the other is not.

a solution even more plausible.

In reality, however, it is quite rare to observe a large-scale side payment from a victim in order to combat transboundry pollution.[2] There have been several previous attempts to explain why we rarely observe such international transfer provision even in simple unidirectional pollution from one polluter to one victim. One of them stems from the potential to weaken a victim's reputation as a negotiator. Mäler (1990) contends that, if a victim has to deal with two or more other countries on transboundary pollution issues, or to deal with the same country on several such issues, too great willingness to provide a side payment may give the country a reputation as a "weak" negotiator and increase the costs of other agreements. Another possible explanation would be the existence of significant transaction cost that works against materializing any kind of international agreement. Also, it has been pointed out that considering a transboundary environmental issue and other international relations at the same time could lead to a situation where a cooperative relationship can be more easily attained, even in the absence of a side payment. The analysis of an "issue linkage" or an interconnected game has been conducted by Folmer, van Mouche and Ragland (1993) and Cesar and de Zeeuw (1995).

The aim of this study is to present another possible explanation for the scarcity of cooperative relationships based on side payments in international environmental issues. Its main thesis is that the existence of asymmetric information between a victim and a polluter can be a source of difficulty in establishing such a relationship. In environmental economics, several previous studies have tackled the problem of asymmetric information, mostly concerning the effectiveness of abatement technologies or the resulting abatement cost, by resorting to the "mechanism design" approach (e.g., Ellis 1992). In contrast, this study investigates the information asymmetry about the environmental damage cost that a polluter incurs from its own emissions. Even though physical environmental damages of a polluter may be observable to a victim, the polluter's evaluation of its damages would be very difficult to infer from outside. In the context of transboundary pollution, few previous studies have examined the issue of asymmetric information about damage cost. In fact, in a typical analysis of unidirectional pollution, the environmental damages of a polluter have usually been assumed away. This might not be a realistic assumption as the acid rain is also a serious domestic environmental issue in China, for instance.

The structure of this paper is as follows. In the next section, we describe our issue and also present our simplified framework as preliminaries for the following analysis. In section III, we examine a signaling game where a polluter makes an announcement about

[2] An exception for this statement would be the Montreal Protocol where developing countries were made concessions in order to go along with it (Missfeldt 1999).

its non-cooperative abatement level and a victim responds with its offer of a side payment as its effort to establish an international agreement, and demonstrate that there is a possibility that the polluter and the victim cannot agree upon a Pareto-superior agreement. In section IV, we discuss the results of our signaling game and consider several ways to extend our simple model. The last section concludes this study with final remarks.

II. ISSUE

First of all, in order to focus on the issue with asymmetric information rather than the "free-ride" problem which could arise among multiple polluters or victims, we assume that there are only one polluter and one victim in this unidirectional pollution problem. Let e be the level of pollutant abatement effort made by the polluter.[3] Its abatement cost function is assumed to be increasing and strictly convex in e. In our unidirectional pollution, the emissions of the pollutant cause pollution not only in the victim country but also in the polluter country itself. We suppose that the damage cost functions of the victim and the polluter are both decreasing and convex in e. For simplicity, the number of possible types of the polluter's damage cost is limited to two. In an asymmetric information case, the victim does not know which of these two types is true, and they are distinguished only by their marginal damage costs.

If there is complete information upon the type of the polluter, we can easily identify its truthful abatement level in the non-cooperative situation, where the polluter takes only its own cost into account, as the level that minimizes the sum of its abatement and damage cost. Such an abatement level can be implicitly given by the usual marginal condition[4], which is

$$MAC(e) = MDC_p^{\,i}(e), \qquad [1]$$

where $MAC(e)$ is the marginal abatement cost function, and $MDC_p^i(e)$ (the superscript i signifies its type and $i = L$ or H) is the negative of the marginal damage cost function of the polluter.[5]

On the contrary, the abatement level in the cooperative case, where the damage cost of the victim is added to the consideration, is given by

[3] In our framework, we suppose that the pollutant emissions in a victim country are completely contained for domestic reasons.

[4] Given the assumptions on the curvatures of the abatement and damage cost functions, the first-order conditions [1] and [2] are not only necessary but also sufficient for cost-minimization.

[5] Precisely speaking, the latter signifies the value of environmental damages avoided by one additional abatement effort.

$$MAC(e) = MDC_p^{\,i}(e) + MDC_v(e), \qquad [2]$$

where $MDC_v(e)$ is the negative of the marginal damage cost of the victim. This is simply a variation of the "Samuelson condition" for the provision of a public good. When there is not too significant a difference between $MDC_p^{\,L}(e)$ and $MDC_p^{\,H}(e)$ relative to the size of $MDC_v(e)$, we can depict these marginal cost curves as in Figure 1 for instance.[6] Given the marginal conditions above, the four abatement levels, $e_N^{\,L}$, $e_N^{\,H}$, $e_C^{\,L}$ and $e_C^{\,H}$, in Figure 1 are respectively the non-cooperative abatement levels and the internationally optimal or cooperative abatement levels for the two different types of the polluter.

Although the cooperative abatement level certainly increases the total welfare over the non-cooperative level, there is currently no world government to force the polluter to take into account the damage cost of the victim as well as its own. In order to realize the cooperative abatement level in essentially a non-cooperative international setting, therefore, the victim needs to provide a sufficient amount of side payment. Such a side payment has to leave the polluter, now saddled with the higher abatement assignment, at least as well off as its non-cooperative welfare. Using Figure 1, we can see that the "minimal" amounts of side payment required for attaining the cooperative abatement levels are the area (C+E+F) and the area (B+C+D), for the high and low cost types of the polluter, respectively. With such compensation amounts, each type of the polluter becomes indifferent between the non-cooperative and the cooperative abatement levels, and all the gain from the cooperation accrues to the victim.

However, if there is asymmetric information on the type of the polluter, by behaving strategically a certain type of the polluter may be able to gain a strictly positive profit after receiving a side payment in an international agreement. In our setting, the high damage cost type polluter may improve its welfare by pretending itself as the low type. Let us see how this is possible. If there is no international agreement, the high damage polluter definitely loses by pretending as the low type and choosing $e_N^{\,L}$, instead of choosing its truthful non-cooperative abatement level, $e_N^{\,H}$. Unless a side payment is provided, the area (A) in Figure 1 signifies the amount of its welfare loss by pretending as the low cost type. However, the victim needs to pay at least the area (B+C+D) in order to induce the low damage polluter to attain the cooperative abatement level of $e_C^{\,L}$, which actually requires the high damage polluter to bear the extra cost of only the area (C) over its reduced damage cost. Therefore, the high damage type polluter that succeeds in

[6] If there is a very wide difference between these two curves relative to the size of the marginal damage cost of the victim, the $MDC_p^{\,H}$ curve might locate above the $MDC_p^{\,L}$+MDC_v curve for any abatement level. In the following analysis, however, this case is precluded by the assumptions on the relevant parameter values.

convincing the victim that he is the low damage type could gain the area (B+D) even if the international agreement is implemented with the minimal side payment.

In the face of such a potential strategic action, the victim may try to implement a counter-measure to mitigate loss arisen from this behavior. To an economic problem with a hidden characteristic, the idea of the "revelation principle" has been widely applied (Fudenberg and Tiorle 1991). In our unidirectional pollution context, this principle essentially implies that a victim can come up with a certain menu of various contracts, each of which is intended for a particular type of the polluter, and that allowing the polluter to voluntarily choose his favorite contract will, in fact, result in the revelation of its true type and the highest possible welfare for the victim, even though some information rent will usually accrue to the polluter, depending on its actual type. In each contract, the amount of side payment from the victim and the polluter's abatement level in the agreement is clearly specified. Then, a contract is immediately agreed by the two parties to implement, which does not explain the observed infrequency of international agreements with side payments. In our signaling model the offer of this sort of two-dimensional contract is not allowed. In particular, the abatement level in a potential international agreement is determined exogenously, possibly by an outside agency or by a separate negotiation between the two countries, in such a manner that the polluter's choice of its non-cooperative abatement level must be honored. More precisely, we assume that the targeted level of abatement is selected to be an internationally optimal level, based on the polluter's non-cooperative abatement choice.

Henceforth, without the loss of generality, we restrict our attentions to a very tractable case of a quadratic abatement function for a polluter and linear damage cost functions for the two countries. We suppose that the possible types of the polluter's damage cost are limited to two types as above, and, moreover, that the difference between these two types is entirely represented by the difference in the slopes of these damage cost functions. Using the similar notations, the damage cost function of the polluter, $DC_p^i(e)$, is defined as

$$DC_p^{\,i}(e) = \theta_i(e_p^{\,u} - e). \qquad [3]$$

Here, θ_i is either θ_L or θ_H, depending upon its types, and $e_p^{\,u}$ is the abatement level above which the pollutant emissions become completely harmless for the polluter. The actual value of the parameter, θ_i, is privately known to the polluter, but the victim merely has an *ex ante* subjective probability distribution regarding the type of the damage cost function of the polluter. This "prior belief" is defined such that the victim originally expects to face the low damage type and the high damage type with the probability of p ($0 < p < 1$)

and 1-p, respectively. We assume that this belief is common knowledge between the two parties for the sake of the analysis in the next section.

On the other hand, the damage cost function of the victim, $DC_v(e)$, is expressed as

$$DC_v(e) = d(e_v^u - e). \qquad [4]$$

Similarly to e_p^u above, e_v^u is the polluter's abatement level above which the pollutant emissions are harmless for the victim.[7] The parameter d is just a constant and known by both countries. Finally, the abatement cost function of the polluter, $AC(e)$, is expressed as

$$AC(e) = \frac{1}{2}ce^2, \qquad [5]$$

where c is a known constant. With these simply specified set of functions, Figure 1 can be rewritten as Figure 2. Then, the non-cooperative abatement level of the polluter in the absence of a side payment can be easily derived from [1], [3], and [5] as $e_i^N = \theta_i/c$ for $i = L$ or H. On the other hand, the cooperative abatement level can be derived from [2], [3], [4], and [5] as $e_i^C = (\theta_i+d)/c$ for $i = L$ or H.

III. SIGNALING GAME AND ITS EQUILIBRIA

In this section, we model our problem as a simple signaling game and examine its perfect Bayesian equilibria.[8] The game tree in Figure 3 depicts the strategic interactions that take place between the polluter and the victim. As is well known in the game theory literature, a game of incomplete information can be transformed with the introduction of the initial move by "Nature" to a game of imperfect information (Harsanyi 1967). An important restriction of this transformation in our context is that the victim's prior belief on the polluter's damage cost must be common knowledge. With this transformation, Nature moves initially to determine the type of the polluter's domestic damage cost. Knowing the result of this Nature's move, the polluter at the node P_L^0 or P_H^0 makes a committed announcement regarding the non-cooperative level of abatement. Here, we assume that this announcement is committed in the sense that the polluter has to implement the chosen non-cooperative abatement level without any assistance in case that the agreement is not eventually reached.[9] We will discuss the alteration of this

[7] We suppose that the level of abatement efforts will not reach e_p^u or e_v^u under any circumstance. Then, due to the linear specification of the damage cost function, these values are actually unimportant in our analysis.
[8] For the definition of a perfect Bayesian equilibrium and other game-theoretic concepts in this section, see Fudenberg and Tirole (1991).
[9] Equivalently, the announcement can be made concerning its own damage cost as long as its non-cooperative abatement level of the polluter is based on this announcement when an agreement eventually fails.

assumption in the next section.

After this action of the polluter, the cooperative abatement level in the international agreement is determined so as to maximize the global gain from the agreement according to the announced non-cooperative abatement level by the polluter. That is, the internationally optimal level of abatement is selected as the abatement level in a potential agreement. In our simple game, this decision is made exogenously. Then, the victim chooses the amount of side payment, s, which will be offered to the polluter in exchange for attaining the cooperative abatement level. Finally, the polluter chooses either to accept this offer of a side payment or to reject it. In case of the acceptance, the two countries engage in the international agreement which implies the implementation of the internationally optimal abatement level by the polluter and the provision of the side payment by the victim. In case of the rejection, the polluter implements its announced non-cooperative abatement level without any assistance from the victim.

The respective payoffs for the two countries in Figure 3 are obtained by calculating the corresponding areas in Figure 2. The left and right entries in each parenthesis are the payoffs of the polluter and the victim, respectively. We suppose that the payoff for conducting its own non-cooperative abatement level is simply zero for each type of the polluter. On the other hand, we evaluate the victim's payoffs by supposing that its welfare level at e_L^N is standardized to zero.[10] Let us further denote θ_H-θ_L by θ for the sake of notational convenience.

The game tree in Figure 3 is depicted in a slightly abbreviated way in order to focus on the interactions after the choice of e_L by the polluter. In our setup, the choice of e_H essentially reveals that the polluter is the high cost type because the low cost type would never choose e_H. The low cost type polluter always acts honestly at the node P_L^0 because the low cost type polluter can secure itself the payoff of zero by choosing e_L while it would lose $\theta(\theta+2d)/2c$ by being required to abate up to e_H^C with the amount of side payment that leaves the high cost type polluter just break-even, which is represented by the area (k+n+o). This low cost type polluter's payoff is obtained by calculating the negative of the area (f+l+t), which is the area (k+n+o) minus the area (f+k+l+n+o+t) in Figure 2. Suppose that the high cost type polluter acts honestly at the node P_H^0, it also receives the area (k+n+o) from the victim and just breaks even. However, the high cost type has an incentive to lie and state e_L at the node P_H^0 because of the possibility that it is treated as the low cost type and required to abate only up to e_L^C with the side payment that yields the polluter a strictly positive payoff of the area (f+l), which is given by subtracting

[10] These assumptions on payoffs do not lead to any loss of generality because only the relative magnitude of each payoff matters to the choice of an action by each player.

the area (k) from the area (f+k+l).

In fact, merely by the considerations of the belief at the information set V, it is clear that under any circumstances we cannot have a "separating" equilibrium where both types of the polluter act honestly at their initial nodes. Let us suppose that, when the information set V is reached, the victim believes that the polluter is the low cost type with the probability r ($0 \leq r \leq 1$) and it is the high cost type with the probability 1-r. If they are behaving honestly at the initial nodes, the consistency requirement of the victim's posterior belief specifies $r = 1$. Then, the victim will offer the side payment that leaves the low cost type just break-even. However, such an amount of side payment will induce the high cost type to switch its strategy and choose e_L^N, thus annihilating the possibility of this separating equilibrium.

In fact, depending upon the victim's prior belief concerning the polluter's type, the game yields two different kinds of perfect Bayesian equilibria. When p is rather large, the game has a "pooling" equilibrium where the two types of the polluter choose the same action, e_L^N, at the initial nodes. In this equilibrium, an international agreement is always implemented on a relatively small scale. On the other hand, when p is sufficiently small, we have a "hybrid" or "semi-separating" equilibrium. In such an equilibrium, the high cost type polluter is playing a mixed strategy at the node P_H^0. Especially, it randomizes between e_H^N and e_L^N, and e_L^N is the action always taken by the other type at the node P_L^0. Moreover, in this equilibrium it is possible that there is no international agreement reached. Let us see how we can obtain these results.

The action by the polluter at the end of the game is straightforward. Each type of the polluter accepts a side payment from the victim and engages in an agreement only if its payoff from the acceptance is greater than its payoff from the rejection. For the low cost type polluter, this implies that it ought to accept s if s-$d^2/2c \geq 0$, that is, $s \geq d^2/2c$, and reject s otherwise. The high cost type polluter would rather accept s if s-$(d-\theta)^2/2c \geq -\theta^2/2c$, which can be transformed to $s \geq (d^2-2\theta d)/2c$, and reject s otherwise. In this study, we suppose $d \geq 2\theta$. That is, the damage cost of the victim is at least twice as large as the difference in the damage costs of the two types of the polluter. When this is not the case, there is a possibility that a negative amount of side payment is offered by the victim and accepted by the high cost type polluter in the equilibrium. This is a consequence of our assumption that the initial action of the polluter cannot be reversed after an agreement eventually fails.

Given this strategy of the polluter at the last nodes and its belief over the types of the polluter, the victim determines the amount of side payment. The belief is characterized by the value of r, and the objective of the victim is to maximize its expected payoff. Let

Pr(s) be the probability that the offer of s is accepted by the polluter. According to the polluter's response toward s at its last nodes, there are three possible values for Pr(s); Pr(s) = 1 when $s \geq d^2/2c$, Pr(s) = 1-r when $d^2/2c > s \geq (d^2-2\theta d)/2c$, and Pr($s$) = 0 when $s < (d^2-2\theta d)/2c$. Then, the victim's problem is expressed as

$$\underset{s}{Max}\left(\frac{d^2}{c}-s\right)\Pr(s)\,. \qquad [6]$$

The payoff function to the victim is depicted in Figure 4, specifically when it is willing to use a mixed strategy. In order for the victim to employ a mixed strategy, the expected payoff from offering $s = d^2/2c$ must be equal to the expected payoff from offering $s = (d^2-2\theta d)/2c$. Thus, a mixed strategy is employed by the victim only when

$$(1-r)\left(\frac{d^2}{c}-\frac{d^2-2\theta d}{2c}\right)=\frac{d^2}{c}-\frac{d^2}{2c}\,. \qquad [7]$$

By solving [7] for r, we obtain $r = 2\theta/(d+2\theta)$. As we will see below, this particular value of r turns out to be an important threshold and we denote it by r^*. If the victim believes that there is more chance of the polluter's being the low cost type than r^*, it offers $s = d^2/2c$. On the contrary, when it believes that the polluter is less likely to be the low cost type than r^*, the victim offers $s = (d^2-2\theta d)/2c$. Then, Only when $r = r^* = 2\theta/(d+2\theta)$, the victim employs a mixed strategy. Let us suppose that, as its mixed strategy, the victim chooses $s = d^2/2c$ with the probability w ($0 < w < 1$) and $s = (d^2-2\theta d)/2c$ with the probability 1-w, respectively.

Now, we consider the action taken by the high cost type polluter at the node P_H^0. When it chooses e_H^N, the international agreement aimed at e_H^C will be implemented with the side payment which makes the high cost type polluter just break-even because the victim is certain that no low cost type polluter had chosen e_H^N at the node P_H^0. Consequently, the high cost type polluter obtains zero by choosing e_H^N. Then, the expected payoff from choosing e_L^N must also be zero for the high cost type polluter to randomize at P_H^0. In such a hybrid equilibrium, therefore, the following equation has to be satisfied:

$$w\left(\frac{d^2}{2c}-\frac{(d-\theta)^2}{2c}\right)+(1-w)\left(\frac{d^2-2\theta d}{2c}-\frac{(d-\theta)^2}{2c}\right)=0\,. \qquad [8]$$

Solving [8] for w, we obtain $w = \theta/2d$.[11] Hence, for a mixed strategy to be the equilibrium strategy for the high cost type polluter, we need $w = \theta/2d$ as the victim's mixed strategy.

Finally, we must specify the belief of the victim when the information set V is reached. This posterior belief has to be determined by the Bayes' rule, given the victim's prior

belief and the equilibrium strategy of the polluter. Let us suppose that the probability with which the high cost type polluter chooses e_L^N at the node P_H^0 is u ($0 < u < 1$).[12] As we have seen above, in order for the victim to employ its mixed strategy, we must have $r = 2\theta/(d+2\theta)$. Hence, by the Bayes' rule, the following equation must be satisfied:

$$\frac{2\theta}{d+2\theta} = \frac{p}{p+(1-p)u}. \qquad [9]$$

Solving [9] for u, we obtain $u = dp/2\theta(1-p)$. While the constraint $0 < u$ can be trivially satisfied, the constraint $u < 1$ provides a condition for obtaining the hybrid equilibrium. That is, we have the hybrid equilibrium only if $u = dp/2\theta(1-p) < 1$, which yields the condition upon p as $p < 2\theta/(d+2\theta)$.

In summary, depending on the prior belief of the victim concerning the polluter's type, we have two different kinds of perfect Bayesian equilbriua for our particular signaling game. The first is the pooling equilibrium where both the low cost and high cost types of the polluter choose e_L^N as their non-cooperative abatement level. This equilibrium realizes if $p \geq 2\theta/(d+2\theta)$ as in Figure 5. In this pooling equilibrium, the victim does not update its posterior belief. The polluter's initial move is followed by the offer of $s = d^2/2c$ by the victim and this offer will be accepted by the polluter. Hence, in this equilibrium, there is always an international agreement on a relatively small scale. The other possible equilibrium is the hybrid equilibrium which occurs when $p < 2\theta/(d+2\theta)$. In this equilibrium, the high cost type polluter employs the mixed strategy as its local strategy at its initial move, randomizing between e_L^N and e_H^N with the probabilities $u = dp/2\theta(1-p)$ and $1-u$, respectively. At the information set V, the victim's posterior belief is $r^* = 2\theta/(d+2\theta)$ on the left node and $1-r^*$ on the right node, and it offers $s = d^2/2c$ with the probability $w = \theta/2d$, and $s = (d^2-2\theta d)/2c$ with the probability $1-w$. The polluter accepts this offer only when its payoff from the acceptance is greater than its payoff when it rejects the offer, that is, $s \geq d^2/2c$ for the low cost type and $s \geq (d^2-2\theta d)/2c$ for the high cost type, and otherwise it rejects the offer, implying that an international agreement can fail in the latter case.

IV. DISCUSSIONS

The most interesting result obtained in the previous section is that there is a possibility that an international agreement is not reached successfully despite the fact that it represents a Pareto improving change. In the hybrid equilibrium, an agreement fails with

---

[11] If $u = 1$, we have a pooling equilibrium.

[12] Note that, because we assumed $d \geq 2\theta$, it is always the case that $0 < w = \theta/2d < 1$.

the probability $r^*(1\text{-}w) = \{\theta(2d\text{-}\theta)\}/\{d(d+2\theta)\}$, because the low cost type polluter rejects the offer of $s = (d^2\text{-}2\theta d)/2c$ from the victim. Let us define $q = r^*(1\text{-}w)$. Interestingly, the value of q does not depend on the level of the victim's prior belief concerning the polluter's type. This is because, no matter what the victim's prior belief might be, the value of r has to equal $r^* = 2\theta/(d+2\theta)$ for a mixed strategy to become the equilibrium strategy for the victim, which is required for the realization of the hybrid equilibrium. On the other hand, the victim's prior belief is critical in determining whether the hybrid equilibrium actually emerges. As long as p is sufficiently small, which means that the polluter is believed sufficiently likely to be the high cost type, we have a non-zero probability that a mutually beneficial international agreement is rejected by the polluter.

Furthermore, we can easily derive $\partial r^*/\partial d = -2\theta/(d+2\theta)^2 < 0$, which implies that, the more concerned the victim is about its own environmental damages, the less likely we have the hybrid equilibrium given an exogenous probability distribution over the types of the polluter. We can also obtain $\partial q/\partial d = 2\theta\{d(\theta\text{-}d)+\theta^2\}/\{d(d+2\theta)\}^2$, where q is the probability that an international agreement fails in the hybrid equilibrium. Then, a straightforward calculation shows that $\partial q/\partial d < 0$ always holds for $d \geq 2\theta$, which we have assumed above. This result implies that the likelihood of an eventual disagreement goes down with the increase in d. Both $\partial r^*/\partial d < 0$, and $\partial q/\partial d < 0$ indicate that an international agreement is more likely to be reached if the victim cares about its own environmental damages more significantly. These results are somewhat intuitive in that, the more important this environmental issue is to the victim, the more strongly it would seek an international agreement even if it might allow the high cost type polluter to pretend as the low cost type.

On the other hand, we can derive $\partial r^*/\partial \theta = 2d/(d+2\theta)^2 > 0$, which implies that, the greater the difference between the two polluter's types is, the more likely the hybrid equilibrium realizes.[13] Therefore, as the value of θ increases, we are more likely to observe the hybrid equilibrium although whether we have the hybrid equilibrium or the pooling equilibrium also depends on the actual value of p. Moreover, we can derive $\partial q/\partial \theta = 2d\{d^2\text{-}\theta(d+\theta)\}/\{d(d+2\theta)\}^2$. Then, it follows that $\partial q/\partial \theta > 0$ for $d \geq 2\theta$. Hence, the likelihood of the actual failure of the international agreement in the hybrid equilibrium goes up with the increase in θ. As opposed to the change in d, both $\partial r^*/\partial \theta > 0$ and $\partial q/\partial \theta > 0$ indicate that the two countries have more difficulty in reaching an agreement when the difference between the two types of the polluter is relatively large. These results imply the fact that the high cost type polluter finds it more difficult to successfully disguise itself as the low cost type in such a case.

The structure of the signaling game above is admittedly very simple. We can easily think of several ways to extend the model. First, we can consider more than two types of the polluter in terms of its damage cost with different probability distributions over its types as a prior belief of the victim. This extension will certainly complicate the model but would not change the basics of strategic interactions between the polluter and the victim over a resulting form of an international agreement. That is, as long as certain types of the polluter have interest in misrepresenting themselves and the victim loses by such a disguise, the victim may attempt to "play hard" by offering a relatively small amount of side payment as a measure to discipline such a polluter and not to yield to the polluter's disguise too easily.

The second possible extension would be to include a repeated bargaining process over the amount of side payment. Recently, there have been many important works on bargaining with asymmetric information, and they provide us with several insights into such a bargaining. In a bargaining situation, we usually introduce discount factors which measure the impatience of negotiating parties and make it more attractive for them to reach an agreement sooner than later. In a bargaining game where only an uninformed party makes proposals and there are continuous types of the receiver, a "screening" equilibrium typically emerges. In such an equilibrium, the uninformed player screens the other player's private information through time via the sequence of offers (Kennan and Wilson 1993). In our context, the victim starts offering a relatively small amount of side payment which can be accepted only by a polluter with sufficiently high domestic damage cost, and, every time the polluter rejects its previous offer, the victim will gradually increase its offer amount for the purpose of screening the polluter's type.

Additionally, our setting is an example of a so-called "gap case" because the total welfare gain from an agreement is always strictly greater than the cost to implement it, irrespective of the polluter's type. As Muthoo (1999) demonstrates for a general gap case with infinite opportunities of making proposals, the two parties will always agree on a mutually beneficial trade in some finite time. That is, there is no possibility that the players disagree and fail to reach an agreement indefinitely. In our specific context, if the victim possesses the infinite opportunities of proposing side payments, the polluter will accepts the offer sooner or later. However, the delay in reaching an agreement still represents an overall efficiency loss.

We have also assumed above that the polluter needs to commit to its initially announced non-cooperative abatement level in the sense that it has to implement this level of abatement when an agreement eventually fails. However, if the costs of reneging

[13] It is always the case that $r^* < 1/2$ since we have assumed $d \geq 2\theta$.

on its initial announcement, such as the cost of losing international reputation, are relatively insignificant, the polluter has an incentive to avoid its own welfare loss that arises from actually implementing too small a non-cooperative abatement level. If the polluter's renouncement is completely costless, we only observe the pooling equilibrium where the polluter always chooses e_L^N at the node P_H^0. The high cost type polluter will never do worse by claiming e_L^N than by acting honestly because it is guaranteed the payoff of zero by choosing e_L^N in this case. Therefore, the belief of the victim at the information set V simply coincides with its prior belief. A similar calculation to the previous section shows that, if $p \geq (2\theta d-\theta^2)/(d^2+2\theta d-\theta^2)$, the victim offers $s = d^2/2c$, and if $p < (2\theta d-\theta^2)/(d^2+2\theta d-\theta^2)$, it offers $s = (d-\theta)^2/2c$. Although this outcome is qualitatively different from the one above, there is still a possibility that an international agreement fails when $p < (2\theta d-\theta^2)/(d^2+2\theta d-\theta^2)$ because the low cost type polluter surely rejects the offer of $s = (d-\theta)^2/2c$.

V. CONCLUSION

The stalemate in addressing an international environmental issue can potentially be explained by the incentive conflict arisen from the asymmetric information on the environmental preference of a polluter. Without the information on its damage cost, a victim or any international agency cannot judge exactly what would be the internationally optimal level of abatement and the appropriate amount of side payment. In this study, we examined a signaling game between a polluter country and a victim country over an agreement to mitigate unidirectional transboudary pollution. In our simple analytical setting, we have seen that the existence of asymmetric information can prohibit the realization of a Pareto-superior international agreement. Specifically, such a situation is likely to occur when the uncertainty over the polluter's damage cost is sufficiently significant compared to the size of the victim's damage cost.

References

Baumol, W. and W. Oates. 1988. *The Theory of Environmental Policy*. 2nd Edition. Cambridge: Cambridge University Press.

Cesar, H. and A. de Zeeuw. 1995. "Issue Linkage in Global Environmental Problems." In *Environmental Policy for the Environmental and Natural Resources: Techniques for the Management and Control of Pollution*, ed. A. Xepapadeas. Cheltenham: E. Elger.

Ellis, G. 1992. “Incentive Compatible Environmental Regulation.” *Natural Resource Modeling* 6: 225-256.

Folmer, H., P. van Mouche, and S. Ragland. 1993. “Interconnected Games and International Environmental Problems.” *Environmental and Resource Economics* 3: 313-335.

Fudenberg, D. and J. Tirole. 1991. *Game Theory*. Cambridge: MIT Press.

Harsanyi, J. 1967. “Games with Incomplete Information Played by Bayesian Players.” *Management Science* 14: 159-182, 320-334, 486-502.

Kaitala, V. and M. Pohjola. 1995. “Sustainable International Agreements on Greenhouse Warming - A Game Theory Study.” In *Control and Game-theoretic Models of the Environment*, eds. C. Carraro and J. Filer. Boston: Birkhauser.

Kennan J. and R. Wilson. 1993. “Bargaining with Private Information.” *Journal of Economic Literature* 31: 45-104.

Mäler, K. 1990. “International Environmental Problem.” *Oxford Review of Economic Policy* 6: 80-108.

Missfeldt, F. 1999. “Game-theoretic Modelling of Transboundary Pollution.” *Journal of Economic Surveys* 13: 287-321.

Muthoo, A. 1999. *Bargaining and Its Applications*. Cambridge: Cambridge University Press.

FIGURES

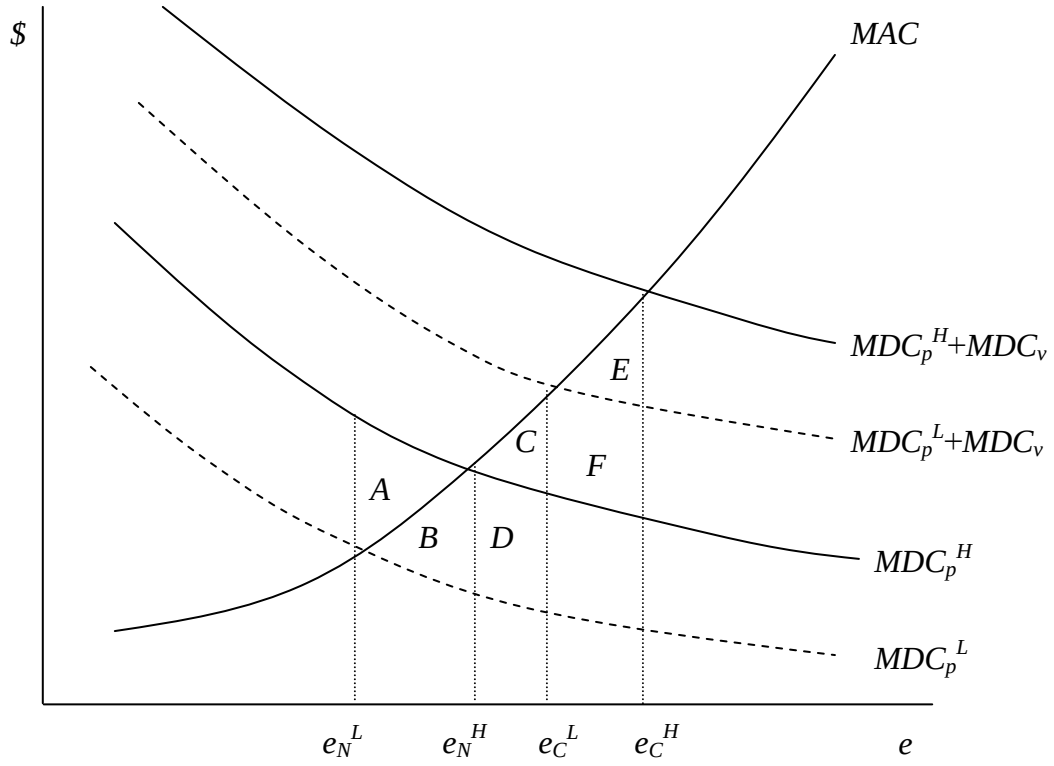

FIGURE 1 Illustration of Potential Gain from Disguising

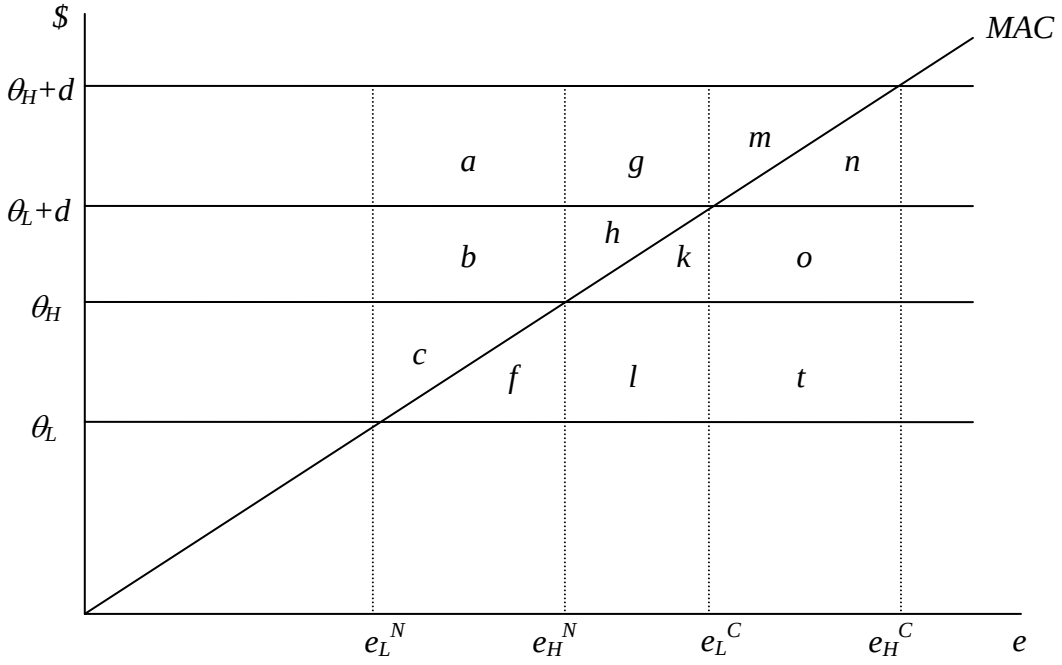

FIGURE 2 Case with Linear Damage and Quadratic Abatement Functions

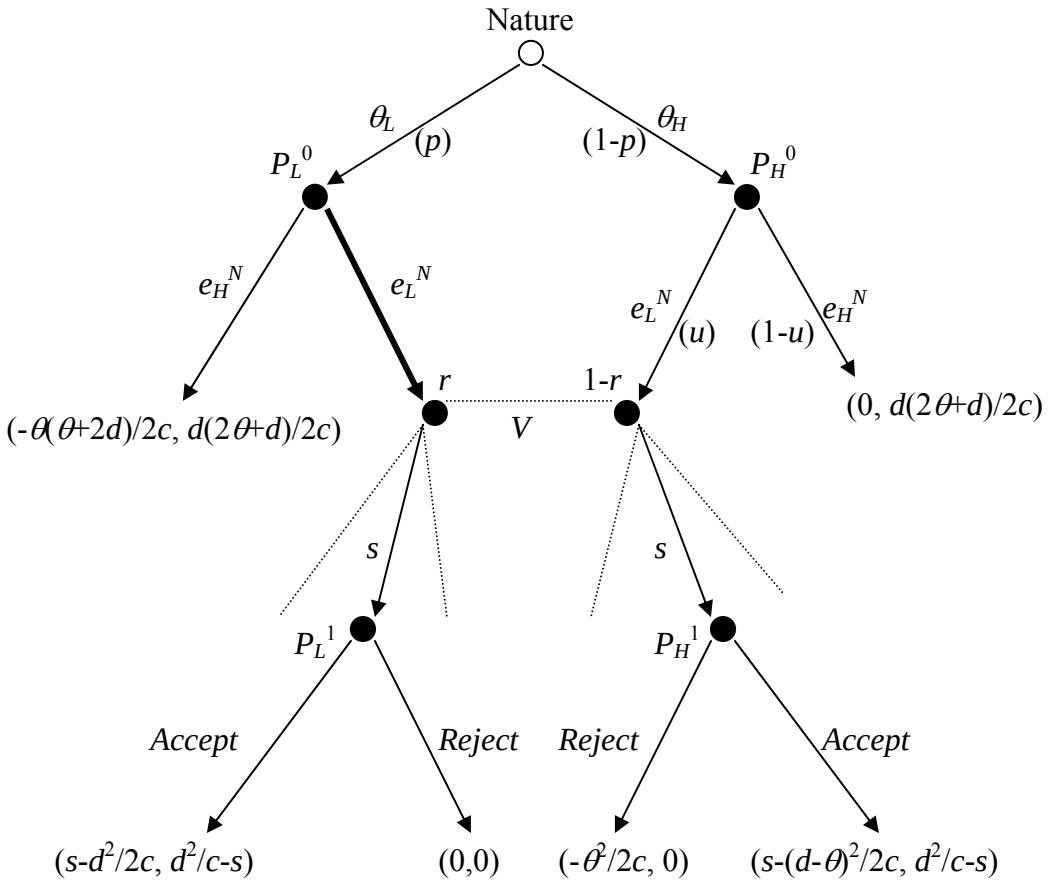

FIGURE 3 Tree of Our Signaling Game

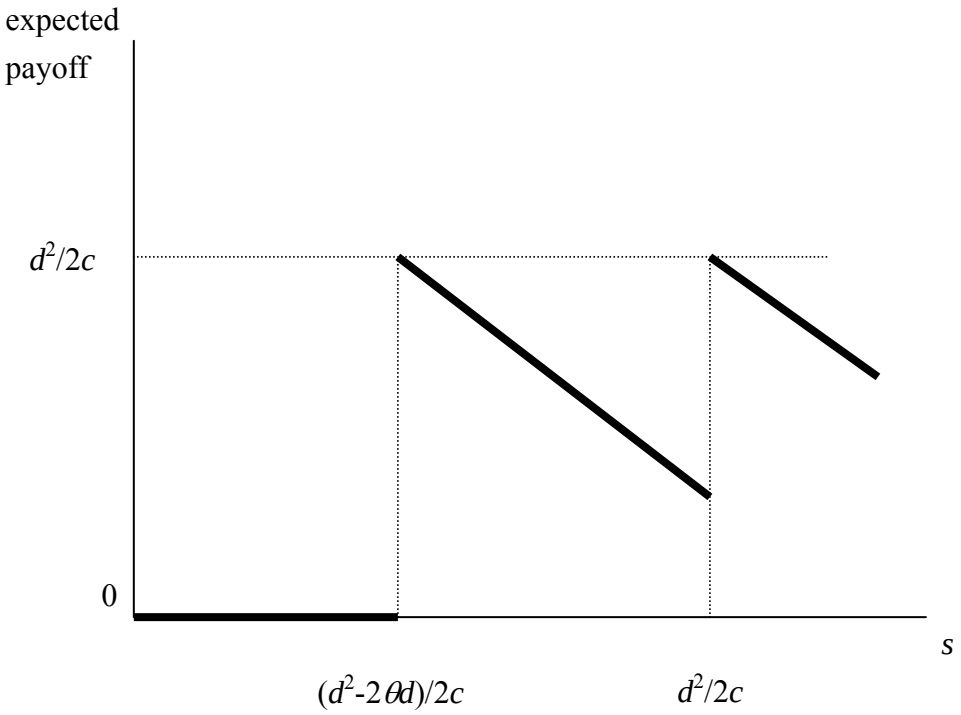

FIGURE 4 Illustration of the Victim's Payoff in the Hybrid Equilibrium

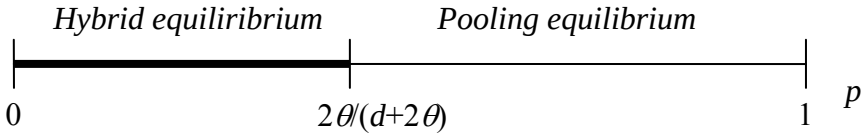

FIGURE 5 The Illustration of Equilibria Differentiated by the Prior Belief