Draft, October 8, 2009

REWARDING THE PUNISHER:

ENTICING COMPLIANCE WITH INTERNATIONAL TREATIES

Omri Ben-Shahar and Anu Bradford

The University of Chicago Law School

Introduction

Some states sign, ratify and comply with international treaties. Others fail to do the same. Non-compliance is a pertinent problem undermining all areas of international law, including international environmental law where states frequently have an incentive to defect from their international commitments. A state acting in self-interest hopes to free-ride on the efforts of other states to preserve the environment, while continuing itself to deplete the common (limited) resource pool. This leads to the familiar problem known as the "tragedy of commons" where individual states' behavior ultimately destroys the shared resource. [1]

This paper studies a mechanism for enhancing international cooperation to preserve a public good such as clean atmosphere. The set up is familiar: some states refuse to join a treaty seeking to preserve public good or, should they join the treaty, defect from their commitments. These states are called *Violators.* Other states are eager to protect the environment and seek to establish and enforce international norms aimed at preserving it. These states are called

Enforcers. Enforcement, though, is costly, which means that some violations would go unpunished and, consequently, undeterred in the first place.

This paper focuses on the cost of enforcement as a key impediment for achieving international environmental cooperation. It offers a game-theoretic framework for examining the incentives of the Enforcer to impose a sanction on a Violator. The model developed in this paper shows that any sanction-based system to induce compliance is inadequate if the Enforcer's threat to punish the Violator is not credible. This is the case each time the cost of carrying out its threat is higher for the Enforcer than the cost of tolerating the non-compliance of the Violator. Thus, the cost of imposing sanctions will always lead to some equilibrium level of non-compliance. Given that the threat of sanctions cannot eliminate non-compliance, the model examines whether Enforcer can achieve further compliance by introducing a system of rewards. It shows that the introduction of pecuniary incentives leads to a superior level of compliance. A system of positive incentives may make everyone better off—the Violator and, more surprisingly, the Enforcer, which finances the system of rewards.

The key insight, which in the model accounts for the superiority of rewards over sanctions, is the idea that a reward fund can be used not only to bribe Violators and furnish them with incentives to comply, but also to compensate Enforcers for the costs of inflicting sanctions. In effect, a dollar of reward doubles the deterrent effect relative to a dollar of sanction. Intuitively, this is similar to using a defendant's bail money to fund bounty hunters. The defendant recognizes that by fleeing he would not only forfeit the bail, but now he is more likely to be apprehended. The bail fund provides a double deterrent.

[1] Garret Hardin.

After demonstrating the optimal use of sanctions and rewards through a simple model, this paper applies the insights of the model into the negotiations of a global climate change treaty ("GCCT"). The pursuit of an effective GCCT, designed to replace the United Nations Framework Convention on Climate Change and the adjoining Kyoto Protocol, is one of the most pressing tasks facing the international community.[2] The greatest challenge underlying the impending negotiations is the difficulty of securing the cooperation of developing countries, in particular that of China, which is the biggest emitter of greenhouse gases ("GHG") in the world. Developing countries did not undertake binding commitments to reduce their GHG emissions under the Kyoto Protocol, compromising the effectiveness of the Treaty. Of course, a related challenge is to induce the developed countries to participate in the active enforcement of the GCCT. The paper examines how the combination of sanctions and rewards might, in the end, be a more effective strategy in ensuring that developing countries will abandon their status as Violators and that developed countries will join forces as Enforcers in a global fight to halt climate change.

I. Model

A. Framework of Analysis

A group of states share a common pool (e.g., water resource, fishery, atmosphere). The interaction between the states is modeled as a game with the following components:

(i) Players: There are *N* identical violator states, labeled V_i, $I = 1,2,...N$, and one enforcing state, labeled *E*, for a total of $N+1$ players. Each player is a homogenous entity, i.e., any domestic diversity is resolved prior to the start of the international interaction. While this formulation is a

[2] The Kyoto Protocol entered into force in 2005. To date, 184 states have ratified it. These states do not include the United States. In December 2009, states will gather in Copenhagen, seeking to launch

simplification of the global political map, it isolates the incentives of the *E* state, which is the focus of the analysis. It can be thought of as one large entity such as the United States or the European Union that have the capacity to enforce international rules.

(ii) Strategies: This model assumes that *V* consume and the *E* punishes. Specifically, each V_i can consume from the common pool and further its depletion. Let $x_i \geq 0$ denote V_i's consumption. We interpret x_i as the increment of consumption above the maximum allowed quota set by international agreements. That is, we normalize $x_i = 0$ as the (potentially positive) allowable level of consumption. The enforcer state, *E*, does not consume above its allowed quota $x_i = 0$, it only punishes. *E* can punish V_i with any sanction that inflicts a cost of $c_i \geq 0$ on V_i.

(iii) Payoffs: V_i's payoff is its utility from consumption of x_i net of the cost of sanction c_i:

$$u_i(x_i,c_i) = f(x_i) - c_i$$

where $f(x)$ is V_i's instantaneous utility from its consumption, and it is assumed that $f(0) = 0, f' > 0, f'(\infty) = 0$ and $f'' < 0$, i.e., positive but decreasing marginal returns from consumption. We also assume that the sanction can potentially be greater than $f(x_i)$, which means that a state may end-up with negative utility, in which case the sanction is collected from its other resources.[3]

The enforcing state *E* derives negative utility from the Violator states' consumption **and** from punishing them. It is assumed that the disutility from consumption is linear, i.e., each unit of additional consumption by any V_i hurts just as much.[4] In order to punish V_i with a sanction c_i,

negotiations of a GCCT.

[3] This framework assumes that states are identical with respect to their utility function. Nothing in our results would depend on this simplifying assumption.

[4] We recognize that in the context of climate change, the negative effect from GHG is not linear. That is, there is a certain level of pollution that the atmosphere can sustain without triggering a negative reaction. After a certain "dose" of GHG, the harm can rapidly increase. This does not change the result on our model, as we can assume that we have reached a level of emissions where GHG emissions cause disutility

E has to incur a cost, *A(c_i)*, of the following structure:

$$A(c_i) = \begin{cases} 0 & if : c_i = 0 \\ \underline{A} + \alpha c_i & if : c_i > 0 \end{cases}$$

E incurs no cost if it decides not to punish. But any positive level of penalty involves a "fixed cost", $\underline{A} > 0$, and a "variable cost" of $\alpha > 0$ per unit of penalty.[5] The fixed costs stem from establishing a system to administer the process of punishment (*i.e.,* the basic process of verifying non-compliance and obtaining authorization to punish). Variable costs are a function of the extent and complexity of the violation. The imposition of sanctions levies an immediate cost on the sanctioning state. Thus, *E*'s payoff is:

$$U(\{c_i\},\{x_i\}) = -\beta\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} A(c_i)$$

where β is *E*'s constant marginal cost from each additional unit of consumption expended by *V* states. It is assumed that in the short-run, there is no resource depletion constraint, i.e., Violator states' combined consumption does not exceed the amount of available resource.

This payoff framework assumes that the *E* does not have the opportunity to consume from the common pool, and that V_i do not suffer from the depletion of the pool. Both unrealistic assumptions are not necessary, and both are made merely for analytical simplicity. In reality, *E* enjoys consumption but is willing to limit its consumption to the maximum level set by the

(and that we have not reached saturation levels of emissions where additional emissions would not further deplete the environment) and that it is thus socially desirable to eliminate additional emission of GHG.

[5] The assumption that the variable cost is linear is not material. The same result will arise if the

international standards if it can participate in a successful compliance mechanism; Similarly, each V_i may suffer disutility from the pools's depletion, but the disutility is more than offset by the utility it derives from consumption. Thus, we rely on the familiar Prisoners' Dilemma account in making the simplifying assumptions.[6] It allows us to focus the analysis on the realistic phenomenon where some states are eager to reduce their consumption of the common resource pool, despite the economic burden it entails, whereas other states are eager to continue their consumption, despite the harm they also experience in the long run from the depletion of the common pool. Thus, we can isolate the problem of how a party with an enforcement motivation can implement cooperation among the non-complying parties.

The timing assumptions of the game are as follows. Initially, every V_i makes its choice of x_i. E has then to decide whether to sanction. It is assumed that the sanction can have an "incapacitating" effect, i.e., it is in the power of E not only to hurt V_i, but also to suspend the consumption.[7]

B. Equilibrium

(i) Nash Equilibria[8]

The game has many Nash equilibria of the following character: E imposes no sanction on

variable cost decreases marginally.

[6] For an introduction and discussion of the Prisoners's Dilemma game in the Game Theory literature see Nalebuff and Dixit, *Thinking Strategically* (1991), pp. 11-14; Luce and Raiffa, *Games and Decisions* (1957), pp. 94-102.

[7] This assumption does not imply that one state can infringe on the sovereignty of another. It merely allows to reduce the model to a "one-shot" interaction. In a repeated play setting this feature does not have to be assumed, and instead it is derived endogenously.

[8] Nash Equilibrium is the most commonly applied solution concept in Game Theory. It is an assignment of strategy to each player, such that given that every other player conforms to his assignment, no player has an action that yields a higher payoff. For a formal analysis see Fudenberg and Tirole, *Game Theory* (MIT Press, 1991), pp. 11-14.

V_i if this state chooses $x_i = 0$. *E* imposes a sanction on V_i if this state sets $x_i > 0$; each V_i chooses $x_i = 0$. This is an equilibrium, because given the threat by the *E*, it is in the best interest of each V_i to choose zero consumption; and given that all V_i choose zero consumption, the *E* does not have to bear any disutility of punishing. There are many sanctions with which these equilibria can be supported, and they all share the feature that, in equilibrium, they are not carried out.

The problem with this family of equilibria is its credibility. The threat of sanctions by *E* is successful in deterring violator states only when the violators deem it credible and hence believe that their noncompliance would generate a punishment. However, this threat will not be credible if it is not in *E*'s interest to carry it out. Some consumption by V_i is cheaper to bear than to punish. *E* may be better-off tolerating the violation than imposing the sanction, given the cost of punishing. In the model, this is because there is a discrete cost *A* to levying even a minor sanction. As long as the negative effect of V_i's consumption is small, *E* prefers to tolerate it. Thus, we need to restrict our attention to the "credible" equilibrium.

(ii) Sub-Game Perfect Equilibrium[9]

Any player V_i realizes that a threat by *E* to punish any level of consumption is not credible. V_i knows that if it chooses a relatively modest consumption level, *E*'s best interest is not to punish, thus save the cost of punishment, and endure the modest infringement. Formally, for every level of consumption *x* of a violator state, let $\hat{c}(x)$ be the "minimum required sanction" to terminate *x* (MRS), which is defined by:

$$\hat{c}(x) = f(x)$$

[9] Sub-game perfect equilibrium is a Nash Equilibrium that allows only credible threats. For a player's strategy to part of a sub-game perfect equilibrium, he must not have a better action at any stage of the game, no matter what the other players choose. See Fudenberg and Tirole, *Game Theory* (MIT Press,

Since $f(x)$ is the utility that V derives from consumption, the sanction has to be at least of that magnitude to convince the consuming state to suspend its consumption. In this sense, E can "unilaterally" terminate any V_i's consumption: it merely imposes the MRS.

The minimum cost that E has to bear in order to terminate a consumption of x is $A(\hat{c}(x))$. E will pursue a punishment if and only if:

$$A(\hat{c}(x)) < \beta x$$

Since βx is E's cost of allowing x consumption to continue, E will punish only if the cost of the MRS is less than the cost of continued consumption. Denote by $\hat{x}$ the "critical" level of consumption, i.e., the level of x that satisfies:

$$A(\hat{c}(\hat{x})) = \beta\hat{x}$$

This is the level of consumption at which it is just as costly for E to tolerate it as it is to terminate it.[10]

We can state the following proposition:

Proposition 1. *The unique sub-game perfect equilibrium is for E to punish V_i by $\hat{c}(x_i)$ if and only if $x_i > \hat{x}$, otherwise not punish at all; and for each V_i to choose $x_i = \hat{x}$.*

Proof. For strategies to be a sub-game perfect equilibrium, they must subscribe for each player an action which is optimal **from that point on**. For E, a choice of sanction has to be made once x_i is already determined. Thus it can condition its sanction on the actual x_i, and the only rational

1991), pp.92-96.

[10] Under the assumptions that $f'' < 0$ and $f'(\infty) = 0$ such a level must exist.

condition is the one prescribed by the equilibrium strategy. If $x_i > \hat{x}$, E would prefer to bear $A(\hat{c}(x_i))$ than βx_i.[11] If $x_i \leq \hat{x},$ then $A(\hat{c}(x_i)) \geq \beta x_i$ and E's optimal course of action is no punishment. Given this credible program, the best V_i can do is to exploit E's tolerance to the limit, and choose the highest x_i that will not bring about a punishment, which is $\hat{x}$. To verify that this is the unique sub-game perfect equilibrium, notice that if there were another equilibrium, in it E would have to choose a value different than $\hat{x}$ as its threat point. But if it threatens to punish consumption that is lower than $\hat{x}$ the threat is not credible; and if it is willing to tolerate consumption that is greater than $\hat{x}$ it is not optimizing, and could do better by punishing. Thus, there is no other equilibrium. Q.E.D.

Remarks. (i) In equilibrium, no punishment is carried out, and each violator state consumes $\hat{x}$, for a total consumption of $N\hat{x}$. Figure I depicts this equilibrium diagrammatically. How does the equilibrium outcome depend on the parameters of the model? The amount of consumption $\hat{x}$ depends on the costs of sanctioning. Other things equal, the greater is $\underline{A}$ (the fixed cost of a sanction) or α (the marginal variable cost of a sanction), the higher is each V_i's consumption. Figure II depicts the change of $\hat{x}$ that results from increase in $\underline{A}$ or in α. Similarly, the more benefit V derives from consumption, or the less cost E suffers, the higher is the equilibrium consumption. Figure III depicts this effect.

[11] As x_i increases above $\hat{x}$, both βx_i and $A(\hat{c}(x_i))$ increase, but βx_i increases faster, due to the assumption that $f'' < 0.$

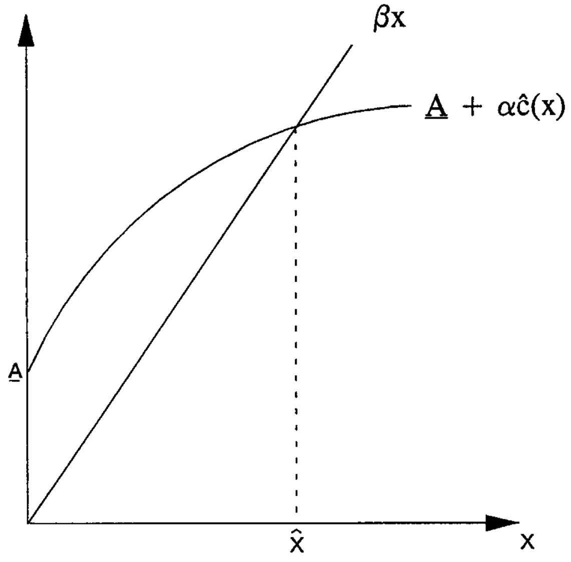

FIGURE I

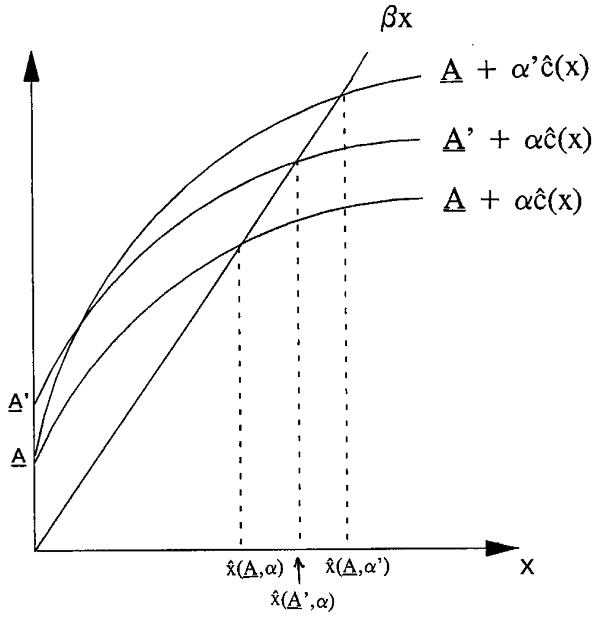
βx
$\underline{A} + \alpha'\hat{c}(x)$
$\underline{A}' + \alpha\hat{c}(x)$
$\underline{A} + \alpha\hat{c}(x)$
$\underline{A}'$
$\underline{A}$
$\hat{x}(\underline{A},\alpha)$
$\hat{x}(\underline{A},\alpha')$
$\hat{x}(\underline{A}',\alpha)$
x

FIGURE II

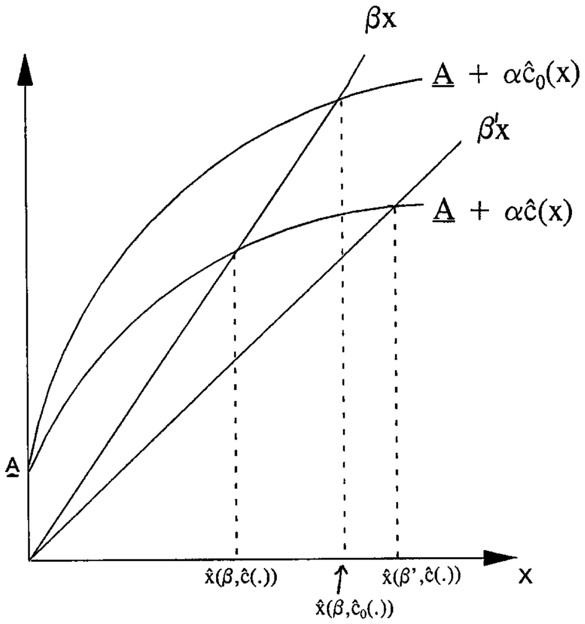
βx
$\underline{A} + \alpha\hat{c}_0(x)$
$\beta' x$
$\underline{A} + \alpha\hat{c}(x)$
$\underline{A}$
$\hat{x}(\beta,\hat{c}(.))$
$\hat{x}(\beta',\hat{c}(.))$
$\hat{x}(\beta,\hat{c}_0(.))$
x

FIGURE III

(ii) The fixed cost of penalty $\underline{A}$, which allows positive levels of consumption, was assumed to be an exogenous parameter. Alternatively, it may be plausible to assume that V can influence the magnitude of $\underline{A}$ strategically. Notice that V has an interest in inflating $\underline{A}$: the greater $\underline{A}$ is, the higher will be its equilibrium consumption and payoff. Thus, if V can apply measures that effectively increase $\underline{A}$, such as threatening to counter-punish E, the result would be to increase this state's consumption. Similar effects follow if V can influence the magnitude of α, the variable cost of the sanction.

II. Incentives for Compliance: Mechanisms of Rewards

A. General: “Stick” versus “Carrot”

The equilibrium analysis above was founded on the assumption that enforcement of international norms is accomplished through threats of penalties that an enforcing state imposes. As far as this state's “stick” reaches, so does the compliance of the potential violator states. This section considers a different, complementary, approach to enforcement: the “carrot” approach. According to it, E offers pecuniary rewards to each V_i in return for cooperation. If V_i defects, it is not sanctioned directly, but merely loses its reward.

There are two reasons why such an extension is worthwhile pursuing. First, it may be that in some areas the “stick” approach is not available: the enforcer state simply has no incentive or means with which to sanction a violator state. Second, and more interestingly, even if a sanction system is operable, a reward system can improve Pareto-efficiency: it may increase the payoff to every state. The Violators will not be worse-off than under the sanctions regime, and the Enforcer will be strictly better-off, an improvement that will induce the Enforcer to finance the reward system from its own self-interest. The basic reason is simple: a reward is a mere transfer, whereas a punishment involves real resources and a deadweight loss. The shift to a reward scheme avoids the deadweight loss and the saving can be divided among all. Further, we will show that a particular reward scheme can have the additional advantage of bolstering the credibility of the Enforcer's threat to levy sanctions.

We will assume throughout this section that *E* continues to stipulate the type of sanctions that were discussed in Section II. Thus, every violation of more than $\hat{x}$ would lead *E* to inflict a sanction, because it would be in *E*'s interest to sanction such a violation. However, we will restrict our attention to ways in which violations can be reduced below the equilibrium level of $\hat{x}$ where it is not in *E*'s interest to punish V_i.

The analysis proceeds as follows. First, we consider a simple reward mechanism, and examine the situations in which it works. Then we extend the analysis to examine a situation in which it is too costly for the Enforcer to reward Violators.[12] To induce compliance in this situation, we develop a model that combines sanctions and rewards: Violators are rewarded for complying; yet if they were to defect, the reward mechanism is employed to reward the Enforcer for punishing the Violator. This lowers the costs of enforcement and provides a more effective deterrent in the first place.

B. A Simple Reward System

Consider that *E* establishes a reward fund. It endows the fund with the cash amount of $Nf(\hat{x})$. Each V_i receives $f(\hat{x})$ from the fund if its consumption is 0 (i.e., does not exceed the maximum limit set by an international treaty), and receives nothing if its consumption exceeds 0. Such a system can lead V_i to comply (consume 0) as V_i gains nothing by deviating. However, for this system to work, it must be the case that *E* desires to participate in it. It costs *E* an amount of

[12] We assume that the Violators exhibit no interest in preserving the pool, and that the entire burden

$Nf(\hat{x})$ to keep this fund; it saves E $N\beta\hat{x}$ of damage that otherwise would have occurred. If $f(\hat{x}) < \beta\hat{x}$, then E is strictly better-off participating in setting-up the fund.[13]

Intuitively, this participation condition implies that the fund can be established if the benefit to any V_i from the consumption which it is willing to forgo is less than the benefit to E from eliminating this consumption. Only then can E "bribe" V to cease their consumption, and this outcome will increase total welfare. In some areas of environmental protection, it is plausible to assume that the benefit to the Enforcer from eliminating the consumption of the Violators is greater than the cost of lost consumption to the Violators. However, this may require both E and V_i to take a long-term view on the costs and the benefits of the consumption and preservation of the common pool (*i.e., E* not to apply a high discount rate in evaluation the value of preserving the common pool). If this was the ordinary situation, then it would be easy to achieve full compliance, as it would be strictly in the interest of the Enforcer to set-up this reward mechanism. All that this mechanism does is it eliminates the fixed cost of punishing. Once the fund is set, there is no cost for the enforcer state to withhold payments (in fact, the enforcer state enjoys a cost saving in the magnitude of the withholding). Absent a fixed cost, the credibility of the threat to punish even the smallest deviation from the agreed standard is restored, and full compliance is attained.

However, in many cases this mechanism does not work. If the value of lost consumption to V exceeds the benefit to E from eliminating this consumption, E cannot set a sufficiently endowed fund to implement compliance. This is likely to be the case when V cares more about

of generating the incentives to comply thus falls on the Enforcer.

[13] This also implies that E is better off using a reward to eliminate $\hat{x}$ than using a sanction. This is because $\hat{x}$ is defined as a level for which a sanction is too costly to impose. Only when $\hat{x}$=0 would it be the case that sanctions are less costly than rewards.

its short-term benefit of immediate consumption or when *E* applies a discount rate in assessing the detrimental future effects of *V*'s depletion of the common pool. This is also likely to be the case when *E* operates alone as an enforcer state, unable to induce all benefiting states to participate in setting up the fund and contributing to it. Note, that it may still be socially valuable to eliminate consumption: once we count the presumed long-term benefit to violator states from preserving the pool (which they currently discount to zero or close to zero), the total benefits can exceed the cost of lost consumption. However, as *E* needs to pay more than $N\beta\hat{x}$ to all *V* states, it will not participate. We turn, therefore, to examine a reward mechanism that can resolve this difficulty.

C. Rewarding the Punisher

One way to overcome the problem of *E*'s inability to fund the full economic loss of *V* states is to lower *E*'s costs in participating in the fund. This requires two modifications to the reward scheme. First, the scheme has to include rewards *and* sanctions. V_i are offered a reward for compliance; if they fail to comply and continue the violation (namely, set the level of consumption above zero), they lose the reward and will be inflicted a costly sanction. Second, If *E* has to impose a sanction, the cost of the sanction would be financed, at least in part, from the resources of the fund that were intended to reward V_i and that have now been freed as a result of V_i's non-compliance. We call this compound reward mechanism a system of "Rewarding the Punisher": V_i are rewarded for complying; if they still choose to defect, *E* is rewarded for punishing them.

In order for this Rewarding the Punisher mechanism to work, *E* must be given a sufficient incentive to punish. The reward fund that is set up to compensate V_i can potentially furnish such incentives for *E*: since a defection by any V_i frees some of the fund's committed resources, those can be utilized to finance the cost of punishment for the enforcer state. Consequently, the sanction that has to be imposed on a Violator is relatively small: it has to equal the Violator's gain from defection, which equals its value from consumption **minus** the value of the foregone reward. We show, perhaps surprisingly, that when a defection by some V_i leads not only to the deprivation of the reward, but also to an imposition of a penalty by *E*, more defections can be deterred than under a mechanism relying on simple rewards or sanctions alone.

Before we move on to demonstrate the idea formally, we illustrate the logic of the Rewarding the Punisher mechanism through a simple example. Assume that V_i enjoys a benefit of $1000 from violation and *E* suffers a harm of $700 from violation. Assume further that *E* can punish V_i, but in order to inflict a sanction of $c *E* has to bear a punishment cost of $200+c. In this scenario, there is no credible threat to punish. A sanction has to be at least $1000, which would cost *E* $1200 to impose, far more than the $700 burden *E* suffers with continued violation. Similarly, *E* cannot offer any reward that would lure V_i to comply. Now, suppose that *E* sets up a fund of $700. If V_i ceases the violation, it would receive the full $700 from the fund. However, if the V_i refuses to cease the violation and continues with any positive level of violation, it would not receive any reward and instead will be punished by *E.* The fund would reimburse *E* for the cost of the sanction it imposes up to the amount of $700. To analyze V_i's incentive under this scheme, we first have to identify how severe would be the sanction that *E* would be willing to impose. Expecting to be reimbursed up to $700, the maximum sanction that *E* would have an incentive to inflict is $500. This sanction would cost *E* $500 + $200 = $700, exactly the amount he would get back as a reimbursement. Thus, turning to V_i's perspective, V_i has to choose between A) violation, which would entail a net payoff of $500 ($1000 of benefit from continued violation minus $500 sanction) and B) compliance, which would yield a payoff of $700, directly from the reward fund. In this scenario, V_i would choose compliance.

It can also be shown, within this numerical illustration, that *E* does not have to endow the fund with a full $700, representing the entirety of *E*'s surplus from compliance. *E* can induce full compliance by V_i with a lesser fund. Here, the minimum necessary fund would be $600. This would enable *E* to threaten a sanction of only $400, which in turn would make V_i's payoff from violation equal $600 ($1000 benefit minus $400 sanction). This is exactly the payoff for V_i from compliance. Thus, any reward offered to V_i strictly exceeding $600 would make V_i better off complying.

To demonstrate this same insight more formally, we begin with the "easiest" case in which *E* contributes to the fund the maximum amount that is consistent with its incentives—*E*'s entire benefit from compliance. Because this can be thought of as unrealistic, we will subsequently show that similar deterrence can be obtained with a smaller contribution by *E*.

Suppose the fund is endowed by $N\beta\hat{x}$, the highest endowment *E* is willing to make. Each V_i is rewarded by $\beta\hat{x}$ if complies, but as we are working under the assumption that $\beta\hat{x} < f(\hat{x}),$ we know that V_i has the incentive to defect. That is, we are assuming that the benefit to *E* from eliminating consumption is less than the benefit to V_i from continuing violation (or else, a simple reward mechanism discussed in the previous section would suffice.) Thus, V_i would prefer to continue violation and forgo the reward. To eliminate this incentive to defect, *E* can threaten to punish V_i by not only depriving V_i of the reward, but also by inflicting a sanction of no less than $f(\hat{x}) - \beta\hat{x}$ on V_i. Denote this sanction by c^*. This is the size of the net gain to V_i from defection: the value of consumption $f(\hat{x})$ minus the cost of the forgone reward $\beta\hat{x}$. If the sanction is at least c^* — the size of this net gain—it can deter V_i from defection and induce it to accept the reward instead. Put differently, if V_i continues its consumption, its payoff is $f(\hat{x}) - c^*$, which is the value of consumption minus the net gain from defection. If, instead, V_i ceases consumption, its net payoff is $\beta\hat{x}$, the reward it will collect. To induce V_i to cease consumption it must be that

$$f(\hat{x}) - c^* \leq \beta\hat{x},$$

which implies that $c^* \geq f(\hat{x}) - \beta\hat{x}$.

We know from section II that this threat is not credible: it may be cheaper for E to ignore the deviation than to sanction it. This is because the total cost for E to fund both the reward and the punishment is $c^* + \beta\hat{x}$, which exceeds E's harm from an ongoing consumption of $\hat{x}$. To overcome this problem, the fund can offer to finance E's cost of punishment. The fund will reimburse E up to an amount of $\beta\hat{x}$, which the fund has available as a result of the violator state's deviation. As long as $c^* \leq \beta\hat{x}$, this mechanism will lead to compliance. That is, if the cost of imposing the necessary sanction c^* is less than the amount available in the fund to reimburse the Enforcer, the Enforcer's threat to impose the sanction is credible and the Violator's defection would be deterred.

In fact, it may not be necessary for E to contribute it entire benefit $N\beta\hat{x}$ to the fund. We can state and prove the following proposition:

PROPOSITION 3. *A mechanism of "rewarding the punisher" can implement full complianceE at a minimum cost of K^*, which is the solution to:*

$$A[f(\hat{x}) - K^*/N] = K^*/N.$$

Proof. First, we show that the mechanism works with a fund of $N\beta\hat{x}$, the largest fund E is willing to support. To deter a defection of $\hat{x}$, E has to apply a sanction of c^*. E has an incentive to pursue this penalty: it spends $A(c^*)$ to punish, but receives $\beta\hat{x}$ as a reward from the fund. Since $c^* < \hat{c}(\hat{x})$,[14] and since $A(\hat{c}(\hat{x})) = \beta\hat{x}$, it follows that $A(c^*) < \beta\hat{x}$, the reward for sanction exceeds the cost of it. Second, E can deter a defection of $\hat{x}$ with a smaller fund. Let K denote the magnitude of the fund, from which each V_i can be rewarded K/N.[15] To deter deviation, E has to impose a sanction which equals the Violator's net gain from continuing its consumption rather than taking the reward, $f(\hat{x}) - K/N$. The cost of applying this sanction is $A[f(\hat{x}) - K/N]$. The cash incentive which the fund can supply E to induce E to punish will be, at most, K/N. Thus, E will punish if and only if:

$$A[f(\hat{x}) - K/N] \leq K/N.$$

For high enough K, the right hand side will be greater than the left hand side.[16] As K decreases, the left hand side increases and the right hand side decreases, thus, the lowest fund that implements compliance, K^*, is the one in which this condition appears with equality. Q.E.D.

Remarks. (I) *Why Does the Mechanism Work?* This mechanism improves the credibility problem of the threat to punish small deviations by eliminating E's cost of carrying out the threat. Indeed, the fund's endowment is furnished by E but this investment is a sunk cost for E, at the time of the

[14] Recall that by definition, $\hat{c}(\hat{x}) = f(\hat{x})$. Since $c^* < f(\hat{x})$, it follows that $c^* < \hat{c}(\hat{x})$,

[15] We are assuming that all V_i are identical. The framework can easily be extended to capture V_i of varying sizes, each receiving a reward commensurate with the burden their idiosyncratic $\hat{x}_i$ inflict on E.

[16] A higher K implies a higher K/N and thus a higher value at the right hand side. At the same time, a higher K implies a lower $f(\hat{x}) - K/N$ and thus a lower $A[f(\hat{x}) - K/N]$, the value at the left hand side.

decision whether to punish. Since E bears no additional cost in punishing, it will always do so, and more defections are deterred. In a sense *E credibly threatens to punish itself if it does not punish even a small deviation,* hence bolstering the credibility of E's threat to punish and eliminating V_i's incentives to commit such deviations.

(ii) *The Number of States.* As the number of violator states rises, the fund endowment must rise proportionally, such that the per-state reward remains unchanged. This implies that there are no scale effects — cooperation among a larger group of states requires solely a proportionately larger fund, without changing the per-state incentive scheme. In other words, no generality is lost if one views the game as a set of separate interactions, each between E and one of V_i; a state's behavior does not affect others.

(iii) *Extensions*. [COMPLETE] Several factors may be added to the analysis. First, the interaction between states may be repeated across time. This would introduce reputational concerns and enhance the incentives of the large state to utilize sanctions. Second, the parties' information may be incomplete. The enforcer state, for example, may have only incomplete information about the extent of the violations and the identity of the violator. This feature would generally impair the operation of any incentive device. Third, there are various types of sanctions that can be employed, varying by their cost to the enforcer state and the amount of social waste they generate. [Discuss also the applicability outside the commons problem. The model is not limited to the commons issue but also is not a general model for all international enforcement. This is e.g., different in a club good situation.]

III. Application: international cooperation on Climate Change

To illustrate how the model can be applied in practice, we discuss how the combination of sanctions and rewards can be employed to enhance participation in, and compliance with, the GCCT. Consistent with the framework of the model, we first examine the Enforcer's ability to enforce the GCCT through a simple punishment mechanism. As the model predicts, this mechanism, if at all feasible, can effectively deter only the most egregious violations. Any violation below a certain level ($\hat{x}$) remains undeterred because it remains too costly for the Enforcer to inflict a sanction. Next, we discuss situations where a simple reward mechanism can induce compliance with the GCCT. We conclude that this mechanism is superior in comparison to reliance on sanctions but will fail in the absence of adequate incentives to contribute to the fund. Finally, we examine the potential of the Rewarding the Punisher mechanism to furnish the incentives for the Violators to comply with the GCCT.

A. Why Climate Change Cooperation is Difficult?

There is a broad agreement among states that international cooperation to fight climate change is necessary; climate change is a genuine global threat that cannot be solved by any nation alone. States have gradually acknowledged that that they must, collectively, reduce the total quantity of GHG that they emit into the atmosphere. Yet consensus on how to allocate this responsibility among states is missing. No state wants to bear a disproportionate cost of cutting down its emissions while allowing other states to continue to deplete the atmosphere. This tension has been at the center of the efforts to pursue international cooperation on climate change, complicating the efforts to overcome the collective action problem.

Several possibilities to allocate responsibility for GHG emission cuts exist. For instance, states could agree to freeze their emissions on current, or some historical, baseline of consumption.[17] The Kyoto Protocol adopts this method by mandating developed countries to freeze their GHG emissions at the 1990 consumption level. This method of dividing responsibility, however, is controversial as it benefits states that have thus far contributed most to climate change, granting them the highest quota of permitted emissions.

Another possible option to allocate responsibility is to adopt a uniform global tax on GHG emissions or a global cap-and-trade system that obliges all GHG producers to pay an equal amount of their emissions. This follows the “polluter pays” principle. However, developing countries in particular oppose the idea of assigning responsibility equally among states. They maintain that developed countries have historically contributed more to climate change than developing states have. Developed and developing countries also differ in their respective economic and technical capacity to tackle global climate change. This, according to the developing countries, justifies that developed countries bear the primary responsibility of reversing the climate change.[18] However, many developed countries, including the United States, insist that they will not agree to reduce their emissions if developing countries are exempted from the similar responsibility.

The Kyoto Protocol gave in to the developing countries’ demands and exempted them from any obligation to reduce their GHG emissions. The failure to bind developing countries into

[17] See Kyoto, Lieberman-warner bill

[18] This view rests on the familiar international law principle of “common but differentiated

global emission reduction targets led to the refusal by the United States to ratify the Protocol. Thus, too many significant Violators were left outside of the effective treaty framework, compromising the effectiveness of the Kyoto Protocol. Without the full participation of the developing countries—the fastest growing GHG emitters—a post-Kyoto GCCT will do little to halt climate change. Looking forward, the greatest challenge is therefore to entice these Violators to sign onto, and comply with, the GCCT.

Thus, the efforts to negotiate a GCCT face two hurdles: First, Enforcers need to entice the Violators to sign the GCCT. Second, once the Violators have signed the GCCT, Enforcers need to secure their compliance with the GCCT. While we acknowledge that these two stages—participation and compliance— present separate obstacles for effective cooperation, we do not distinguish them analytically in the discussion. Instead, for simplicity, we assume that Enforcers seek to employ the threat of sanctions or the promise of rewards to induce participation in the first place and, resort to the same strategies to induce compliance once the participation has first been secured.

B. Enforcers and Violators

States participating in GCCT negotiations consist of Enforcers and Violators. We assume that any given state's status as an Enforcer or a Violator is determined by both public welfare and public choice considerations. The higher the costs of experiencing climate change and the lower the cost of fighting it, the more likely the state is to assume the role of the Enforcer. Similarly, the more influential the pro-environment lobby and the more marginal the counter-lobby of the carbon-intensive industries in the state is, the more likely the state is to

responsibilities".

assume the role of the Enforcer. When the reverse conditions dominate, the state is expected to assume the role of the Violator.

States' energy infrastructure largely determines the state's cost of fighting climate change. Some states do not have carbon-intensive economies. They are not major producers of fossil fuels (including coal) or energy-intensive products (including aluminum, steel, iron, cement, glass and chemicals). Due to their low carbon/GDP ratio, they would not suffer major costs in moving away from reliance on fossil fuels. France is an example of a country that relies on nuclear energy and that faces relatively low costs of switching to low-carbon economy as a result. These states are likely Enforcers of the GCCT. Other states' economies are carbon-intensive based on their high carbon/GDP ratio. States with larger coal reserves, including China, United States and Australia, would face high costs if required to rein in their consumption of fossil fuels. These states are the potential Violators of the GCCT.

In addition, various states experience the threat of climate change to a different extent or within a different timeframe. Most threatened are small island nations which will become uninhabitable, if at all survive, the changing temperatures caused by climate change. These nations are vocal yet largely powerless proponents on tough international commitments to cut emissions. They support the GCCT but have little capacity to enforce it against larger Violators. Other states will feel the adverse effects of climate change with delay, reducing their sense of urgency in taking decisive measures. This makes them likely Violators.

Similarly, political support for taking action to tackle climate change varies across states. Some states, most prominently the member states of the European Union, are committed to taking action to curb their emissions, predominantly because of the heightened domestic awareness of the dangers of climate change. The political support within Europe is explained by active environmental NGOs and the participation of green parties in many coalition governments. In other states, pro-environment lobby is less influential. Instead, powerful interest groups that stand to lose from tough emission standards dictate the political agenda, causing the state to resist efforts to the flight against climate change. This is one reason why the United States has thus far refused to assume international commitments to reduce its GHG emissions.

Based on these considerations, the most prominent supporter of the GCCT is the European Union.[19] European countries seek ambitious international commitments to rein in GHG emissions and have the incentive to enforce the GCCT against Violators. Under the new administration, it is plausible that also the United States will sign onto the GCCT. However, many in the United States Congress insist on making their approval of the treaty conditional on China's participation in the GCCT. We therefore make an assumption that the United States agrees to cut its GHG emissions only if China agrees to do the same. This idea is consistent with the Waxman-Markey climate change bill that is currently pending before the Congress. [20]

An important group of Violators consists of developing countries. Developing countries argue that they have the right to industrialize (=pollute), just like developed countries did throughout the history of industrialization. The common argument among Violators is that climate change is caused by enforcers who therefore should take responsibility of fixing it.

[19] Members of the European Union are developed countries many of which have a (relatively) low carbon footprint and thus lose less from cutting their emissions than the United States or China, for instance, would. Many European countries, including in particular low-lying countries like the Netherlands, will feel the direst consequences of the rising sea levels and thus have most the gain from curtailing emissions. Numerous European countries also have a strong pro-environment domestic lobby and a relatively weak counter-lobby from carbon-heavy domestic industries.

Second, due to their lower level of development, violators are not in a position to bear the costs of technological change and industrial transformation that mandatory emission cuts would require of them. A prime example of a Violator is China. Last year, China overtook the US as the largest emitter of GHG in the world. Further, China's emissions are growing at a 4.2% rate per year because of its thriving economy.[21] China's comparative advantage in international manufacturing is partly based on low energy costs due to its enormous coal reserves. China's benefit from continuing emission of GHG is therefore high. Given the high costs of switching to alternative fuel sources, China remains reluctant to assume any international commitments that would force it to do so. And should China yield to the international pressure to sign onto such commitments, it would retain an incentive to defect from the agreement given the economic benefits of continuing to rely on coal in fueling its economy.

There are three primary reasons why the United States would seek to enforce the GCCT against China. First, the United States knows that climate change is a global problem that needs a global solution. A ton of CO_2 contributes to the global warming regardless which country emits it.[22] It is therefore particularly crucial that countries with the highest rates of emissions, including China, assume and uphold international commitments to cut down their GHG emissions. The United States knows that a GCCT without the participation of developing countries remains ineffective given that developing countries combined are projected to account for two thirds of global CO_2 emissions in the course of this century.[23] Second, if China stayed outside the GCCT, the United States fears that the carbon-intensive production from developed

[20] Cite and refer to the provision on border tax adjustment.

[21] Michael P. Vanderbergh, *Climate Change: The China Problem*, 81 S. Cal. L. Rev. 905, 914 (2007–2008).

[22] Frankel 7.

countries may migrate to China or some other “pollution haven”. This is a phenomenon known as “leakage”. Leakage means that global emissions would not be reduced; they would merely be shifted from the jurisdiction of Enforcers to that of the Violators. Third, the United States fears that its domestic energy-intensive industries will be placed at competitive disadvantage compared to similar industries in China. If Chinese industries are allowed to continue their production methods, US aluminum, steel, cement, glass, paper and iron industries will not be able to compete against their Chinese counterparts due to the soaring production costs in the US.

The model allows for the (realistic) possibility of several Enforcers and several Violators. However, for simplicity, the below discussion will focus on the United States as a single Enforcer and China as a single Violator. Analytically, we can replace the United States in the discussion by European Union. We nevertheless choose to focus on the United States because of its motivation to condition its own participation on the GCCT on its ability to enforce the GHG emission cuts against China.

A. Punishment Mechanism

What Penalties Can Be Used?

While the enforcement motivation for the United States is clear, the avenues available for pursuing enforcement are less obvious. It is plausible, yet highly doubtful, that an international enforcement mechanism will be incorporated into the proposed GCCT.[24] However, such a mechanism would not be available if China in the end refused to sign the GCCT. International treaties rest on the principle of state sovereignty; a state can only be bound by a treaty to which it

[23] IFA, EIA

[24] This would present a significant improvement to the Kyoto Protocol, which lacks such a mechanism.

consents. An enforcement mechanism tied to the GCCT would thus not be available vis-à-vis non-signatory states. Also, if the GCCT incorporated an enforcement mechanism, Violators would be less likely to sign it in the first place.

Given the limits of any conceivable multilateral enforcement mechanism, the United States and many European Union member states have suggested that they will use trade sanctions as way to compel even the non-signatory states to reduce their GHG emissions. In essence, their proposal is to impose a carbon border tax on products that are imported from countries that do not sign onto the GCCT or assume comparable commitments domestically. If a certain state fails to charge its domestic producers for their GHG emissions (*i.e.,* does not force them to buy emission permits or pay a carbon tax), the price of those products would be "adjusted" to reflect the price that similar domestic products bear once the foreign products cross a border.

To illustrate this, consider that the United States signals that it is prepared to impose a carbon border tax on every Chinese carbon-intensive product P_i that Chinese manufacturers import into the US market. P_i is expected to be cheaper to produce in China than in the United States, given that the United States manufacturers are obliged to internalize the cost of carbon they emit in the manufacturing the similar product. Thus, Chinese manufacturers derive a benefit (x_i) by excluding the cost of carbon that they emit in their production of P_i. To offset this benefit, the United States imposes the carbon border tax, which inflicts a cost of $c_i \geq 0$ on China. Note that the border tax is only effective if it exceeds the benefit China derives from polluting (x_i). Thus, the minimum effective carbon border tax would need to be equal to the amount that China saves in producing P_i though carbon-intensive production measures. [25]

The logic of this kind of sanction is that it would equalize the competitive situation between a domestic producer that is subject to GHG emission caps or a domestic carbon tariff and the foreign producer that is exempted from a similar required on its home market. In other words, the carbon border tax would eliminate the unfair comparative advantage of a country that fails to internalize the price of emitted carbon in the pricing of the product it exports. It would also mitigate the leakage problem as producers that would have relocated to "pollution havens" would not be able to export their products to countries regulating emissions as they would similarly be subject to a carbon border tax. This idea is reflected in the various bills pending in the United States Congress, including the Lieberman-Warner Bill and Waxman-Markey Bill, both of which contain a provision calling for a carbon tax to be imposed at the US border on goods that originate from countries that do not limit the carbon-content of their products.

For the purpose of the below discussion, we assume the carbon border tax to be a principal mechanism on which the United States relies in seeking to enforce the GCCT norms against China. The United States believes that the carbon border tax would be beneficial in two ways. First, the mere *threat* of its imposition may induce China to assume the commitments embedded in the GCCT. Thus, this enforcement mechanism would potentially incentivize China to sign the GCCT. Second, if China failed to sign the GCCT or—if it failed to comply with the GCCT despite its decision to sign it—the threat could be carried out by actually imposing the

[25] Remember that the minimum required sanction that the United States must impose to entice China to limit the carbon-content of its products China's is $\hat{c}(x)$. Since $f(x)$ is the utility that China derives from allowing for unlimited GHG emissions, the sanction has to be at least of the magnitude of $f(x)$ to convince China to curtail its consumption.

carbon border tax. Thus, unlike most international enforcement mechanisms, this mechanism would be available irrespective of China's final decision to sign or not to sign the GCCT.

2. Why Penalties Are Costly to Punishers

Several costs are embedded in the system of applying a carbon border tax. The border measure would require the United States to set up an office to administer the border tax, or to devote new resources to the existing United States Customs and Border Protection agency. This would impose a non-trivial fixed cost on the US. Further variable costs would stem from the actual evaluation of the carbon-content of each product P_i, $I = 1,2,...N$ on the border. This evaluation would be necessary to determine the specific amount of the border tax in each case. This would be a non-trivial task given the potentially high number of trading partners that fail to impose a price on carbon and that export a high number of products to the United States. The evaluation of the carbon-content of any given P_i would be particularly difficult if there is no information on the production methods used to produce P_i. Determination of the carbon content may also be contested, leading to a need to defend the measure in the WTO Dispute Settlement Mechanism.

It is also unclear if the carbon border tax is consistent with WTO rules. The United States may be required to defend the entire border measure in a costly litigation before the WTO, uncertain of its success. At worst, if the United States loses the dispute, the WTO can authorize the United States trading partners to retaliate against the United States exporters. Thus, a significant potential cost of imposing a carbon border tariff on Chinese products is a threat of retaliation from China. This is a primary reason why some United States exporters have opposed the idea of imposing border measures to seek compliance with the GCCT. At worst, measures against China may escalate into a trade war, a prospect that would involve significant costs. On the other hand, were the United States to win the potential WTO litigation and thus be given green light to impose a carbon border tax, the ruling would also give the United States' trading partners a permission to impose a similar border adjustment measures vis-à-vis US producers, which they determine to fall short from their obligations under the GCCT.

Different type of costs stem from the rise of the price of Chinese (carbon-intensive) products in the United States. Even if import-competing industries (the producers of similar, non-carbon intensive products) benefited from "adjusting" the price of the Chinese goods on the border, the price that the consumers in the United States must pay for the Chinese products increases. Similarly, the United States manufacturers that use Chinese products as raw material or inputs suffer from having to internalize the border tax in the pricing of their final product.

To be sure, the United States would also benefit from the imposition of a border tax. Reducing the demand of the Chinese products on the United States market would help United States manufacturers compete against the Chinese. This would preserve domestic employment opportunities that would vanish if the Chinese products enjoyed an unfair comparative advantage by producing carbon-intensive goods. Presumably, the leakage problem would also be mitigated, bringing important benefits in terms of reduced overall emissions. Finally, the border tax would also generate direct revenue to the United States government. This revenue could be invested in the United States domestic economy.

1. The Limits of Sanctions

The numerous uncertainties and complexities embedded in the system of carbon border tariff reduce their attractiveness as a sanction. Yet even if they were feasible to administer, the ultimate limitation for their use stems from the costs the United States incurs in inflicting a sanction. As the model predicts, given the costs of punishing, carbon border tariffs are effective in deterring only China's most egregious violations while failing to deter the more moderate (but nevertheless costly and inefficient) levels of violations. Knowing this, China can exploit the United States' toleration limit by choosing the level of GHG emissions that falls just below the level that would trigger a sanction.

Exactly when is it in the United States' interest to impose the carbon border tax? The model predicts that the United States will punish China by $\hat{c}(x)$ only if the harm caused by China's GHG emissions (x) exceeds the cost of imposing the border carbon tariff on China $(\hat{x})$.Knowing this, China can choose the level of emissions that inflict a harm on the United States which is equal to the cost the United States bears when punishing China $(x_i = \hat{x})$. This way China can exploit the United States' tolerance limit simply by choosing the highest level of GHG emissions that will not trigger a carbon border tariff because the cost of imposing such a tariff will exceed the benefit from eliminating China's emissions.

To do this, China can either reduce the extent of harm the United States suffers from China's emissions or, alternatively, to raise the costs the United States incurs when inflicting a sanction on China. To reduce the overall harm the United States suffers, China may, for instance, limit its GHG emissions in certain sectors of the economy in order to bring down its overall level of emissions while allowing some other sectors—in particular those that are the key to its economic growth— to continue polluting. China may also selectively reduce its emissions only in export-oriented sectors in order to limit its exposure to border measures yet retain an overall level of GHG emissions that continue to deplete the atmosphere.

Alternatively, China could remove the United States' incentives to inflict a sanction by raising the United States' costs of carrying out the punishment. It might, for instance, limit its export of steel, glass and other carbon-intensive products yet continue to export so called "carbon-derivatives"—products whose carbon content is difficult to determine. It is more costly for the United States to administer a carbon border tax on products whose carbon content is complicated, if at all possible, to determine. Finally, China can divert trade in its most carbon-intensive products to markets which it does not expect to impose border measures on its products. Any of these measures would reduce the effectiveness and hence the attractiveness for the United States to administer a costly system of punishment through border tariffs.[26] Finally, China knows that it is much costlier for the United States to impose trade sanctions on China than to impose the same sanctions on smaller trading partners, since China can inflate the costs of the sanctions by limiting the access of United States companies to its market. A concrete threat of counter-punishment alone may lead the United States to hesitate pursuing its sanctioning strategy.

It follows that the United States sanctions China only when the GHG emissions are particularly high or when the costs of imposing sanctions are particularly low. But when the gains (reduced GHG emissions, restoring competitiveness, preventing leakage, tariff revenue

[26] China could also seek to mobilize interest groups within the United States, assuming those interest groups benefit from China's non-compliance (i.e., manufacturers who use Chinese inputs).

etc.) do not offset the costs (administering punishment, higher prices for United States consumers, threat of litigation and counter-retaliation against United States exporters), the United States will not find it rational to punish Chinese importers with a border tax. This leads us to explore the possibility of luring China into compliance through the system of rewards.

B. Reward Mechanism

After showing the limits of sanctions, the model explored the option of Enforcer establishing a reward fund. The key insight of the model is that the reward fund does not only induce the Violator to comply with the GCCT. It can also make the Enforcer better off. A simple system of rewards can eliminate the cost of punishment by offering a transfer payment for the Violator in return for its compliance. For instance, the United States could offer China cash compensation in return for China's commitment to limit its GHG emissions to the level set in the GCCT. Alternatively, in situations where it is too costly for the United States to "buy" China's cooperation, compliance could be achieved through the "Rewarding the Punisher" mechanism: The United States would pledge to reward China for its compliance. However, if China failed to comply, it would lose its expected reward. This lost reward would then be transferred to the United States, which could use the reward to finance the costs of punishing China. Compensating the United States for its costs of inflicting a punishment would make its threat to do so credible. Thus, ultimately China, or any other potential violator, can be deterred by a combination of sanctions and rewards.

1. Setting Up a Reward Fund

The idea of rewards has been discussed in connection with the GCCT. Recently, the British Prime Minister Gordon Brown proposed the creation of an international fund to help developing countries adopt clean technologies and thereby comply with the requirements of the GCCT. According to this proposal, developed countries would provide $100 billion per year by 2020 to finance projects to cut down emissions in developing countries. Brown suggests that the fund could be raised partly by taxing flights and shipping. Thus, instead of exempting the developing countries form the GHG emission cut targets, Brown's proposal calls for developed countries to finance developing countries' efforts to comply with those targets. This proposal falls short of the developing countries' estimate of what they would need to comply with the GCCT. Still, it rests on the idea that transfer payments (of a contested amount) would flow from developed countries to developing countries as a "reward" for their compliance.[27] Similarly, Jeffrey Sachs has called for two global trust funds to be established: A mitigation fund that would offer transfer payments for the purpose of adopting new emission technologies and a technology transfer fund that would provide poorer countries access to (often IP-protected) technologies that can be harnessed to reduce countries' GHG emissions. Sachs has stressed that the setting of the fund would require donor countries to commit approximately 0.5 % of their GNP to the fund. This would amount to $170 billion dollars annually which, according to Sachs, could be directed to recipient countries to compensate them for their efforts to mitigate climate change.

Enforcers would be expected to make the cash compensation conditional on Violators using the transfer payment for agreed upon purpose. The fund could, for instance, require Brazil

[27] The idea of directing technical assistance to developing countries is also consistent with a longstanding principle of "differential responsibilities" that developed and developing countries ought to have under

to undertake measures to reverse deforestation by paying farmers to leave forest intact or provide them alternative farming land; India and China could be required to build plants that are using the carbon sequestration technology or to invest in renewable energy technologies.

There are several possibilities on how any given reward fund could be set up. The resources to the fund could be supplied ad hoc, as the need arises to extend a reward to a potential Violator. Alternatively, commitments could be sought from Enforcers upfront. A reward fund where the funds have been committed upfront—conditionally on securing other Enforcers' commitments—would have the advantage of mitigating likely free-rider problems: either every Enforcer contributes to the fund or no Enforcer contributes. Ex ante uncertainty regarding the exact nature of violations and identity of Violators may induce Violators to participate in this situation. Irrespective of the chosen timing of committing the funds, Enforcers would need to credibly pre-commit the funds either to reward the Violator or, in case the Violator is deprived of the Reward, to reward the Punisher for inflicting a sanction on the Violator.

2. Who to Reward and How Much?

A key question in setting up the fund is to decide *who* would reward *whom, when* and by *how much*. The fund would be most successful if it was collectively financed by numerous Enforcers, including the United States and the European Union as most prominent Enforcers. Each potential Enforcer could reduce its own upfront commitment by convincing numerous other Enforcers to participate as well. Enforcers' relative commitment levels could be determined according to their relative responsibility (based on *i.e.,* total or historical emissions) or their

international law.

capacity (GDP) to finance the fund, the latter being less contentious in not requiring countries to resolve the difficult question of who bears the responsibility for climate change. Adjusting the contributions according to Enforcers' GDP would also approximate the practice of other international organizations, including the International Monetary Fund or the World Bank.[28]

Violators would be the beneficiaries of the fund. These would be the countries which are economically dependent on outside funding in order to finance the technological change required to meet their emission reduction targets. The amount of the reward that the Violator would be entitled to would depend on the resources required to meet the emission reduction target set by the GCCT, adjusted by an (e.g., GDP based) estimate on the extent to which a country could finance these costs without outside assistance.

Decision on whether to reward a Violator requires establishing the metric for compliance. This would be set by the GCCT, and would most likely consist of a numerical target (i.e., reduction of total GHG emission by 20%) against some baseline (i.e., benchmark being the country's 1990 emission levels). Alternatively, in the absence of countries' ability to agree on absolute targets, they could agree to reduce the "emission intensity" of production, as China has suggested. This more modest commitment would require countries to cut their emissions per unit of GDP. Finally, absent the ability to agree on numerical targets, countries could agree to adopt a set of domestic policies aimed at reducing countries' emissions. While the need of a benchmark is equally important in establishing compliance in the sanctioning system, Violators can be expected to sign into more ambitious commitments if they know that they can merely lose a reward (as opposed to be subject to punishment) if they fail to comply.

[28] Note that the IMF for instance sets the countries' quotas based on a weighted average of GDP (weight of 50 percent), openness (30 percent), economic variability (15 percent), and international reserves (5

To evaluate whether Violators have adhered the emission reduction targets established the GCCT, Enforcers need information on the Violators' actual emission levels. This is, inevitably, difficult and costly. Without detailed knowledge on the production processes employed in different countries, it is difficult to know precisely whether countries are complying with their obligations under the GCCT. Thus, part of the reward fund's fixed costs would need to be directed to monitoring and reporting the emissions in order to determine whether country is meeting its target and whether it therefore remains entitled to its rewards.

3. The Limits of the Fund's Success: Violations That Are Too Costly to Bribe

In addition to the difficulties stemming from measuring and monitoring compliance, the reward fund will not provide a solution for situations where it costs more for Enforcers to set up the fund than it costs them to endure Violator's continuing emissions. We learned that only when the benefit the Violator derives from the consumption of the common pool is less than the benefit the Enforcer derives from eliminating Violator's consumption, the Enforcer can "bribe" the Violator to cease its consumption. For instance, if China benefits less from emitting GHG into the atmosphere than the United States benefits from eliminating those emissions, the United States is better of paying China to cease its consumption through the system of simple rewards. However, if the cost of financing the reward fund exceeds the benefit the United States derives from gaining China's compliance ($f(\hat{x}) > \beta\hat{x}$), it will not set up the reward fund. Thus, the largest fund the United States is willing to support is one which equals the cost China inflicts on the United States by continuing its consumption of the common pool. This would be likely in particular if the United States applies a high discount rate in assessing the detrimental future

percent).

effects of China's GHG emissions or if the United States cannot convince other Enforcers to share the burden of financing the fund. Similarly, if the United States believes that China is gradually agreeing to reduce its overall GHG emissions (following a recent statement by the President Hu Jintao that China intends to curb the intensity of its emissions even if it won't agree to absolute cuts in emissions), the United States may conclude that in the presence of even limited cooperation on the part of China, the reward system would be too costly when compared to the harm caused by China's continuing (more controlled) emission levels. In other words, the limits of the simple reward system mirror the limits of the sanctioning system: sanctions or simple rewards can only deter most egregious violations of the GCCT but leave socially undesirable levels of emissions undeterred.

C. Rewarding the Punisher

1. Leveraging Less Money to Generate More Incentive

An obvious objection to the simple reward system is the distrust that Enforcers would agree to finance the fund. The model shows that this is likely to be the case each time Enforcers find that the cost of establishing such a fund would exceed the benefit they derive from gaining Violators' participation in the GCCT ($f(\hat{x}) > \beta\hat{x},$). The model introduces a possible solution for this problem. We may be able create incentives for Enforcers to create the reward fund in the first place if the introduction of the system of rewards can be combined with the credible system of sanctions. This way, the technological assistance the United States would pledge to China would be conditional on China using the assistance to lower its GHG emissions. More importantly, if China failed to comply despite the offer of assistance, the United States could withhold the assistance and turn to the fund itself to lower its costs of enforcing the GCCT against China. At this point, the United States would have no incentive to refrain from punishing China; the funds are pre-committed and thus present a sunk cost for the United States. Knowing this, the United States can make a credible threat to punish at a lower cost and China has a greater incentive to comply even when offered a lower reward in return for its compliance. Thus, under the Rewarding the Punisher mechanism, the United States needs to leverage fewer funds ex ante to create the incentives for China to comply with the GCCT.

2. Challenges of the Rewarding the Punisher Mechanism Fund

The Rewarding the Punisher mechanism would present several challenges that the Enforcers would need to solve before setting up the fund. First, the problem of measuring emissions and monitoring compliance, discussed in connection with the simple reward mechanism, would equally apply to the Rewarding the Punisher mechanism. In order to know how to set the level of reward and how to determine whether a Violator is entitled to that reward, a mechanism of measuring, reporting, and verifying compliance is necessary. Thus, some portion of the reward fund would still need to be channeled towards the task monitoring compliance. However, while the fixed costs of monitoring tasks would remain the same, the model showed that the actual size of the reward would be lower. As the "lost rewards" would be circulated back to the Enforcers to finance (part of) their costs of punishment, the Enforcers would need to set up a smaller reward fund initially.

Additional problems stem from the need to determine when to reward the punisher. Some potential Enforcers may be motivated by the opportunity to collect a reward in situations where punishment is not welfare-enhancing. In addition to frivolous attempts to punish

Violators in order to collect the reward, some Enforcers may seek to collect the reward yet not use it to carry out punishment. This would, obviously, dilute the entire “double leverage” mechanism that makes the Rewarding the Punisher model attractive in the first place. Thus, it may be necessary to monitor not only the conduct of the Violators but that of the Enforcers in order to guarantee the effectiveness of the fund.

Conclusion

[Summarize insights. Note that the insight of the model is relevant for numerous areas of international law where the cost of enforcement presents a key impediment for achieving cooperation. Discuss extensions.]