A THEORY OF INTERNATIONAL COOPERATION

Scott Barrett*

London Business School

November 1997
Revised June 1998

Abstract: This paper develops a coherent theory of international cooperation relying on the twin assumptions of individual and collective rationality. Using a linear version of the *N*-player prisoners' dilemma game, I provide a formal proof of Olson's conjecture that only a "small" number of countries can sustain full cooperation by means of a self-enforcing agreement. Moreover, I find that this number is not fixed but depends on the nature of the cooperation problem; for some problems, three countries will be "too many," while for others even 200 countries will be a "small" number. In addition, I find that the international system is only able to sustain *global* cooperation--that is, cooperation involving 200 or so countries--by a self-enforcing treaty when the gains to cooperation are "small." Finally, I find that the ability of the international system to sustain coooperation does not hinge on whether the compliance norm of customary international law has been internalized by states or whether compliance must instead be enforced by the use of treaty-based sanctions. The constraint on international cooperation is free-rider deterrence, not compliance enforcement.

*I am grateful to Geir Asheim, Olivier Compte, Jeroen Hinloopen, and Marco Mariotti for helpful comments at conference and workshop presentations and to three anonymous referees and the editor of this journal for helping an economist learn how to write a political science paper.

Address. Scott Barrett, Associate Professor of Economics, London Business School, Sussex Place, Regent's Park, London NW1 4SA, UK; tel: 44 171 706-6798; fax: 44 171 402-0718; email: sbarrett@lbs.ac.uk.

1. Introduction

The theory of international cooperation developed in this paper assumes that cooperative arrangements between countries must be both individually and collectively rational: individually rational because the choice of whether to be a party to a treaty is voluntary; collectively rational because diplomats meet face to face and so can exploit fully the potential joint gains from cooperation in a treaty. Individual rationality is a standard assumption in the literature. Collective rationality is a more novel assumption, but it is compelling nonetheless. In this paper I show that the combination of these assumptions has profound implications for the theory of international cooperation.

Two pillars of the received theory are (1) that cooperation can be sustained as an equilibrium of a noncooperative repeated game by strategies of reciprocity (Axelrod, 1984; Axelrod and Keohane, 1985; Keohane, 1986), and (2) that cooperation can only be supported by a "small" number of countries (Olson, 1965; Keohane, 1986). These features of the theory should be compatible but it isn't obvious that they are. Indeed, the "folk theorems" invoked to explain (1) clash with (2); they show that, for small enough discount rates, cooperation can be sustained as an equilibrium for *any* number of players. Olson supports the second pillar of the theory by a convincing, intuitive argument that appeals to the principle of reciprocity, but he doesn't offer a formal proof of the claim, and nor, to my knowledge, has anyone else. So the two pillars remain unreconciled. However, the folk theorems rely only on the assumption of individual rationality; they do not require that agreements also be collectively rational. I show in

this paper that the combination of these assumptions makes features (1) and (2) of the received theory compatible.

In particular, I provide a formal proof of Olson's conjecture that full cooperation can be sustained by means of a self-enforcing agreement only if the number of players is "small."[1]

More than that, I show that whether any given number of countries is "small" depends on the problem at hand. This means that full cooperation can sometimes be sustained by a great many countries and sometimes not even by a few. In showing this, I solve a puzzle in the literature: why some treaties can be sustained by nearly all the countries in the world when others cannot even be sustained by a handful of countries (see Keohane and Ostrom, 1994; Snidal, 1994; Young and Osherenko, 1993). Finally, I show what this means for world welfare. I find that there is an inverse relationship between the maximum number of countries that can sustain full cooperation by means of a self-enforcing agreement and the aggregate gains to cooperation. The international system, hampered as it is by the principle of sovereignty, can only sustain full cooperation among *all* the world's 200 or so countries when the total gains to cooperation are "small"--that is, when a global agreement isn't really needed. I demonstrate these points by analyzing a linear version of the symmetric prisoners' dilemma game, which captures the essentials of the cooperation problem and yet requires amazingly little mathematics. However, I emphasize that the basic insights of the paper can be shown to hold more

[1]Of course, one can always limit cooperation in a repeated game by assuming that discount rates are high enough. I show that cooperation is limited, even for arbitrarily small discount rates.

generally.

What difference does the assumption of collective rationality make to the theory? Cooperation can only be sustained by an international treaty if no country can gain by not being a party to it, and no party can gain by not implementing it. That is, free riding must be deterred and compliance must be enforced. An agreement must therefore specify a strategy--a plan detailing what the parties should do--and this strategy, if obeyed, must succeed in deterring free-riding and enforcing compliance. Moreover, it must be in the interests of the parties actually to behave as the strategy demands. That is, the threat to reciprocate, to harm a country that has deviated from the strategy, must be credible. Essentially, the assumptions of individual and collective rationality define what we mean by a "credible" strategy.

Individual rationality implies that, if every other country plays the equilibrium strategy, each can do no better than to play this strategy; and that, if a country deviated from this strategy "by accident," then this country would want to revert to the equilibrium strategy and so would each of the others want to impose the punishment prescribed by the strategy, given that all other countries obeyed the strategy. That is, when push comes to shove, free-riding and non-compliance are punished; and it is precisely because it is known that this behavior will be punished that no country deviates in equilibrium.

Collective rationality, as the term is used in this paper, implies that an equilibrium

agreement cannot be vulnerable to renegotiation--that there cannot exist an alternative, feasible agreement that all countries prefer to the equilibrium agreement; that, should a country deviate from the equilibrium "by accident," not only would this deviant want to revert to the equilibrium strategy, and not only would every other country behave in the manner prescribed by this strategy, given that all others did so, but all of the countries called upon to punish the defection would actually want to carry out the punishment and would not be tempted to renegotiate the agreement--to choose an alternative, feasible punishment or overlook the defection and not punish the defector at all.

In an infinitely repeated game, strategies capable of deterring a unilateral defection *are* credible (assuming that countries are sufficiently patient), if by "credible" we mean that the strategies are individually rational. This is what the folk theorems tell us. But such strategies will *not* be credible (even for arbitrarily small discount rates) if by "credible" we mean that they are collectively rational, provided N is large enough. The reason is that, the larger is N, the greater will be the harm suffered by the $N - 1$ "other" countries when they impose the punishment needed to deter a unilateral deviation. If N is large enough, it will not be in the collective interests of these countries actually to impose this punishment, should a deviation occur. An agreement which asks its signatories to play this "incredible" strategy would be vulnerable to renegotiation; it would therefore not be self-enforcing.

As just indicated, my analysis is cast in a repeated game setting, and yet Chayes and Chayes (1995) have recently challenged the applicability of the theory of repeated games

to problems of international cooperation. They claim that cooperation is sustained by the international compliance norm and not, as suggested by the theory of repeated games, treaty-based sanctions. The authority to impose sanctions, they note, "is rarely granted by treaty, rarely used when granted, and likely to be ineffective when used" (Chayes and Chayes, 1995: 32-33). Downs, Rocke, and Barsoom (1996; hereafter DRB) disagree that treaty-based sanctions are not needed. They argue that "both the high rate of compliance and relative absence of enforcement threats are due not so much to the irrelevance of enforcement as to the fact that states are avoiding deep cooperation--and the benefits it holds whenever a prisoners' dilemma situation exists--because they are unwilling or unable to pay the costs of enforcement" (DRB, 1996: 387).

It is hard to take sides in this debate, because the Chayes's consider the compliance problem in isolation of free-riding, while DRB conflate these two problems. Compliance and free-riding are different problems. But they are related problems and should be analyzed jointly. Doing so, however, poses an analytical problem: the theory of repeated games does not distinguish between "defection" as a failure to *comply* with an agreement and "defection" as a failure to *participate* in an agreement. The distinction is important, however, because while countries might be compelled, by means of the compliance norm of international law, to comply with the agreements they sign up to, there does not exist an international norm that requires that states *be* signatories to a cooperative agreement. Indeed, the essence of sovereignty is that states are free to participate in treaties or not as they please.

In the second half of this paper I recast the problem of international cooperation as a stage game in which signatories are assumed to choose their actions jointly so as to maximize their collective payoff (as required by collective rationality), in which nonsignatories are assumed to choose their actions independently so as to maximize their individual payoffs (as required by individual rationality), and in which all countries are free to be signatories or nonsignatories (as also required by individual rationality). As noted earlier, the Chayes's and DRB agree that countries comply with the agreements they sign up to; what they disagree on is whether this means that treaty-based sanctions are not needed and whether anything like deep cooperation can be sustained by the international system. I therefore adopt the tactic of assuming that all countries have internalized the compliance norm of customary international law in order to see whether this assumption matters.[2] I show that DRB are right that the international system may fail miserably at sustaining deep cooperation, even assuming that the Chayes's are right that the norms of international behavior suffice to ensure that countries comply fully with their international obligations. Like the earlier result, I also find that only a "small" number of countries can sustain the full cooperative outcome, and that there is an inverse relationship between the maximum number of countries that can sustain full cooperation and the total gains to cooperation.

Because of their different formulations, the repeated and stage game models sustain

[2]To assume that states have internalized the compliance norm is to assume that states will comply with an agreement they have signed up to, whether or not is in their interests to do so. This should be interpreted only as shorthand for the assumption that the compliance norm is sustained outside of the model under consideration. Kandori (1992) shows how norms can be sustained by community enforcement.

cooperation by means of different strategies. To sustain full cooperation as an equilibrium of a repeated prisoners' dilemmma, collective rationality requires that, if a party to an agreement plays Defect, the other parties can do no better collectively than to respond by playing Defect; and that, if this defector subsequently plays Cooperate in a punishment phase, to make amends for its earlier transgression, all the other parties to the agreement still can do no better collectively than to continue to play Defect--that is, to punish the original defection (it is this that makes the agreement "renegotiation-proof"). To sustain full cooperation as an equilibrium of the stage game, collective rationality requires only that the first of the above conditions be obeyed (the second cannot figure in the stage game model, because this game is essentially "one-shot" and so there cannot exist a "punishment phase"): that, if one country plays Defect, all the other countries can do no better collectively than to play Defect. Though different in the details, both strategies have the same basic requirement: that the countries responsible for punishing a unilateral defection must not be able to do better, either individually or collectively, by not carrying out the punishment specified in the treaty. Put differently, both approaches require that cooperation be enforced by credible punishment strategies.

Moreover, for a certain and important class of cooperation problem--one where the cost to participating in a treaty is independent of the number of countries that participate--I show that these conditions are identical. In other words, the compliance norm doesn't buy any additional cooperation.[3] The reason is intuitive. Any punishment to deter non-

[3]This should not come as a surprise. In the model presented here, non-compliance implies that a signatory will play Defect when the agreement requires that it play Cooperate. So a signatory that fails to comply with the agreement will be

compliance must "fit the crime." So the larger the potential compliance failure the larger must be the threatened punishment if non-compliance is to be deterred. The greatest harm that any one signatory can inflict on the others is to do what it would do if it withdrew from the treaty entirely. So if a treaty can credibly threaten to impose a punishment that deters signatories from withdrawing unilaterally, it can easily threaten to impose a punishment that deters signatories from failing to comply with the agreement unilaterally. Once free-riding has been deterred, compliance enforcement comes free of charge.

This result needs to be modified slightly if the cost to each country of playing Cooperate is decreasing in the number of countries that play Cooperate--if there are increasing returns to cooperation. For, in comparison with the case discussed above, if any country plays Defect, the payoff to the others of playing Defect increases (punishing a defection becomes more attractive), whereas if a country plays Cooperate in a punishment phase, the payoff to the others of continuing to play Defect decreases (punishing a defector becomes less attractive). Increasing returns thus makes cooperation a little easier to sustain in the stage game model than in the repeated game model. But the reason for this is not that the assumption of full compliance buys any additional cooperation. The reason is that the stage game lacks a temporal dimension and so can't specify explicitly an appropriate strategy of reciprocity.

The analysis developed in the paper is abstract. Many important features of real world

---

indistinguishable from a country that free-rides on the agreement.

cooperation problems like climate change mitigation and ozone layer protection don't figure in the model--to take an obvious example, I assume that countries are symmetric when they most certainly are not. Moreover, the focus of my analysis is narrow. My interest is in determining the conditions that must hold for full cooperation to be sustained by the anarchic international system. I have little to say in this paper about whether something short of full cooperation can be sustained. But for all of these limitations, the theory is relevant to the real world, as the following example illustrates.

The Montreal Protocol sustains something very close to full cooperation. Nearly every country is a party to this agreement, and in implementing it the most harmful ozone-depleting substances are being phased out around the world. At a recent conference of the parties to the Montreal Protocol, delegates suggested (not for the first time) that this agreement should serve as a model for the climate change negotiations, soon to be convened in Kyoto. The analysis developed in this paper is useful for knowing whether their ambition could be met--whether the success at Montreal could be replicated in Kyoto. The theory tells us that it could be, but only if the underlying payoffs are favorable to international cooperation. Of course, these payoffs are givens, and so it may not be possible for the Kyoto negotiatiors to match the succcess of the Montreal Protocol.[4] To sustain full cooperation requires more than negotiation acumen, more than leadership, more than an active epistemic community, more even that an assurance that

[4] As it happens, the agreement negotiated in Kyoto bears a number of similarities to the Montreal Protocol. Crucially, however, the Kyoto Protocol does not contain a free-rider deterrence mechanism. The Montreal Protocol does, in the form of trade sanctions between parties and non-parties.

countries will obey the compliance norm. It depends also on whether the payoffs are of a magnitude that make the threat to punish deviations from full cooperation credible. This is the central message of this paper.

Before proceding to the substance of the paper, I should perhaps comment on why I specialize by analyzing cooperation as an *international* problem. Certainly, the theory does have relevance to other problems. But the rules of the game of cooperation vary in different situations, and one must take care before extrapolating.[5] Where cooperation among firms is legal, it can be codified in a contract, which can then be enforced by the courts having jurisdiction over the parties. Cooperative arrangements arrived at in this setting need not be self-enforcing. Where cooperation among firms is illegal, it may no longer be possible for firms to negotiate openly, and in this context the notion of collective rationality is less compelling. Finally, local, self-organized collective action problems of the type analyzed by Ostrom (1990) take place in settings where there is at the very least a potential for central intervention.[6] Context matters to the analysis of cooperation, and though the theory developed here will have implications for different settings, I apply it in this paper only to inter-state relations (and indeed only to a subset of these).

2. Individual Rationality in the One-Shot, *N*-Player Prisoners' Dilemma

[5]See the Special Issue of the *Journal of Theoretical Politics*, Vol. 6 (1994), no. 4.

[6]For example, Ostrom begins her study by discussing the inshore fishery at Alanya, where the cooperative which developed rules for managing the community resource had previously been given jurisdiction over such matters by national

The underlying game is assumed to be an N-player prisoners' dilemma, where $N \geq 2$, where countries must choose between playing Cooperate and Defect, and where the payoffs to each of the symmetric players of making these choices (P_D and P_C, respectively) are linear functions of the total number of countries that play Cooperate, z:

$$P_D(z) = bz, \quad P_C(z) = -c + dz. \tag{1}$$

In (1), b, c, and d are parameters, and the payoffs have been normalized such that $P_D(0) = 0$. This linear formulation is obviously special, but it will allow us to obtain very strong results using very little mathematics.

The prisoners' dilemma has three important features, and the parameters in (1) must be restricted to ensure that these are satisfied by the model.

The first feature of the prisoners' dilemma is that play Defect is a dominant strategy in the one-shot game. This means that every player must get a higher payoff when playing Defect than when playing Cooperate, irrespective of the number of other countries that play Defect (Cooperate). Formally, I limit my attention to problems that satisfy:

$$bz > -c + d(z+1) \text{ for all } z,\ 0 \leq z \leq N-1. \tag{2}$$

---

legislation.

The second feature of the prisoners' dilemma is that country i's payoff is increasing in the number of other countries that play Cooperate, irrespective of whether i plays Defect or Cooperate. This implies $b, d > 0$. Furthermore, upon setting $z = 0$ we see that (2) requires $0 > -c + d$, and so, given that $d > 0$, we must have $c > d$.

The third feature of the prisoners' dilemma is that the Nash equilibrium of the one-shot game is inefficient; all N countries would prefer an alternative feasible outcome where at least some countries play Cooperate to the Nash equilibrium in which no country plays Cooperate. I shall strengthen this assumption slightly and assume that the aggregate payoff is strictly increasing in z (this will ensure that the aggregate payoff is maximized when all countries play Cooperate; that is, when $z = N$). A little calculus shows that this requires

$$-c + 2dz > b(2z - N) \text{ for all } z,\ 0 \# z \# N. \tag{3}$$

If the gain to any country i of one more of the other countries playing Cooperate is the same, irrespective of whether i plays Cooperate or Defect, then $b = d$. This situation is illustrated in Figure 1 (see also Schelling, 1978). If, however, the gain to any country i of one more of the other countries playing Cooperate is greater if i plays Cooperate also, then $d > b$. In this case, cooperation would exhibit a kind of increasing returns. I allow for both cases and so assume $d \$ b$.

To sum up, in addition to (1), (2), and (3), the model also assumes:

$$c > d \$ b > 0. \quad (4)$$

With this formulation, the equilibrium of the one-shot, *N*-player prisoners dilemma game is unique: all countries play Defect in equilibrium. This equilibrium is inefficient: every country strictly prefers the outcome in which all countries play Cooperate. The latter outcome, called the full cooperative outcome, maximizes the aggregate welfare of all countries. The problem of international cooperation, at least as defined here, is to sustain the latter outcome as an equilibrium of a repeated game by means of a strategy of reciprocity.

Notice that I have defined the international cooperation problem as one where no country can be excluded from enjoying the benefits associated with cooperation by others. The problem of sustaining international cooperation is thus defined here as a problem of providing an international public good. Protection of the ozone layer and climate change mitigation are examples of global public goods. Other problems of interest are not suited to the model constructed here--international trade agreements being only one example.

3. Individual Rationality in the Infinitely Repeated, *N*-Player Prisoners' Dilemma

Suppose that the one-shot game is repeated infinitely often and that, against this background, the *N* players negotiate an agreement in which they all pledge to play the

famous Grim strategy; that is, they all agree to play Cooperate in period *0* and to play Cooperate in every subsequent period provided no player ever played Defect in the past but that, should Defect ever be played by any player, every player must thereafter play Defect forever.

Grim has two attractive features. The first is that play Grim is a Nash equilibrium: given that the other players play Grim, any player *j* can do no better than to play Grim. To see that this is so in the present model, suppose player *j* deviates in period *t*. It will then get a payoff of $P_D(N - 1) = b(N - 1)$ at time *t*. By (2) we know that $P_D(N - 1) > P_C(N)$. So *j* gains initially from the defection. However, *j* will lose in the long run if the threatened punishment really is carried out. To know whether *j* can gain on balance from defecting, we need only compare the per-period payoff in the cooperative and punishment phases, assuming that the rate of discount is negligibly small. In the punishment phase, *j* gets an average payoff of $P_D(0) = 0$. In the cooperative phase, *j* gets a per-payoff of $P_C(N) = - c + dN$. Inequality (3) tells us that the latter payoff exceeds the former (since (3) must hold for $z = N/2$). So no player can gain by deviating unilaterally from Grim in a cooperative phase.

The Nash equilibrium is a rather weak requirement. For it is reasonable to ask: if a country did deviate "by accident," would every country really play Grim? Suppose that every other country plays Grim in a punishment phase. Will country *i* want to play Grim also? If *i* plays Grim, it will get a per-period payoff of $P_D(0) = 0$. If *i* deviates, it will get a per-period payoff of $P_C(1) = - c + d$. By (2), the former payoff exceeds the

latter. So the threat to implement the Grim punishment is individually rational. Furthermore, this is true for any *N*.

It is of course true that *any* feasible, individually rational outcome of the one-shot game can be sustained as a subgame perfect equilibrium of the infinitely repeated game provided the players are sufficiently patient (see, for example, Fudenberg and Maskin, 1986). For example, the strategy Always Play Defect sustains the equilibrium of the one-shot game as a subgame perfect equilibrium of the repeated prisoners' dilemma. But given that, in the context of international negotiations, the players are able to meet, to deliberate openly on their predicament, to negotiate, it would be collectively irrational for them to choose to sustain a Pareto-inefficient outcome from the set of all outcomes that can be supported as subgame perfect equilibria. So while the one-shot game can't explain how countries could *ever* cooperate, the infinitely repeated game can't explain why countries don't *always* cooperate. Theories built on either edifice will thus lack any cutting power; they won't be able to make sharp predictions.

It might seem from this discussion that the assumption of collective rationality favors cooperation.[7] I show below, however, that this is not so. More than that, I show that this assumption gives the cutting power that we desire in a theory.

4. Collective Rationality in the Infinitely Repeated, *N*-Player Prisoners' Dilemma

[7]Indeed, were I to drop the assumption of individual rationality, collective rationality would sustain *only* the full cooperative outcome.

Though Grim is subgame perfect, it *seems* incredible because it is grossly unforgiving. Indeed, it is precisely for this reason that the famous Tit-for-Tat strategy appeals more to our intuition. But Tit-for-Tat is *not* subgame perfect; it is not an individually rational strategy. If a party deviates and then reverts to Tit-for-Tat, and if all other players play Tit-for-Tat, then the one-off defection results in an "unending echo of alternating defections" (Axelrod, 1984: 176). In other words, the players could do better by deviating from Tit-for-Tat after the one-off deviation has occurred.

Contrary to intuition, Grim can can claim to be superior to Tit-for-Tat. But there is a problem with Grim that individual rationality fails to reveal. As our intuition suggests, Grim is too unforgiving. Though countries do not have an incentive to deviate from Grim unilaterally, they do have an incentive to deviate *en masse*. Grim is not a collectively rational strategy.

To see this, consider the $N = 2$ game and suppose that one of these countries, country j, deviates from Grim "by accident." Then each player will get an average per-period payoff in the punishment phase of 0. Though neither player can do better by deviating in the punishment phase, both players would do better collectively by renegotiating their agreement and restarting a cooperative phase, for they would then each get an average payoff of $-c + 2d$, and by (3) we know that $-c + 2d > 0$. Moreover, consistency demands that the theory allow them to renegotiate. The folk theorems are intended to explain how cooperation might emerge as an equilibrium, but they only allow players to

begin a cooperative phase once (usually, in some period labelled *0*). This is arbitrary. The theory should also allow cooperation to restart following a period of defection. Put differently, the theory should acknowledge that the players cannot make a credible commitment not to renegotiate. A self-enforcing treaty must not only be subgame perfect but also immune to renegotiation.[8]

A strategy that satisfies these requirements is a close cousin of Tit-for-Tat, Getting-Even.[9] This requires that country *i* play Cooperate unless *i* has played Defect less often than any of the other players in the past. The main difference between Tit-for-Tat and Getting-Even is that the latter strategy imposes a punishment that is more proportionate to the harm caused by the deviation. In a 2-player game, if one player deviates for 20 periods and then reverts to cooperation, Tit-for-Tat demands that the other player revert to cooperation immediately after the first player has done so. Getting-Even, by contrast, requires that the other player not revert to cooperation for 20 periods.

To show that Getting-Even is both individually and collectively rational, consider again

[8]It might be argued that it should also not be possible for any coalition of countries, taking the actions of all others as given, to agree to deviate from the agreement; that it should not be possible for any sub-coalition to agree to deviate from this alternative agreement; and so on. In other words, it might be argued that treaties should be coalition-proof Nash equilibria (see Bernheim, Peleg, and Whinston, 1987). However, application of this concept to the infinitely repeated prisoners' dilemma poses certain technical problems, as noted by Bernheim, Peleg, and Whinston.

[9]The concept of a renegotiation-proof equilibrium used here is due to Farrell and Maskin (1989). van Damme (1989) derives the strategy which supports full cooperation as a renegotiation-proof equilibrium of the 2-player prisoners' dilemma. See also Myerson (1991), who gave the above strategy the name, "Getting-Even." My contribution here is to extend the use of this concept to the $N > 2$ case and to apply it to

the N-player game. Suppose j deviates at time t and then reverts to Getting-Even in period $t + 1$. j then gets a payoff of $b(N - 1)$ in period t, a payoff of $- c + d$ in the punishment period, and a per-period payoff of $- c + dN$ from period $t + 2$ onwards. Had j not deviated, it would have gotten a payoff of $- c + dN$ every period from time t onwards. Since we are taking discount rates to be vanishingly small, deviating is individually irrational provided j would get a larger total payoff in periods $t + 1$ and $t + 2$ by playing Cooperate than by playing Defect. If j doesn't defect, it will get $2(- c + dN)$ in these periods. If j does defect and then reverts to Getting-Even, it will get $b(N - 1) - c + d$ in these periods. Play Getting-Even is thus individually rational if $2(- c + dN) > b(N - 1) - c + d$ or $- c + 2dN - bN > d - b$. Setting $z = N - 1$, (3) implies $- c + 2dN - bN > 2(d - b)$. So, provided $d \$ b$, (3) implies that Getting-Even is an equilibrium strategy. Setting $z = N$, (3) implies $- c + 2dN - bN > 0$. So Getting Even is also an equilibrium strategy for $d < b$.

However, Getting-Even is only subgame perfect provided $d \$ b$. To see this, suppose j deviates at time t and then reverts to Getting-Even. In period $t + 1$, j therefore plays Cooperate, while all other players play Defect. Any player i, $i \ldots j$, gets a payoff of b in period $t + 1$ and a payoff of $- c + dN$ in every subsequent period if all players play Getting-Even from period $t + 1$ onwards. If i deviates in period $t + 1$ and then reverts to Getting-Even in period $t + 2$, however, it gets a payoff of $- c + 2d$ in period $t + 1$ and a payoff of $b(N - 1)$ in period $t + 2$; thereafter, i gets $- c + dN$ every period. Deviating is therefore irrational for i provided $b - c + dN \$ - c + 2d + b(N - 1)$ or $d \$ b$. This last

---

international cooperation problems.

requirement holds by (4).

To sum up so far: like Grim, Getting-Even is individually rational. I now show that, unlike Grim, Getting-Even is also collectively rational.

Getting-Even will be collectively rational if all countries have no incentive to renegotiate the agreement. If every country other than *j* plays Getting-Even in a punishment phase, after *j* has reverted to Getting-Even, then they will each get a payoff of *b* per period. If they deviate *en masse*, however, then they will each get - *c* + *dN* per period. It will thus not be in their collective interests to deviate if

$$(b + c)/d \geq N. \qquad (5)$$

Since $d \geq b$ by assumption, (5) implies that $(d + c)/d \geq N$, and this in turn implies that all the countries called upon to punish *j* for cheating cannot do better collectively than to play Defect in the punishment phase, even if *j* plays Defect in this phase also. Agreements that satisfy (5) are not vulnerable to renegotiation. The threats needed to sustain full cooperation in these agreements are credible.

To sum up: I have shown that Getting-Even can sustain full cooperation by means of a self-enforcing agreement if (5) holds. I have *not* shown that there does not exist an alternative strategy that can do better than Getting-Even (that is, a strategy that can sustain full cooperation using a weaker punishment, and so allow full cooperation to be

sustained for a larger N). However, in the appendix I show that Getting-Even cannot bettered, as long as we hold on to the assumptions of individual and collective rationality. Result (5) is robust.

Inequality (5) tells us that the full coooperative outcome can only be sustained as an equilibrium of the repeated game if N is not "too large." Notice that, since (2) must hold for $z = 1$, $(b + c)/d < 2$. So we know that the full cooperative outcome of the generic $2X2$ prisoners' dilemma game can be sustained as an equilibrium of the repeated game. This is not a new result (see van Damme, 1989; and Myerson, 1991), but (5) shows just how special the 2-player game is. It may not be possible for even three countries to sustain the full cooperative outcome by means of a self-enforcing agreement.

Importantly, (5) tells us that the maximimal value of N that can sustain the full cooperative outcome as an equilibrium is not fixed but depends on the parameter values. Consider some examples. Suppose $b = d = 3$ and $c = 4$. Then (2) and (3) hold for $N \geq 2$, but at most 2 countries can sustain the full cooperative outcome as an equilibrium of the repeated game. Suppose instead that $b = 2$, $d = 3$, and $c = 10$. Then (2) and (3) hold for $N = 6$ and $N = 7$, but (5) says that at most 4 countries can sustain full cooperation by means of a self-enforcing agreement. Finally, suppose $b = d = 1$, and $c = 149$. Then (2) and (3) hold for $N \geq 150$ while full cooperation can be sustained as an equilibrium only so long as $N \# 150$. Keohane (1984) has argued that, for international relations problems, the number of players *is* "small," even in the case of global negotiations (in 1984, when Keohane made this argument, there were about 150

countries in the world; today there are almost 200). But the theory developed here shows that whether the international system is "small" depends on the nature of the cooperation problem.

More than this, the theory implies that the number of countries in the world is "small" only with regard to issues for which the total gains to cooperation are "small." In other words, when cooperation is needed most, the international system is least capable of sustaining cooperation by a self-enforcing agreement. To see this, notice that the gains to cooperation are $N[P_C(N) - P_D(0)] = N(-c + dN)$. The gains to cooperation are thus decreasing in c and increasing in d. But from inequality (4) we know that the maximal value of N that can sustain full cooperation as an equilibrium is increasing in c and decreasing in d. So the international system can only sustain full cooperation among *all* countries when the gains to cooperation are "small."

Does this result speak to any real world problems? I have shown elsewhere (Barrett, 1998a) that the aggregate gains to cooperation are small in the case of stratospheric ozone depletion. This is not because the world would not benefit from a ban on ozone-depleting substances. To the contrary, the reason is that the benefit of a ban is so large relative to the cost, that every industrial country would want to ban these chemicals unilaterally, even if no other country did so. The challenge to the Montreal Protocol was to make it attractive for poorer countries also to ban these substances, and for the ban by signatories to be made effective by ensuring that production would not relocate to

nonsignatory countries.[10]

5. Compliance Enforcement and Free-Rider Deterrence

The theory outlined above teaches that cooperation can be sustained by a self-enforcing treaty which incorporates a strategy of reciprocity. But Chayes and Chayes (1991, p. 313) observe that, "not only are formal enforcement mechanisms seldom used to secure compliance with treaties, but they are rarely even embodied in the treaty text." Now, the fact that such enforcement mechanisms are seldom used is entirely consistent with the theory developed here. In equilibrium, no party would deviate from the treaty because the threat to carry out the punishment is credible. Where the theory and practice of international coooperation seem to clash is in the observation that compliance enforcement mechanisms are rarely expressed in black and white. The reason may be that the theory is wrong and such mechanisms are not needed, as the Chayes's argue; or it may be that, as Downs, Rocke, and Barsoom (1996) maintain, the theory is right and the fact that such mechanisms are not incorporated in treaties implies that agreements typically do not improve much on the noncooperative outcome.

To illuminate this debate, I distinguish between free-rider deterrence and compliance enforcement by representing international cooperation as a stage game: in Stage 1, countries choose whether to be signatories or nonsignatories to an international

[10]The former problem requires the use of "carrots" or side payments. For an analysis of how carrots can aid cooperation, see Barrett (1998b). The latter problem is sometimes called "trade leakage," and is discussed in Barrett (1997).

agreement; in Stage 2, signatories choose *jointly* whether to play Cooperate or Defect; and in Stage 3, nonsignatories choose independently whether to play Cooperate or Defect. I assume that the compliance norm has been fully internalized, so that all signatories comply with the obligations they negotiatiate in Stage 3. As noted in the Introduction, this assumption is merely a tactic. I use it to see whether internalization of the compliance norm matters.

As usual, the equilibrium is found by solving the stage game backwards. Assuming that all actions are publicly observable, the strategies of each player will generally be contingent on the history of the game. However, the stage game version of the prisoners' dilemma is special in that the history of the game is irrelevant to nonsignatories; for them, play Defect is a dominant strategy. If signatories were to choose whether to play Cooperate or Defect independently, then they too would play Defect. However, signatories to a treaty do not choose their actions independently. They negotiate their choice of actions and it would be collectively irrational for them to put their signatures on a treaty that did not maximize their joint payoff.

Let k denote the number of signatories, and let signatories be identified by the subscript s and nonsignatories by the subscript n. Then, for the 2-player game, if $k = 1$ the sole signatory will play Defect and get a payoff $P_s = 0$ (if this country played Cooperate instead it would get a payoff of $-c + d$, and by (2), $-c + d < 0$), while if $k = 2$ both signatories will play Cooperate (since $-c + 2d > (b - c + d)/2$ by (3)) and get a payoff $P_s = -c + 2d$ each. Nonsignatories can do no better than to play Defect whatever signatories do

and so they get a payoff $P_n = 0$ if $k = 0$ or $k = 1$.

These payoffs can be worked out by each country before the Stage 1 game is played. So in Stage 1, each country will know the consequence of choosing to be a signatory or nonsignatory, taking as given the choice of the other country to be a signatory or nonsignatory. Assuming that a country will accede to a treaty if, in doing so, it is not made worse off, there is a unique equilibrium. It is that both countries are signatories and that both play Cooperate. The institution of the treaty coupled with the compliance norm thus transforms the dilemma game into one in which full cooperation is sustained as an equilibrium.

But full cooperation will not always be sustained as an equilibrium of the transformed game. Suppose the payoff functions are given by $P_D = 3z$ and $P_C = -4 + 3z$. Then we obtain the above result for $N = 2$. Not so if $N = 5$. For the transformed game, nonsignatories will play Defect in equilibrium. If there is only one signatory, it too will play Defect (if this country plays Defect it gets $P_S = 0$; if it plays Cooperate it gets $P_S = -1$). However, if there are two or more signatories, they will each get a higher payoff if they both play Cooperate (for example, if $k = 2$, each signatory gets $P_S = 0$ if they both play Defect and $P_S = 2$ if they both play Cooperate). And so on. It can be shown that, in equilibrium, $k^* = 2$ signatories play Cooperate and $N - k^* = 3$ nonsignatories play Defect. The full cooperative outcome is *not* sustained as an equilibrium of this game, even though the compliance norm is assumed to have been fully internalized.

To generalize even further, suppose the payoff functions for the N-player dilemma game are given by eqs. (1). Then signatories will play Cooperate provided the payoff they each get by playing Cooperate exceeds the payoff they each get by playing Defect, or $k > c/d$; otherwise, signatories can do no better collectively than to play Defect. Because play Cooperate is not an equilibrium of the one-shot prisoners' dilemma game, we know that $c/d > 1$ and so $k^* \geq 2$. As in the repeated game model, cooperation can always be sustained as an equilibrium for the special 2-player case.

Since, by assumption, full cooperation requires that all players play Cooperate, it must be true that $N > c/d$. Let k^0 be the smallest integer greater than c/d. Then we know that $k^* \geq k^0$. But when $k = k^0$, no nonsignatory would wish to accede to the treaty. To see this, notice that, if $k = k^0$, a nonsignatory gains by acceding to the treaty if $(d - b)k^0 > c - d$. But, by (2), $(d - b)z < c - d$ for all z, $0 \leq z \leq N - 1$. This is a contradiction. Once there are k^0 signatories, it would be irrational for another country to accede to the treaty. Hence, the equilibrium number of signatories must be $k^* = k^0$ (assuming that the solution is "interior"). Figure 2a illustrates the solution for $k^* < N$ and Figure 2b for the case where $k^* = N$.

Full cooperation can only be sustained as an equilibrium of this transformed game if signatories can do no better collectively than to play Defect when $k = N - 1$ and to play Cooperate only when $k = N$. The latter requirement holds by (3). The former holds provided $0 \geq -c + d(N - 1)$ or

$$(d + c)/d \geq N. \qquad (6)$$

Notice that (6) can be interpreted as saying that an agreement to play Cooperate would only come into force (that is, would only be legally binding on the countries that had ratified it) if all N countries have ratified it. Hence, k^* can be interpreted as the minimum participation level prescribed by international treaties. Of course, the case where $k^* = N$ is special. And it is a feature of most treaties that the actual number of parties usually exceeds the number prescribed by the minimum participation clause. This suggests that in the majority of treaties the minimum participation clause may serve as a coordination device rather than as a mechanism for deterring free-riding.[11]

Upon comparing (5) and (6) one finds that, if $b = d$, then the maximum number of countries that can sustain the full cooperative outcome as a self-enforcing agreement will be the same for both models. If, however, $d > b$--if there are increasing returns to cooperation--then a smaller number of countries can sustain the full cooperative outcome as an equilibrium in the repeated game model as compared to the stage game model. However, as noted in the introduction, this does not mean that the assumption of full compliance buys any additional cooperation. The stage game model is essentially one-shot; it does not allow for reactions, and so it cannot describe fully an appropriate strategy of reciprocity.

[11]See Barrett (1997), where the minimum participation clause actually emerges as an equilibrium.

The main reason for using the stage game model is to show that the vital qualitative insight of the repeated game model holds here as well. Recall that the total gain to cooperation, $N(-c + dN)$, is decreasing in c and increasing in d. By contrast, k^* is increasing in c and decreasing in d (ignoring the integer problem). This means that, for N given, k^* will tend to be "large" ("small") when the total gain to cooperation is "small" ("large"). The international system is able to sustain less cooperation the greater is the potential gain to cooperation--that is, the greater is the need for cooperation (see also Barrett, 1994).

Notice that, in equilibrium, nonsignatories get a higher payoff than signatories. Nonsignatories (of which there are $N - k^*$) free-ride. The underlying game of whether to play Cooperate or Defect is a prisoners' dilemma game, but the transformed game of whether to be a signatory or nonsignatory to the treaty is a chicken game. Each country would prefer to free-ride, but if too few countries are parties to the treaty, it is in the interests of nonsignatories to accede. Though the players are symmetric by assumption, in equilibrium they behave differently. Some are signatories and play Cooperate; some are nonsignatories and play Defect. The model can't identify which countries will be signatories and which nonsignatories (though the identities of these countries can be determined if countries make their stage 1 choices in sequence; the first $N - k^*$ countries to choose will all choose not to be signatories and the last k^* to choose will all choose to be signatories), but as the countries are symmetric this doesn't matter.[12]

[12]This will not be true when countries are strongly asymmetric; see Barrett (1998b).

The essential lesson of the stage game is that, despite the assumption of full compliance, a self-enforcing treaty may only be capable of sustaining $k^* < N$ signatories. Free-riding may be a problem for international cooperation, even if compliance isn't. At the very least, sticks are needed to deter free-riding, though the constraints on individual and collective behavior may be such that the full cooperative outcome cannot be sustained by international treaty. Large sticks may be needed to deter free-riding but large sticks may not be credible.

Though I am unable to settle the dispute about compliance, the theory developed here does broaden the debate. It suggests that, even if the Chayes's are right that compliance isn't a problem, they may be wrong that sanctions are not needed to sustain cooperation or that the international system sustains anything like full cooperation. It suggests too that Downs, Rocke, and Barsoom may be right that full cooperation typically hasn't been sustained, but that they may be wrong in implying that the reason for this is weak enforcement. Free-rider deterrence may be the greater problem.

What constrains cooperation in the stage game, as in the repeated game, is the assumption that signatories negotiate a collectively rational agreement. If we drop this requirement, then the assumption that the compliance norm has been internalized will ensure that full cooperation can always be sustained as an equilibrium. For if signatories could be sure of complying with *any* agreement, then to sustain full coooperation as an equilibrium would only require an agreement which says that each country will play Cooperate provided all others do, but that, should any other country

play Defect instead, then all the other countries will punish this defection. In general, however, such an agreement will not be collectively rational. Should one country play Defect, it will not generally be collectively rational for the remaining $N - 1$ countries to punish the deviation.

6. Conclusions

The central idea behind the theory presented here is that the institutions that sustain international cooperation must be both individually and collectively rational: individually rational because the international system is anarchic; collectively rational because countries cooperate explicitly and can renegotiate their treaties at any time. When combined, these requirements give the theory of international cooperation great cutting power. The theory predicts that the full cooperative outcome of the N-player, prisoners' dilemma can only be sustained by a self-enforcing treaty when N is "small." For global problems (that is, problems for which N is "large"), the theory predicts that full cooperation can only be sustained by a self-enforcing treaty when the gains to cooperation are "small."

These are powerful if depressing predictions. They are not, however, context-free. In a richer environment than analyzed here, it is possible that more cooperation could be sustained by a self-enforcing treaty. For example, I have shown elsewhere (Barrett, 1997) how linking the provision of a global public good to international trade allows the space of punishment strategies to be expanded. The credible threat of trade sanctions

may be able to sustain cooperation where the threat to withdraw provision of a public good cannot. In fact, it is by the threat of imposing trade sanctions that free-riding has been deterred in the Montreal Protocol. Moreover, the threat of trade sanctions has also helped to enforce compliance with the agreement. But even where the strategy space can be expanded, the twin requirements of individual and collective rationality may prevent countries from sustaining full cooperation. Certainly, there should be no presumption that the international system, attached as it is to the principle of sovereignty, is always capable of sustaining full cooperation. *That* conclusion, however unwelcome, does seem robust.

Appendix

Getting-Even, as defined in this paper, assumes that, were j to deviate, then *all* the $N - 1$ other countries must play Defect in a punishment phase. In doing so, these countries harm themselves as well as j, and this is what makes sustaining full cooperation more difficult as N increases. So the question arises: can an alternative strategy--one that harms the $N - 1$ other countries less--sustain full cooperation?

This won't be possible for $N = 2$, because obviously j must be punished for deviating and when $N = 2$ there is only one other country that can do so. However, it isn't obvious that, when $N > 2$, *all* the other $N - 1$ countries must play Defect in a punishment phase. Let us then suppose that m of the $N - 1$ other countries play Defect in the punishment phase (so that $N - m - 1$ of the $N - 1$ other countries play Cooperate in the punishment phase). Call this the m-Getting-Even strategy.

Full cooperation can be sustained as an equilibrium if two conditions are satisfied. First, we require that j cannot do better than to play m-Getting-Even given that every other country does so; that is, we require

$$\max (b(N - m - 1), -c + d(N - m)) \# -c + dN. \qquad \text{(A.1)}$$

By (2), $b(N - m - 1) > -c + d(N - m)$. So (A.1) implies

$$b(N - m - 1) \# - c + dN. \tag{A.2}$$

We also require that each of the $N - 1$ other players cannot do better than to play m-Getting-Even in a punishment phase. That is, we require

$$b(N - m) \$ - c + dN \tag{A.3}$$

for the m countries that play Defect in the punishment phase and

$$- c + d(N - m) \$ - c + dN \tag{A.4}$$

for the $N - m - 1$ other countries that play Cooperate in the punishment phase. But (A.4) reduces to $- dm \$ 0$, implying that we must have $m = 0$. Of course, if $m = 0$--if none of the $N - 1$ other countries plays Defect in a punishment phase--then j will not be punished. So the m-Getting-Even strategy cannot sustain full cooperation as an equilibrium, except for the special case where $m = N - 1$ (for in this case, (A.4) drops out and (A.3) reduces to (5)), provided we require that *all* the $N - 1$ countries not want to renegotiate the agreement.

Now, it might be argued that this requirement is overly strong. Suppose we allow transfers between the $N - 1$ countries called upon to punish j. Then renegotiation will be prevented if the $N - 1$ other countries receive *on average* at least as large a payoff when implementing the strategy as when reverting to full cooperation. However, collective

rationality will in this case require that the $N - 1$ other countries choose m so as to maximize their aggregate payoff in the punishment phase. That is, instead of (A.3) and (A.4) we require

$$\max_{m} \{mb(N - m) + (N - m - 1)[- c + d(N - m)]\} \geq (N - 1)(- c + dN) \tag{A.5}$$

Solving the LHS of (A.5), the first order condition requires

$$b(N - 2m) + c - d[2(N - m) - 1] = 0 \tag{A.6}$$

The second order conditions for a maximum require $2(d - b) < 0$. However, by assumption, $d \geq b$. Hence, the solution to the maximization problem must lie at a corner; (A.5) will require either $m = 0$ or $m = N - 1$.

Of course, (A.2) must hold, and this implies

$$m \geq [b(N - 1) - (- c + dN)]/b \tag{A.7}$$

By (2), the numerator on the RHS of (A.7) is positive. So the solution must require $m > 0$. $m = N - 1$ will be the solution to the LHS of (A.5) if the aggregate payoff of the $N - 1$ other countries is at least as high when $m = N - 1$ as when $m = 0$. Upon substituting, we

require

$$b\,\$ - c + dN. \tag{A.8}$$

But this is the same as (5). Hence, there does not exist an alternative individually and collectively rational strategy that can improve on Getting-Even.

References

Axelrod, R. (1984), *The Evolution of Cooperation*, New York: Basic Books.

Axelrod, R. and R.O. Keohane (1985), "Achieving Cooperation Under Anarchy: Strategies and Institutions," *World Politics*, **38**: 226-254.

Barrett, S. (1994), "Self-Enforcing International Environmental Agreements," *Oxford Economic Papers*, **46**: 878-894.

Barrett, S. (1997), "The Strategy of Trade Sanctions in International Environmental Agreements," *Resource and Energy Economics*, **19**: 345-61.

Barrett, S. (1998a), "Montreal v. Kyoto: International Cooperation and the Global Environment," paper prepared for the Workshop on Global Public Goods: A New Approach to International Development Cooperation, United Nations Development Programme, New York City, 22 June 1998.

Barrett, S. (1998b), "Cooperation for Sale," mimeo, London Business School.

Bernheim, B.D., B. Peleg, and M.D. Whinston (1987), "Coalition-Proof Nash Equilibria I. Concepts," *Journal of Economic Theory*, **42**: 1-12.

Chayes, A. and A.H. Chayes (1991), "Compliance Without Enforcement: State Regulatory Behavior Under Regulatory Treaties", *Negotiation Journal*, 7: 311-31.

Chayes, A. and A.H. Chayes (1993), "On Compliance", *International Organization*, 47: 175-205.

Chayes, A. and A.H. Chayes (1995), *The New Sovereignty*, Cambridge, Mass: Harvard University Press.

Downs, G.W., D.M. Rocke and P.N. Barsoon (1996), "Is the Good News About Compliance Good News About Cooperation?" *International Organization*, **50**: 379-406.

Farrell, J. and E. Maskin (1989), "Renegotiation in Repeated Games," *Games and Economic Behavior*, **1**: 327-60.

Friedman, J. (1971), "A Noncooperative Equilibrium for Supergames," *Review of Economic Studies*, **38**: 1-12.

Fudenberg, D. and E. Maskin (1986), "The Folk Theorem in Repeated Games with Discounting and Incomplete Information," *Econometrica*, **54**: 533-554.

Kandori, M. (1992), "Social Norms and Community Enforcement," *Review of Economic Studies*, **59**: 63-80.

Keohane, R.O. (1984), *After Hegemony*, Princeton: Princeton University.

Keohane, R.O. (1986), "Reciprocity in International Relations," *International Organization*, **40**: 1-27.

Keohane, R. O. and E. Ostrom (1994), "Introduction," *Journal of Theoretical Politics*, **6**: 403-428.

Myerson, R.B. (1991), *Game Theory: Analysis of Conflict*, Cambridge: Harvard University.

Olson, M. (1965), *The Logic of Collective Action*, Cambridge: Harvard University.

Ostrom, E. (1990), *Governing the Commons*, Cambridge: Cambridge University Press.

Schelling, T.C. (1978), *Micromotives and Macrobehavior*, New York: W.W. Norton & Co.

Snidal, D. (1994), "The Politics of Scope: Endogenous Actors, Heterogeneity and Institutions," *Journal of Theoretical Politics*, **6**: 449-472.

van Damme, E. (1989), "Renegotiation-Proof Equilibria in Repeated Prisoners' Dilemma," *Journal of Economic Theory*, **47**: 206-217.

Young, O.R. and G. Osherenko (eds.) (1993), *Polar Politics*, Ithaca: Cornell University Press.