NBER WORKING PAPER SERIES

THREATS AND PROMISES

Jonathan Eaton
Maxim Engers

Working Paper No. 4849

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 1994

We thank B. Peter Rosendorff, Marie C. Thursby, and participants at the NBER-Universities Conference on International Trade Rules and Institutions, Cambridge, MA, December 1993, and at the Summer Meetings of the North American Econometric Society, Quebec, Canada, June 1994, for comments. We gratefully acknowledge the support of the National Science Foundation under grant number SES-9111647. This paper is part of NBER's research program in International Trade and Investment. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Working Paper #4849
September 1994

THREATS AND PROMISES

ABSTRACT

Global environmental concerns have increased the sensitivity of governments and other parties to the actions of those outside their national jurisdiction. Parties have tried to extend influence extraterritorially both by promising to reward desired behavior and by threatening to punish undesired behavior. If information were perfect, the Coase theorem would suggest that either method of seeking influence could provide an efficient outcome. If the parties in question have incomplete information about each other's costs and benefits from different actions, however, either method can be costly, both to those seeking influence and in terms of overall efficiency. We compare various methods of seeking influence. A particular issue is dissembling: taking an action to mislead the other party about the cost or benefit of that action. By creating an incentive to dissemble, attempts to influence another's behavior can have the perverse effect of actually encouraging the action that one is trying to discourage.

Jonathan Eaton
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
and NBER

Maxim Engers
Department of Economics
Rouss Hall
University of Virginia
Charlottesville, VA 22901
and NBER

I. Introduction

The correction of externalities requires that individuals affected by the decisions of others have ways of influencing those decisions. Within national boundaries laws and contracts typically perform this function. A third party, the legal system, can enforce laws against undesired actions, or subsidize rewards for desirable actions. It can also enforce long-term contracts among the interested parties. Parties who are in different countries, or who are themselves governments of different countries, may lack a third party enforcement mechanism. The parties involved must then rely on their own devices to try to influence the decisions of others.

One way to do this is through what Schelling (1960, 1965) calls "brute force," taking direct physical control of the action in question. In this paper we consider less extreme forms of seeking influence: rewarding others for refraining from undesired actions, or punishing others for carrying them out.[1] In an international context, the punishments involved are usually called sanctions.

Our main aim is to understand and to evaluate international institutions to preserve the quality of the environment. Treaties to protect the environment have contained various provisions both to reward countries for pursuing environmentally sound policies, and to punish them for polluting.[2]

The problems of limiting nuclear proliferation provide further motivation. Recent United States policy has varied between threatening to punish North Korea for violating the International Atomic Energy Agreement and

---

[1]Dixit (1987) provides a lucid discussion of these issues.

[2]The Ottawa convention governing the use of chlorofluorocarbons is an example. It calls both for sanctions against violators and rewards for compliance. Parson (1991) provides a description.

promising to reward it for compliance.

In a situation of symmetric information the Coase Theorem (1960) suggests that the only issue is distributional: Seeking influence by either method ensures an efficient outcome.[3] As is now widely recognized, however, if the relevant parties have incomplete information about the costs and benefits of various actions to each other then the question of how to achieve an efficient outcome is more complex.[4]

Our purpose here is to consider the relative benefits of seeking influence by promising rewards and by threatening punishments, both from the perspective of the sender and of efficiency in general. To do so we consider the interaction over time of two parties, called, in the tradition of the literature on sanctions, the sender and the the target.[5] The sender would like to discourage the target from taking an action that benefits the target but harms the sender. The sender can do so by threatening sanctions if the target takes the action, or by promising a reward for not taking it. Both means are costly: Sanctions impose a cost on the sender as well as on the target, while rewarding the target requires giving up resources.

We use two devices to aid exposition. First, to help identify the antecedents of pronouns, we assign a gender to each party, treating the sender as feminine and the target as masculine. Second, since our focus is on preventing environmental degradation, we call the offending action of the target pollution (although in principle it could be any action that the target

---

[3]We explore the use of sanctions in a situation of complete information elsewhere (1992).

[4]Our analysis relates closely to the literature on sequential bargaining under incomplete information. Fudenberg and Tirole (1991) provide a discussion and references to the literature.

[5]See Hufbauer et al. (1990). Other contributions to the literature on sanctions are Daoudi and Dajani (1983) and Kaempfer and Lowenberg (1988).

could take that is to his own benefit and to the sender's detriment).

Given the range of possible rewards and punishments, and alternative information structures, there are myriad possible specifications of the relationship between the parties. We do not attempt to provide even a partial taxonomy. Instead, we focus on situations that illustrate clearly some possible merits and deficiencies of using alternative methods of influence.

Since the most serious problems emerge when the sender does not know the extent to which the target benefits from polluting, we focus on this informational asymmetry. Specifically, the sender is unsure whether the target benefits from polluting by only a small amount, in which case he is "clean," or by a large amount, in which case he is "dirty." A basic issue is what the target reveals about his type through his response to the sender's attempt to influence him. We therefore consider their repeated interaction.

We consider three tools that the sender might use to discourage pollution: (i) a reward for not polluting; (ii) a mild punishment that is enough to deter a clean but not a dirty target from polluting; and (iii) a draconian punishment so severe that it would deter either type of target, but at a higher cost than the mild one. We consider situations in which the sender has access to these three tools in various combinations.

Among other things, we consider the following questions: To what extent do different combinations of tools work to the sender's advantage? What inefficiencies can arise in different situations? When would the sender benefit by committing herself to a policy for the duration of the relationship rather than by choosing a policy each period? Could she ever do better by committing herself to "laissez-faire," taking no action to influence the target?

No single method among the cases we consider dominates any of the others,

either in terms of benefiting the sender or in avoiding inefficient outcomes. What is best for the sender or most efficient overall depends on the specific situation. We find that the sender often could do better by committing to a policy over the two periods. In some cases the sender faces the dilemma that her attempts to seek influence lead to a worse outcome for her, and to more pollution, than would a policy of laissez-faire, if she could commit to such a policy.

Basic to our results is the way that the sender would treat targets of different types. Unless the sender has access to the draconian sanction, she treats a dirty target more favorably than a clean one: If using rewards, she would promise a larger one to a dirty type, since he requires greater compensation not to pollute, while she would not incur the cost of the mild sanction in dealing with a dirty target, since it would not work.

Because the sender treats a dirty target more favorably, a clean one has an incentive to "bluff," i.e., to try to pass himself off as dirty: A clean target might pollute rather than accept a reward that exceeds his benefit from polluting to try to elicit a higher reward later. Alternatively, a clean target might pollute and suffer sanctions to try to make the sender think that is not worth renewing them.

When the sender has access to both a draconian sanction and a mild sanction, however, the clean target no longer has an incentive to make the sender think that he is dirty. Once the sender is sure enough that she is dealing with a dirty target she simply resorts to the draconian sanction. Hence in this case the dirty target, rather than the clean one, has the incentive to conceal his type, acting as a "wolf in sheep's clothing:" The dirty target mimics the clean target to convince the sender that she does not need the draconian sanction to deter pollution. Having fooled the sender the

target then goes ahead and pollutes, suffering only the mild sanction.

We proceed as follows: Section II presents the basic structure of our analysis. Sections III through VI consider specific situations in detail. Section VII offers some concluding remarks.

II. The Basic Structure

We use D to denote the amount of damage done to the sender by the target's pollution. This amount is known to both. The target's benefit from polluting could either be a high amount H if he is dirty or a low amount L (where $H > L > 0$) if he is clean. The true amount is known by the target, but the sender is unsure. She initially believes that the target is dirty with probability θ_1 and is clean with probability $1-\theta_1$. The harm that the target's action inflicts on the sender exceeds whatever benefit the target derives from it. Hence $D > H > L$. From a social perspective, then, the target's action is inefficient.[6]

The two parties interact for two periods, the smallest number allowing us to examine how the target might modify his actions in one period to influence the sender's subsequent beliefs. For simplicity we assume no discounting.

Within each period the two parties interact as follows: The sender begins period i believing that the target is dirty with probability θ_1. On the basis of this belief she either promises the target a reward for not polluting, threatens to punish him if he does pollute, or does both in some combination. In order to punish the target that period, the sender must, at

---

[6]If D is less than H the problem, with rewards, is trivial: If $D > L$ the sender offers L. This offer is accepted by a clean target and rejected by a dirty one. The clean target has no prospect of eliciting a higher reward later. If $D < L$ the sender offers no reward. Pollution always occurs. In both cases the outcome is efficient.

this point, incur a sunk cost C regardless of whether or not she then punishes the target. This cost could represent, for example, the cost of maintaining a military, which the sender must incur before knowing whether or not the target will pollute. We assume that the cost of the sanction C is less than than the amount of damage D that the sender sustains from pollution (or the sanction would never be used).

Once the sender has made her threats or promises, the target chooses whether or not to pollute. Polluting raises the target's payoff, depending on his type, by H or L, and lowers the sender's by D.

What happens next depends on the threats and promises made by the sender at the beginning of the period and the target's subsequent action. If the sender had promised a reward then it is paid if and only if the target has refrained from polluting.[7] If the sender threatened a punishment and incurred the sunk cost C then the sender punishes the target, at no additional cost to herself but at a cost P to the target, if and only if the target has polluted.[8] Hence the target's receipt of either the reward or the punishment is contingent upon the target's behavior, as is the sender's payment of the reward. The sender must incur the cost of arranging the sanction regardless of what the target does and whether or not the sanction is actually implemented.

---

[7]We assume that the sender can make a commitment actually to pay R contingent on the target's not polluting. There are various reasons why the sender would follow through on her promise. She may, for example, have to deal with multiple targets, and wish to maintain a reputation for honesty. Alternatively she may be able to put the payment in escrow under the control of a third party instructed to make payment to the target if he desists from polluting and to return the reward to the sender if he pollutes.

[8]Since at this point the incremental cost of punishing is zero, the commitment to impose sanctions is weakly credible. We have explored the implications of introducing a (strictly positive) incremental cost to sanctions (continuing to assume that the sender is committed to imposing them if the target acts), but deemed the additional insights yielded by this modification not worth the added complications.

We consider four specific situations which do not exhaust all logical possibilities, but together illustrate basic issues that arise in other, more complicated, circumstances. First are the two pure cases of a sender seeking influence only through rewards or only through a single sanction. We focus on the only interesting situation, in which the sanction is mild, i.e., $H > P > L$.[9] We then consider a sender having access to both methods of influence in combination. Finally, we consider the situation of a sender with access to either of two sanctions, one of which is mild while the other is draconian, meaning that the harm that it does to the target exceeds the gain from polluting of even a dirty target, i.e., $P > H$. The draconian sanction, however, requires a larger sunk cost F than the mild sanction, but this sunk cost is still less than the damage done to the sender by the target's pollution, i. e., $F < D$.

We characterize the relationship between the sender and target in terms of the perfect Bayesian equilibrium of their interaction: Whenever parties would make decisions, their strategies are optimal given their beliefs, and they update their beliefs using equilibrium strategies and observed actions according to Bayes' Rule.[10]

For each of our four cases we derive the sender's expected total cost of dealing with the target, including the cost of paying rewards, imposing sanctions, and the damage if she fails to deter pollution. We also examine how inefficient the outcome is.

We now turn to the specifics of each case.

---

[9]If the sanction is draconian ($P > H$) then the sender would use it, and successfully deter all pollution. If it is totally ineffectual ($P < L$) it would never be used.

[10]See, for example, Fudenberg and Tirole (1991, Chapter 8).

III. Rewarding Good Behavior

We first consider a sender who is trying to deter pollution by offering a reward for not polluting. What happens depends upon the relationship between the sender's initial belief about the target and the payoffs. Depending on parameter values, there are three kinds of equilibrium outcome.

1. Complete Pooling: If $\theta_1 > \bar{\theta}$, where $\bar{\theta} = (H-L)/(D-L)$, then the sender simply offers H each period. Whatever the target's type, he accepts.

In this outcome there is no pollution, and the sender never learns anything about the target's type. The clean target successfully exploits the sender's ignorance for two periods.

2. Bluffing: If $\theta_1 < \hat{\theta}$, where $\hat{\theta} = \bar{\theta}(H-R)/(D-R)$ and R can be any number between L and H, then the sender offers R the first period. The dirty target rejects this offer while the clean target rejects it with probability π^R, where π^R satisfies:

$$\bar{\theta} = \frac{\theta_1}{\theta_1 + (1-\theta_1)\pi^R} . \qquad (1)$$

If the offer is accepted then in period 2 the sender offers only L, which is accepted. If the offer is rejected the sender then offers H with probability π^H and L with remaining probability $1-\pi^H$, where π^H satisfies:

$$R = (1-\pi^H)L + \pi^H H \qquad (2)$$

The high offer is accepted by either type while the low offer is accepted only by a clean target.

In this case not only does the dirty target pollute in period 1, but so might the clean target, who pollutes to try to increase its reward in period 2. In the second period the clean target does not pollute, while the dirty target does not pollute only if the sender makes a high offer.

3. The Hold-Up: If $\bar{\theta} > \theta_1 > \hat{\theta}$ then the sender offers H in period 1, which is accepted, and L in period 2, which is accepted only by the clean target.

There is no pollution in the first period, while in the second period the dirty target pollutes while the clean one does not.

We first describe the parties' optimal strategies at each point in their interaction, and then show how this behavior yields these three equilibrium outcomes.

A. Optimal Strategies

We begin with the second period. Since this is the last period of their interaction the target has no incentive to alter his decisions to affect the sender's future beliefs. Hence a clean target will accept any reward above L not to pollute and a dirty one will accept any reward above H not to pollute. Depending on the target's type, offers below these amounts are rejected, and lead to pollution. (We assume that offers just equal to the target's benefit from polluting are accepted.)

For the sender, offering a reward of L is better than offering any reward between L and H while offering H is better than any remaining offer. A reward of H will be accepted by the target regardless of his type, and so will deter pollution for sure. An offer of L will be accepted by a clean target but rejected by a dirty one. Based on her initial beliefs and on what happened the first period, the sender at this point believes that the target is dirty with probability θ_2. Hence she believes the low offer will deter pollution with probability $1-\theta_2$, but will fail to do so with probability θ_2.

What the sender does depends upon how this probability compares with the threshold level $\bar{\theta}$. The sender will offer L if $\theta_2 < \bar{\theta}$ and offer H if $\theta_2 > \bar{\theta}$. If $\theta_2 = \bar{\theta}$ then she is indifferent between these two offers. In this case we let π^H denote the probability with which she makes the high offer, so that she makes the low offer with probability $1-\pi^H$.

The dirty target's payoff in the final period is H regardless of which offer the sender makes: She either rewards him for not polluting with an amount H, or offers L, in which case he rejects the offer and pollutes, deriving a benefit of H. Since the clean target accepts either offer, his benefit is whatever the sender offers, so he prefers the high offer.

We now turn to the first period, beginning with the target's decision. Since the dirty target enjoys H the second period regardless, he has no incentive to alter the sender's beliefs about his type. Hence he behaves just as in the second period, rejecting any offer below H but accepting H or above.

For the clean target things are more complicated. He will obviously accept any offer above H: It will more than compensate him for the benefit of polluting and accepting it will not diminish the sender's belief that he is a dirty type (since he is doing the only thing that a dirty type would do).

By accepting a reward below H, however, a clean target reveals his type

to the sender. Since acceptance means that $\theta_2 = 0$, the sender will offer only L in period 2. Hence accepting any offer $R < H$ yields the clean target R+L over the two periods. (Obviously the target will reject an offer below L since that would leave him less than the payoff 2L, which he can get by rejecting all offers).

If the clean target rejects a low reward R, where $L \leq R < H$, his payoff over the two periods is 2L if the sender offers a reward L in the second period and L+H if the sender offers a reward H in the second period. These two payoffs bracket R+L, the payoff from accepting the offer. Hence what the clean target does depends on the implications of rejecting the low offer for the sender's offer the next period.

If the sender makes a low offer and the target rejects it, from Bayes' rule the sender's posterior belief that the target is dirty is:

$$\theta_2 = \frac{\theta_1}{\theta_1 + (1-\theta_1)\pi^R} \qquad (3)$$

where π^R is the probability that a clean target rejects an offer of a low reward.

B. Equilibrium Outcomes

We now show how this behavior gives rise to the three equilibria.

1. Complete Pooling

If $\theta_1 > \bar{\theta}$ then the sender will offer H in period 2, even if the clean

target rejects low offers with probability one (π^R=1). Even though she knows that the clean target would always bluff, she is sufficiently sure that the target is dirty that, if an offer of R in round 1 were rejected, she would offer H in period 2. In this case, by rejecting any low offer the clean target would get L+H, which exceeds what he would get by accepting a low offer and blowing his cover. Hence, regardless of his type, the target would always reject low offers in period 1. Since the sender would learn nothing about the target's type from the rejection of a low offer, she would offer H for sure in period 2: Making a low offer in period 1 would cost her D+H over the two periods, while offering H, which is accepted by either type, costs her C^R - 2H. Thus her best strategy is to offer H each period, which is accepted by the target regardless of his type.

The sender's cost and the amount of pollution

In this equilibrium the sender's cost is 2H, the dirty target's benefit from polluting. Since there is no pollution the outcome is efficient. The sender could not improve her situation if she could commit to some different strategy, since offering H in each period is the best strategy even if commitment were possible.

If $\theta_1 < \bar{\theta}$ then things become much more complicated. At the low end of this range the sender makes a low offer (strictly below H but not strictly below L), and the clean target bluffs with positive probability. At the high end the sender offers H in period 1 to forestall bluffing, but then offers L in period 2. By credibly threatening to bluff in response to a low first-period offer, the clean target "holds up" the sender for a high offer. We now discuss these two outcomes in greater detail.

2. Bluffing

Suppose that the sender offers a reward R, where $L < R < H$, in period 1. If the clean target always rejected this offer then, since $\theta_1 < \bar{\theta}$, the sender is sufficiently sure that the target is clean that after rejection she would offer L in period 2 But the clean type would then always want to accept the reward R in period 1. But if the clean target always accepted the reward R in period 1 then the sender would know that a rejection meant that the target was dirty, so she would offer H in period 2. The clean type would then reject R in period 1. The only way out of this conundrum is for the sender and the clean target to play mixed strategies, with the sender responding to rejection of a low reward in period 1 by mixing between rewards of H and L in period 2, and the clean target mixing between accepting and rejecting R in period 1. The sender's probability of making a high offer in period 2 must leave the clean target indifferent between accepting and rejecting low offers in period 1, while the clean target's probability of rejecting low offers in period 1 must leave the sender indifferent between offering rewards H and L in period 2.

Since the sender is indifferent only when $\theta_2 = \bar{\theta}$, π^R must satisfy (1).[11] Mixing by the clean target means that he must be equally willing to accept and to reject low offers, which means that π^H must satisfy (2). There are a continuum of equilibrium outcomes in which π^H varies from 0 to 1, with the

[11]If the clean target were to reject low offers with higher probability then the sender would be sure enough that she was dealing with a bluffer that she would offer only L in period 2, in which case the clean target should always accept a low offer. But if the clean target were to reject low offers with lower probability then the sender would be so sure that she was dealing with a dirty type that she would offer only H in period 2, in which case the clean target should always reject a low offer. Only when the clean target rejects with this probability can he provide the sender an incentive to behave in a way that is consistent with his decision.

corresponding R varying from L to H.[12]

The cost to the sender and the pollution amount

The total expected cost to the sender in the bluffing equilibrium is:

$$C_L^R = (1-\bar{\theta})(R+L) + \bar{\theta}(D+H) \qquad (4a)$$

where

$$\bar{\theta} = \theta_1+(1-\theta_1)\pi^R = \theta_1/\bar{\theta} = \theta_1(D-L)/(H-L) \qquad (5)$$

denotes the unconditional probability with which the target rejects a low offer in period 1, either because he is dirty or because he is clean but bluffing. Substituting (5) into (4a) we get:

$$C_L^R = (1-\theta_1)(R+L) + 2\theta_1 D + \theta_1\frac{(D-H)(D-R)}{H-L} \qquad (4b)$$

We can index the continuum of bluffing equilibria by R, which can range between L and H, and the corresponding π^H. In the one that is best for the sender R = L and $\pi^H = 0$. In this case the clean target fails to grab any informational rent from the sender's ignorance about his type, since his payoff is just 2L (what he would get if the sender were not trying to

---

[12]If the first-period offer R equals L then the sender uses a pure strategy, always offering L when this offer is rejected. The clean target is indifferent between accepting and rejecting the first-period offer. If π denotes the probability that he rejects this offer, there are a continuum of equilibria in which π varies between one and π^R as given in equation (1). We focus on the best one for the sender, which has $\pi = \pi^R$. This involves no real loss of generality since the target is indifferent among all these and the sender can force $\pi = \pi^R$ by choosing R to be slightly above L.

influence him). In the worst one for the sender R = H and $\pi^H = 1$, in which case the clean target receives the dirty target's reward for one period.

Note, from (4a), that bluffers make themselves as costly to the sender as truly dirty targets. Hence bluffing raises the sender's cost. This point is made clearer by splitting the sender's expected cost of dealing with the target (4a) into three terms:

$$C_L^R = (1-\theta_1)(R+L) + \theta_1(D+H) + (\bar{\theta}-\theta_1)[(D+H)-(R+L)]. \qquad (4c)$$

The first is the sender's expected cost of compensating a clean target if he were always to reveal his type in period 1, the second is her expected cost of dealing with a dirty target, and the third is the cost created by the clean target's bluffing. In a world in which the target's type were automatically revealed to the sender at the end of the first period, only the first two terms would remain since the target could not bluff.

How does the sender's expected cost of dealing with the target respond to changes in exogenous parameters? As one would expect, her cost increases with D, the damage she herself suffers from pollution, with L, the extent to which the clean target benefits from pollution, and with θ_1, the initial likelihood that the target is dirty. Increases in D and L not only raise her cost directly, but increase the amount of bluffing. An increase in H, the dirty target's payoff from pollution raises her direct costs but reduces the incentive to bluff. The reason is that, as H rises toward D, the cost of buying off a suspected dirty target approaches the cost of pollution itself. The sender must be increasingly sure that she is dealing with a dirty type to find offering H instead of suffering D worthwhile. Hence less and less bluffing can occur in order for her to be willing to offer H. The effect of

an increase in H in curtailing bluffing more than offsets its effect on the sender's direct cost. Hence the more a dirty target benefits from pollution the better off is the sender.

In the bluffing equilibrium pollution occurs in period 1 with probability $\bar{\theta}$ and in period 2 with probability θ_1. Hence offering rewards fails to achieve the efficient outcome of no pollution. Because of bluffing there is even more pollution than would occur if the sender could commit to offering only L both periods. This strategy would eliminate the incentive for the clean target to bluff. Only a dirty type would ever pollute, so that pollution would occur each period with probability $\theta_1 < \bar{\theta}$.

The sender's dilemma

A slight variation of the model shows how the sender might do better if she could convincingly refrain from trying to exert any influence over the target at all. Let L be slightly negative, so that, in the absence of any incentive to bluff, the clean target would strictly prefer not to pollute. By committing to laissez faire the sender could achieve a cost of $2\theta_1 D$. But as long as L is close to zero, her desire to pay a suspected dirty target not to pollute creates an incentive for a clean target to bluff, even though pollution is costly for him. Her cost even in the best outcome for her (when $L = R = 0$ in expression (4b)), then rises.

3. The Hold Up

Instead of making a low offer in period 1, however, the sender could make an offer H that would be accepted for sure by either type of target. She would learn nothing about the target's type in period 1, so that $\theta_2 = \theta_1$.

Since we are now dealing with the case in which $\theta_1 < \bar{\theta}$, she would then offer only L in period 2. Her total cost offering H in period 1 would then be:

$$C_H^R = H + \theta_1 D + (1-\theta_1)L. \qquad (6)$$

Comparing (6) with (4b), this strategy is better than offering a low reward in period 1 if θ_1 exceeds the critical value $\hat{\theta}$.

The sender's cost and the amount of pollution

In this outcome the sender again would benefit if she could commit to offering L both periods. Here, however, the outcome is more efficient than with commitment since commitment would entail more pollution, in contrast to the bluffing equilibrium, where committing to pay just L would reduce pollution.

IV. Punishing with a Mild Sanction

Now consider a sender who is trying to influence the target's decision by threatening a mild punishment. At the beginning of each period the sender decides whether or not to spend C to allow her to punish the target if he pollutes that period. If she makes the expenditure and the target pollutes she inflicts a cost P on him at no further cost to herself. Here we assume that $H > P > L$, so that the threat of the punishment deters a clean but not a dirty target.

As with seeking influence with rewards, there are three kinds of equilibrium outcomes, depending on the sender's initial prior about the target's type:

1. Complete Pooling: Again, pooling occurs if θ_1 exceeds a threshold $\bar{\theta}$, where now $\bar{\theta} = (D-C)/D$. If the sender is initially this sure that the target is dirty, she never imposes sanctions and never learns the target's type. Either type of target always pollutes.[18]

2. Bluffing: If $\theta_1 < \bar{\theta}^2$ then the sender threatens sanctions in period 1. The dirty target balks, suffering the punishment, as does the clean target with positive probability. The sender renews the threat for sure if it worked the first period and renews it with positive probability even if it failed. The new threat always works against the clean target but never against the dirty.

3. Delayed Sanctions: If the sender initially thinks that the target is dirty with probability θ_1, where $\bar{\theta}^2 < \theta_1 < \bar{\theta}$, then she waits until the second period to threaten sanctions. Both types of target pollute the first period as does the dirty target in the second period.

A. Optimal Strategies

To derive the parties' equilibrium strategies we again begin with second period. At this point the threat of sanctions always works against the clean target, but not against the dirty one. The sender believes that the target is

---

[18]We ignore another perfect Bayesian (pooling) equilibrium in the range $\theta_1 > \bar{\theta}$ which can arise if $P > 2H$. The sender threatens sanctions in period 1, renewing them if and only if the target target did pollute. Neither target pollutes in period 1, while both do in period 2. The beliefs needed to support this equilibrium have the contrived property that period 1 pollution makes the sender think that the target is more likely to be clean, even though the dirty target has a stronger incentive to pollute.

dirty with probability θ_2, so that the expected cost of threatening sanctions is $C + \theta_2 D$. If she does nothing either type of target will pollute, so her cost is D. Comparing these costs determines the threshold $\bar{\theta}$.

Either type of target would prefer that the sender not threaten the sanction. The threat lowers the payoff of the clean target from L to 0 and of the dirty target from H to H - P.

Turn now to period 1. The threat of sanctions would not deter a dirty target from polluting both because the benefit of polluting exceeds the harm of the punishment and because pollution will make the sender at least as confident that the target is dirty.

The clean target's decision is more complicated. Knuckling under gives away his type, so that $\theta_2 = 0$ and the sender will definitely renew the threat in period 2. His total payoff over the two periods is zero. But if he pollutes the sender may think it sufficiently likely that he is dirty that she may decide not to renew the threat. His payoff then is $L-P+(1-\pi^S)L$, where π^S is the probability that the sender will renew the threat if the target pollutes.

One possibility is that $P > 2L$. In this case the clean target's gain from pollution is so low relative to the pain inflicted by the sanction that it is not worth suffering the penalty even one period in order to pollute both periods. The threat will always deter a clean target.

More interesting is the case in which $P < 2L$. Here the clean target would be willing to pollute and suffer the penalty in period 1 if he were sure that it would lead to the lifting of sanctions the next period; the clean target has an incentive to bluff. Here bluffing occurs to get sanctions removed rather than to get a higher reward.

If the target does pollute in period 1, the sender believes that the

target is dirty at the beginning of the next period with probability given by (3), where now π^R is the probability that the clean target balks at the threat and pollutes in period 1.

B. Equilibrium Outcomes

The three equilibria emerge from this behavior as follows:

1. Complete Pooling

If $\theta_1 > \bar{\theta}$ then the sender is sufficiently sure that she is dealing with a dirty target that she will not threaten sanctions in period 2 even if she knew that the clean target always bluffed in period 1. Hence the clean target *would* necessarily bluff. The sender's payoff threatening sanctions in period 1 is D+C. In period 2 she would not threaten sanctions, so her payoff would be D. This is worse for her than refraining from the threat at the beginning of period 1. The threat is not worth using in either period, so the sender suffers the cost of pollution each period.

2. Bluffing

If $\theta_1 < \bar{\theta}^2$ the sender threatens the sanction in period 1 and the clean target sometimes bluffs to try to get the sender to drop sanctions. If the threat fails the first time the sender nevertheless renews it with positive probability. Mixing occurs for the same reason that it does when the sender uses rewards: If the clean target always bluffed then the sender would remain sufficiently sure that the target was clean that she would always renew

sanctions. But then the clean target would have no reason to bluff. But if the clean target never bluffed then the sender would always drop sanctions if they failed in the first period, giving the clean target reason to bluff. The only equilibrium outcome is in mixed strategies: To make the parties indifferent between their respective choices: (i) the clean target must bluff with the probability π^R at which $\theta_2 = \bar{\theta}$ (which continues to be given by expression (1)) and (ii) the sender must renew sanctions after pollution occurs with probability $\pi^S = (2L-P)/L$.

The cost to the sender and the amount of pollution

In the bluffing outcome the expected cost to the sender of threatening sanctions in period 1 is:

$$C^S = (1-\bar{\theta})2C + \bar{\theta}(2D+C).$$

$$= (1-\theta_1)2C + \theta_1(2D+C) + (\bar{\theta}-\theta_1)(2D-C)$$

where

$$\bar{\theta} = \theta_1+(1-\theta_1)\pi^R = \theta_1/\bar{\theta} = \theta_1 D/(D-C)$$

is the probability that the target will balk at sanctions in period 1, either because he is dirty or because he is clean but bluffing.

In the second expression for C^S the sender's expected cost of dealing with the target is again the probability-weighted sum of the cost of dealing with a target who reveals himself to be clean, the cost of dealing with a dirty target, and the cost from bluffing. Again, dealing with a target who acts dirty is more costly than dealing with a target who comes clean the first

period, so that a third cost is due to the bluffing caused by the sender's attempts at influence.

Again, if the sender could commit to a policy, here imposing the sanction each period regardless of the first-period response, then the clean target would have no incentive to bluff. The sender's cost would be $2(\theta_1 D+C)$, which is less than her cost in the bluffing outcome. There is also less pollution.

3. Delayed Sanctions

When $\bar{\theta}^2 < \theta_1 < \bar{\theta}$ and the sender does not impose sanctions in period 1 then both types of target would pollute. Since she learns nothing $\theta_2 = \theta_1$ and she would impose sanctions in period 2. Her expected cost from taking this course of action is:

$$C^D = (1+\theta_1)D + C,$$

which is lower than the cost of imposing sanctions if $\theta_1 > \bar{\theta}^2$.

Here potential bluffing causes the sender to eschew sanctions the first period. The clean target exploits the sender's ignorance, but for one period, rather than two, as in the pooling equilibrium. As in the equilibrium with bluffing, the sender would benefit from an ability to commit to imposing sanctions in period 2 regardless of the sender's period 1 actions. Unlike the situation with rewards, where the threat of bluffing led to a hold up and less pollution than would occur if the sender could commit to a course of action, here the potential for bluffing delays sanctions, so that there is more pollution.

C. Mild Threats vs. Promises

How does the sender fare threatening mild punishments rather than making promises? To answer this question we compare her payoffs in the pooling, bluffing, and intermediate outcomes, although the ranges over which these outcomes occur typically differ between the two situations.

In the pooling equilibrium with threats the sender's cost is 2D, instead of 2H, the pooling payoff with rewards. Hence if the sender is very sure that the target is dirty she is better off in a regime of rewards than of mild sanctions.

In the intermediate equilibrium with threats the sender's cost is $(1+\theta_1)D + C$. With rewards it is $H + \theta_1 D + (1-\theta_1)L$. It follows that rewards are less costly as long as the cost of sanctions C exceed the clean target's benefit of polluting L.

The comparison of the two bluffing outcomes is more complicated. Comparing each component of the sender's cost, the relative cost of dealing with a clean target who acquiesces in the first period is higher or lower using rewards as C is higher or lower than L. Since sanctions ultimately do not deter a dirty type, the cost of dealing with a target who acts dirty is always higher using sanctions. Whether more bluffing occurs when the sender uses sanctions is ambiguous. Sanctions are relatively immune from bluffing when C is low, so that the sender is quite likely to impose them even when she is quite doubtful that the target is dirty. Rewards are relatively immune when H is near D, so that the sender must be quite sure that the target is dirty before she offers H. Given that bluffing is going on, the sender will find that it is less costly to influence the target with threats if sanctions do not require any direct cost (i.e., if $C = 0$), even though, unlike rewards,

sanctions fail to deter a dirty target. As the cost of implementing sanctions rises, however, rewards become the lower cost method of achieving influence.

V. Combining Threats and Promises

What happens if the sender can use both the promise of a reward and the threat of the mild sanction in combination? In general the analysis is much more complicated, and we do not provide a complete characterization. Rather we discuss an interesting point that the joint use of the two instruments raises, showing how the sender's ability to use both can make her worse off than if she were restricted to using just one or the other in isolation.

For simplicity we make the additional assumption that $C < L$. Under this restriction, the sender will always threaten sanctions in period 2. They deter a clean target more cheaply than an offer of L, and if she plans to reward a dirty target not to pollute she need offer only H-P, rather than H, for a net savings of P-C. (Recall that $C < L < P < H$).

Hence her only decision is whether to offer H-P and deter pollution by both types or not to offer anything, thereby deterring only the clean type. The threshold probability is now $\bar{\theta} = (H-P)/D$. If θ_2 exceeds this amount the sender will offer the reward as well as impose sanctions, while if θ_2 is less she will only impose sanctions.

As in the previous cases, if the sender's initial prior exceeds this amount there is complete pooling. Compared with the case of the simple reward, either target's payoff is lower since by using the sanction the sender lowers the transfer that she makes to either type. The outcome is nevertheless inefficient, since the threat imposes a real resource cost.

Consider what happens in the bluffing equilibrium, however. Having

access to the sanction changes the sender's payoff from (4b) to:

$$C_L^R = (1-\theta_1)R + 2\theta_1 D + \theta_1 \frac{(D-H+P)(D-R)}{H-P} + 2C. \qquad (4d)$$

The threat acts to reduce the sender's cost in that she no longer has to pay L to a nonbluffing clean target (although she must incur the cost of the threat over two periods). Moreover, the threat lowers the cost of buying out a suspected dirty target in period 2. But as a consequence the sender is more prone to go ahead and offer the reward to a possible dirty type. There is thus more scope for bluffing than if the sender used only rewards in isolation. Since bluffing hurts the sender, the net effect of having access to the penalty (comparing the equilibria that are best for her in each case) can easily be negative. This happens, for example, if L is near zero. As a consequence, when the sender initially strongly suspects that the target is clean, so that a bluffing equilibrium emerges, her total cost using both instruments can be higher than if she could use just the reward. She would be better off destroying her own ability to use the punishment. Even though threatening the mild punishment in conjunction with promising rewards lowers her direct cost of dealing with each type of target, threats can increase bluffing to such an extent that she is worse off.

VI. Mild and Draconian Sanctions: Wolves in Sheep's Clothing

We now add to the sender's arsenal a draconian sanction that imposes so much harm on the target that it would deter either type from polluting. Threatening it costs the sender an amount F. The sender still has access to the mild sanction that would deter the clean but not the dirty target. To

keep things simple we assume that threatening the mild sanction is costless.

In period 2, then, the sender's relevant decision is whether to impose the draconian sanction and deter pollution for sure, at cost $F < D$, or to impose only the mild sanction, risking pollution by the clean target, at an expected cost $\theta_2 D$. What she does thus depends upon the relationship between her belief the target is dirty and the threshold level $\bar{\theta} = F/D$.

In period 1 the clean target knows that whatever the sender does in period 2, he will not want to pollute. His payoff is zero regardless. He consequently does not care whether the sender learns his type, and so will not pollute as long as either sanction is threatened in period 1.

Hence, if the dirty target balks at the mild sanction in period 1, he blows his cover, $\theta_2 = 1$, and the sender threatens the draconian sanction in period 2. His payoff over both periods from balking at the mild sanction is thus H-P.

If he acquiesces, however, his payoff is zero now but H-P if the sender uses only mild sanctions the next period. Hence, if he thinks she would drop them in period 2 he would be indifferent between polluting now and polluting later. A mixing outcome can then emerge. Denote by π^W the probability with which the dirty target acquiesces to the mild sanction. When confronted with a target who acquiesced in the face of the mild sanction, the sender will suspect that he is dirty with probability:

$$\theta_2 = \frac{\theta_1 \pi^W}{1-\theta_1 + \theta_1 \pi^W} .$$

As long as the dirty target acquiesces to the first period sanction with a probability π^W low enough to keep θ_2 below $\bar{\theta}$ he will dissuade the sender from

the draconian threat in period 2.

The dirty target will want to choose π^W in this range. If he acquiesced with higher probability then the sender would impose the draconian sanction even if he acquiesced the first period. His payoff from this is 0, while balking the first period would bring him H-P.

Hence the dirty target is willing to mix between acquiescing and balking to the mild sanction with any probability between 0 and $(1-\theta_1)F/[\theta_1(D-F)]$ (or one, if this amount exceeds one).

The sender is not indifferent to what the dirty target does, however. Her payoff as a function of π^W is:

$$C^D - \theta_1\pi^W D + \theta_1(1-\pi^W)(D+F) - \theta_1[D + (1-\pi^W)F],$$

which falls as π^W rises. The dirty target's mimicking of the clean target to reduce sanctions the next period works to the sender's advantage.

The reason is that, if she imposes the mild sanction in period 1, the dirty target is going to pollute exactly once, either in period 1, in which case she will impose the draconian sanction in period 2, or wait until period 2. In the second case she avoids the cost of the draconian sanction. While the dirty target is indifferent among a continuum of mixing probabilities, the sender could coax him to the best one for her that is consistent with no period 2 sanctions by offering a very small reward for not polluting in period 1.

The sender could, of course, threaten the draconian sanction in period 1. Her payoff would be $F+\theta_1 D$ if she dropped it in period 2 and 2F if she maintained it for both periods. She does better threatening the mild sanction in period 1 as long as $\theta_2 < (F/D)(2-F/D)$. In one period of interaction she

would impose the draconian sanction as long as $\theta_2 > F/D$. With repeated interaction, however, she can use the dirty target's fear of subsequent draconian threats to dissuade him from polluting currently. Hence for higher values of θ_1 she does better not making the draconian threat in the first period. Rather, she uses the possibility of threatening it in the second period to deter pollution in the first.

VII. Conclusion

We have shown how one's attempts to influence the actions of another can be costly, futile, and even self-defeating. These problems can arise whether one is seeking influence by promising to reward good behavior or threatening to punish bad behavior.

Of the various situations we consider, the sender does best when she can threaten a punishment that imposes more harm on the target than the maximum benefit that he could obtain from taking the undesired course of action. In this case the target's attempts to dissemble work in the sender's favor. In order to avoid the threat the target behaves to try to signal that the threat is unnecessary.

In the absence of a draconian penalty, however, offering rewards or threatening mild punishments can work to the sender's disadvantage. These encourage behavior to increase the reward or to avoid the punishment: With rewards the target wants to show that he needs a large reward to desist. With mild punishments the target wants to show that they will not work. Bluffing is most pervasive when the sender is most sure that the target is not really benefiting that much from the action that she is trying to stop.

Our analysis suggests just some of the quandaries that limitations on our

knowledge of others cause in trying to influence them. We have assumed that all parameters except the target's benefit from taking an undesired action are common knowledge. Other possible differences in information might pertain to the target's suffering from the available penalties, as well as the damage done to the sender by the target's actions, and the sender's cost of imposing sanctions and paying rewards. These forms of informational asymmetries pose additional problems. Our analysis preserves their exploration as topics for future research.

REFERENCES

Coase, R. "The Problem of Social Cost," Journal of Law and Economics, 3 (1960): 1-44.

Daoudi, M.S., and Dajani, M.S. Economic Sanctions: Ideals and Experience. London: Routledge & Kegan Paul, 1983.

Dixit, Avinash K. "How Should the United States Respond to Other Countries' Trade Policies?" In U.S. Trade Policies in a Changing World Environment, edited by Robert M. Stern. Cambridge, Mass.: MIT Press, 1987.

Eaton, J. and M. Engers "Sanctions," Journal of Political Economy, 100 (October, 1992): 899-928.

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

Hufbauer, Gary C., Schott, Jeffrey.J., and Elliott, Kimberly A. Economic Sanctions Reconsidered. 2nd ed. Washington D.C.: Institute for International Economics, 1990.

Kaempfer, William H., and Lowenberg, Anton D. "The Theory of International Economic Sanctions: A Public Choice Approach." American Economic Review 78 (September 1988): 786-793.

Parson, Edward A. "Stratospheric Ozone and CFCs: International Institutions." Manuscript. Cambridge, Mass.: Harvard University, 1991.

Schelling, Thomas C. The Strategy of Conflict. Cambridge, Mass.: Harvard University Press, 1960.

Schelling, Thomas C. Arms and Influence. New Haven, Conn.: Yale University Press, 1965.