The Stability of International Environmental Coalitions with Farsighted Countries: Some Theoretical Observations

by

Giulio Ecchia* and Marco Mariotti**

Abstract: we study a three-country model of international environmental agreements where countries may choose either to limit their emissions or to behave noncooperatively. First, we provide a taxonomy of various kinds of strategic situations. Then, by applying some recently developed game-theoretic techniques, we show that if countries are 'farsighted' then there is scope for self-enforcing cooperation in several such situations.

We would like to thank C. Carraro and Francesca Moriconi for comments. There have also been very helpful discussions on related topics with S.Brams. Financial support from Fondazione ENI Enrico Mattei is gratefully acknowledged.

*University of Bologna

**University of Manchester, UK

Address for correspondence: Marco Mariotti, University of Manchester, Economics Department, Manchester M13 9PL, UK (tel: 0161-275-4875; fax: 0161-275-4812; e-mail: m.mariotti@man.ac.uk)

1. Introduction

Environmental issues have become the subject of complex international negotiations. The impulse to subscribing International Environmental Agreements (IEAs) comes from the recognition on the part of various countries that they have an interest in behaving cooperatively in the use of environmental resources. The strength of such an impulse is documented both by the number of IEAs in force (more than a hundred) and by the number of countries participating in the negotiations (for example, the Framework Convention on Climate Change, held in Rio De Janeiro in 1992, is signed by the representatives of 154 countries). However, the incentive for cooperation in the use of environmental resources is limited. Although the situation resulting from cooperation is better, for most or even all countries, than the situation resulting from *completely* noncooperative behaviour, each country would prefer the situation in which it behaves noncooperatively while the other countries cooperate amongst themselves. This free-ride feature reduces the incentive to sign IEAs, in the case where they are binding, or the incentive to abide by an agreement once it is signed, in the case where they are not binding. It is generally argued that IEAs do not have binding force, and, although the demarcation between the two cases is not completely clear-cut, we will follow this tradition here[1]. Assuming then that IEAs are not binding, it is necessary that they should be *self-enforcing*. Loosely speaking, this means that all countries must find it in their interest to abide by the agreement rather than free-ride. However, as soon as one tries to make this concept more precise, several questions arise. For example, when considering the opportunity to free-ride, how should a country allow for the possibility of retaliation on the part of the other countries? Or, should one consider only individual or also simultaneous deviations? In this paper, we propose to define more precisely, by using game-theoretic ideas, a self-enforcing international agreement, and to explore the scope for implementing them in different strategic environments.

This topic has been studied in some recent papers (e.g. Barrett (1994a,b, 1995), Carraro and Sinicalco (1993), Hoel (1992), Botteon and Carraro (1995), Stahler (1994)). We will find it convenient to present our main ideas in comparison with the notion of self-enforcing IEAs contained in those papers.

2. Static Self-Enforcing International Agreements

[1] To what extent an agreement should be considered binding or not depends on the severity of the punishment associated with the breach of contract. If it is true that often no explicit positive punishment is written for a country which does not abide by the IEA it has signed, it should also be recognized that there are various disdvantages (in terms of international relations or even in financial terms) for such a misbehaving country.

In this section we present the broad structure of the models considered in the literature (e.g Chandler and Tulkens (1994), Barrett (1994a), Carraro and Sinicalco (1993), Hoel (1992), Botteon and Carraro (1995)), and the notion of self-enforcingness considered, which we call 'static' for reasons that will become apparent later.

There is a set N of n countries, all of which emit a pollutant which is damaging for the shared environmental resources. For each country, a damage and a cost-abatement function are defined. The damage functions,

(1) $d_i: E \rightarrow \Re$,

express the damage suffered by country $i \in N$ as a function of the emissions of all countries. Here $E = E_1 \times E_2 \times ...\times E_n$ denotes the space of potential emissions on the part of all countries; so, $d_i(e_1, e_2,...,e_n)$ represents the damage suffered by country i when the emissions of country $j \in N$ is e_j (alternatively, one can define a benefit function of the abatement levels). The function $d_i(.)$ is assumed to be increasing and concave in all arguments.

The cost-abatement functions,

(2) $c_i: E_i \rightarrow \Re$,

express the the cost incurred by a country to set its emissions at a given level. They are assumed to be decreasing and concave.

Clearly, much is left exogenous in this description: both the choice sets E_i and the specific functional forms $c_i(.)$ and $d_i(.)$ will depend on technology, economic structures, political variables and so on. Moreover, an important simplification is usually made, which is to assume that $d_i(.)$ has the structure

(3) $d_i(e) = f_i(g(e))$, with $e \in E$, $g: E \rightarrow \Re$, $f: \Re \rightarrow \Re$.

That it, there exists an 'aggregator' function g(.) of the vector of emissions such that the damage for each country only depends, through f(.), on the aggregator level and not on the specific vector of emissions that determines it. Even more in particular, a 'Cournot hypothesis' is made, whereby the aggregator function g(.) is simply the summation operator,

(4) $g(e) = e_1 + e_2 + ...+ e_n$ for all $e \in E$.

The usefulness of these simplifications will now become apparent. The *noncooperative outcome* is simply defined as the vector of emissions e* that constitutes the Nash equilibrium of a strategic form game $G = (E_i, u_i(.))_{i=1,...,n}$ between the n countries, where E_i is the strategy space of country i and $u_i(.)$ is its payoff function defined by

(5) $u_i(e) = d_i(e) + c_i(e_i)$.

Assuming interior solutions, a Nash Equilibrium is characterized by the first-order conditions

(6) $\partial u_i/\partial e_i = 0$,

or, with the aggregator simplification above,

(7) $(\partial f_i/\partial g)(\partial g/\partial e_i) + \partial c_i/\partial e_i = 0$,

and finally with the 'Cournot simplification' of g(.),

(8) $(\partial f_i/\partial g) + \partial c_i/\partial e_i = 0$.

This set of first-order conditions has the following remarkable property. Since the functional forms $f_i(.)$ and $c_i(.)$ are country-specific and do not depend on the other countries' emission levels, the best-reply functions $e_i: E_{-i} \rightarrow E_i$ determined by (8) are constant, that is, the optimal noncooperative level of emissions for country i only depends on the parameters that specify the functional forms $f_i(.)$ and $c_i(.)$, but not on the emission levels of the other countries. This property of 'orthogonality' is noted and discussed at length by Carraro and Siniscalco (1993).

The fully cooperative outcome is defined as the solution e^{FC} to maximizing the total benefits

(9) $U(e) = u_1(e) + u_2(e) + ... + u_n(e)$.

One also defines the outcomes resulting from intermediate levels of cooperation, where, say $k \leq n$ countries sign an agreement to cooperate and the remaining (n - k) countries behave noncooperatively. More precisely (we follow here Barrett (1994a)), the nonsignatories play a 'reduced' noncooperative game by taking the emission levels of the signatories as given, while the signatories maximize their joint payoffs taking into account the behaviour of the nonsignatories. Note that, with the simplifying asumptions made above, the noncooperative outcome of the reduced subgame between the (n - k) nonsignatories has the same property of orthogonality noted before, and each nonsignatory country best-reply now is parametrized on the aggregate emission levels of the

signatories. Only in order to make the main points more simply, let us assume symmetry between countries (the issue of asymmetry is an important one: see for example Carraro and Botteon (1995)). Then the noncooperative solution of the reduced game only depends on the number k of signatories. One can then define the equilibrium payoff of each of the nonsignatories as $\nu(k)$. Similarly, one can then define the payoff of the signatories as $\sigma(k)$. Obviously $\sigma(n) = u_i(e^{FC})$, and $\nu(0) = u_i(e^*)$.

The question which is now addressed is: can full cooperation be supported in a self-enforcing agreement? Or, more in general: how many signatories can there be in a self-enforcing agreement? In order to answer these question one has first to define an appropriate notion of self-enforcingness. The one that seems to most popular at the moment has been developed in the analysis of cartels (d' Aspremont and Gabsewicz. (1986), Donsimoni *et al.*(1986)). The idea is simple: for a coalition of signatories to be stable, two conditions must be met. First, no signatory country can gain from withdrawing from the coalition and acting noncooperatively. Let us call this type of stability *lower stability*: there is no force that acts in the direction of *shrinking* the coalition. Second, no nonsignatory country can gain from joining the coalition and acting cooperatively. Let us call this type of stability *upper stability*: there are no forces that tend to broaden the coalition (in addition, the coalition must be also *viable*, in the sense that it makes positive profits with respect to the benchmark represented by the noncooperative outcome profit level). Thus broadly and informally stated, these conditions are relatively uncontroversial, but 'the devil is in the detail': let us see how expressions such as 'gain from joining' and 'gain from withdrawing' are formalized. In terms of out previous notation, the two stability conditions are translated in:

(a) (Lower stability) $\nu(k - 1) \leq \sigma(k)$;

(b) (Upper stability) $\nu(k) \geq \sigma(k + 1)$[2].

These conditions are deceptively appealing. There are two issues to be considered, '*myopia*' and the problem of *joint deviations*.

In a sense, these definitions embody a feature of 'myopia' on the part of the countries, in that each of them performs its calculations ignoring the possible reactions of the other countries to its decision of altering the existing coalition structure (either by moving to free-riding as a signatory or by moving to cooperation as a non-signatory). Countries only look 'one step ahead'. For example,

[2] Donsimoni *et. al.* (1986) have shown conditions under which coaltions which are stable in this sense exist.

suppose that there are three countries, and the 'status quo' is a coalition formed by A and B, while country C free-rides. Upper stability requires that country C has no incentive to join A and B in the emission-reducing agreement. Now suppose that the coalition {A,B} is *not* upper stable, because indeed country C can gain by forming the grand coalition {A,B,C}. But what if, when C does so, either A, or B, or both, can then gain by abandoning (i.e. free-riding on) the grand coalition just formed? In other words, what if, when C joins {A,B}, lower stability is lost? If so, it may happen (and it seems likely) that C will not be better off, compared to the initial situation, after A or B have abandoned the coalition it just joined. But then, the initial situation may not suffer from lack of upper stability after all! If C looks 'two-steps' ahead, it should refrain from joining {A,B} in the anticipation of the dire consequences to come. Similar examples can be given regarding upper stability. These considerations seem to undermine the whole logic of definitions (a) and (b).

The second issue is the problem of joint deviations. Here each *individual* country compares profits and losses from its decision to join or to abandon a coalition. But in a context in which countries can communicate and agreements can be signed, it would seem sensible to allow for the possibility that *coalitions of countries* may discuss their joining or abandoning a given coalition of signatories. Of course, such collective agreements will have to be themselves self-enforcing, and should therefore be subjected to the same scrutiny to which the initial agreement is subjected. Once again, the logic of definitions (a) and (b) does not seem convincing. For, it may be the case, for example, that two nonsignatories would gain from joining an upper stable coalition together (after signing an agreement to do so), while either country individually would not.

Having made these points, however, it must be said that it is not at all clear or uncontroversial how to capture the flavour of the above discussion in a formal definition. One approach that has been followed in the literature to introduce some elements of farsightedness has been to model the situation by means of an *infinitely repeated game* (see e.g. Barrett (1994a,b, Stahler (1994))) to which the idea of *renegotiation proofness* is applied. This has the merit of endogenizing the sequence of countermoves following a deviation. But such a modelling strategy is problematic in two respects. First, a (subgame perfect) equilibrium may need to be supported by the threat of long retaliations which, given the nature of the game being played, are not necessarily very realistic. The main reason why we depart from the supergame approach, however, is precisely that what it seems to capture is the effect of balancing future 'punishments' with current rewards from free-riding. What we are seeking to represent, on the contrary, is merely how a country's willingness to sign an IEA is affected by the the ultimate consequences that its decision leads to. We aim to ignore deterrence

issues and time-preference and, unlike the case of repetaed games, we assume that payoffs are only received when a stable structure is achieved. Such a stable structure should be seen as a long-term situation, while the moves and countermoves that lead to it are short-term and payoff-irrelevant.

The principal difficulty in this project seems to be the recursive nature of the proposed logic. Consider again our discussion of the myopia feature of stability. Why should the reasoning of country C be limited to two steps ahead? Certainly, the situation which is determined after it has joined countries A and B in the emission limiting agreement, and after country A or B has in turn defected, may not be the final one. What is needed is a definition of self-enforcigness where *each* potential status quo is treated symmetrically. At the methodological level, it should be observed that, this does not enatil assuming that real countries are infinitely farsighted. Rather, one wants to avoid that any arbitrary fixed degree of farsightedness is crucial to determine the outcome of the model. For example, one of the main messages of the literature sketched so far has been the fact that, in general, the fully cooperative outcome e^{FC} cannot be supported if the stability conditions (a) and (b) are to be satisfied. If it turned out that this result depended crucially on countries looking exactly one step ahead, then one would be skeptical in considering partial cooperation as a truly stable outcome.

These issues have been amply studied in the more recent abstract game-theoretic analysis of coalition formation (Bernheim *et al.* (1987), Ray (1989), Greenberg (1990), Chwe (1994), Xue (1994), Ray and Vohra (1992), Mariotti (1995), to mention but a few). In the sequel of this paper we apply to the particular problem at hand the notion of farsightedness developed by Mariotti (1995). Because the analysis becomes quickly very complex when coalitional stability is studied in this way, we will limit ourselves to considering a simplified three-country model. This will be enough to draw some insights on the strategic factors affecting international environmental agreements.

3. The model

There is a set of three countries, $N = \{1,2,3\}$. Each country may choose between two strategies: either set the level of emissions at a 'cooperative' level (strategy C), or behave uncooperatively (strategy P, for 'Pollute'). One interpretation of these choices is in terms of the models sketched in the previous section. That is, given whatever coalition structure is currently prevailing, a country can either decide to free-ride (strategy P) by producing its noncooperative level of emissions, or to join the coalition of signatories (strategy C), and set whatever level of emissions is optimal for the coalition, given the behaviour of the nonsignatories. For our purposes, a drawback of this interpretation is that, for any country, if the other two countries are not cooperating (there is no

coalition) then the strategies P and C should be equivalent from the point of view of that country: the best response of a one-country coalition is just the noncooperative optimum. This would make the complexity of our model somewhat too low to derive interesting results. We then add the following feature to the above interpretation: when none of the other countries is behaving cooperatively, each country has the option of *unilaterally* reducing its own emissions (strategy C), rather than set the noncooperative level of emissions (strategy P). This extension is not a real limitation of the model, for we will not necessarily assume that it is in a country's interest to play P, even if it is available. Indeed, we will consider, among others, situations where a country gains from playing P rather than C no matter what the other countries do (that is, playing P is a dominant strategy).

Each country i has a payoff $u_i(s)$ associated with any strategy profile $s \in \{C,P\} \times \{C,P\} \times \{C,P\}$. We assume that pollution on the part of other countries affects a country negatively irrespectively of the behaviour of the country (polluting or cooperative) and of the identity of the polluters: so, $u_i(.)$ increases with the number of other countries who play cooperatively. That is, for each $i \in N$ and for $s_i \in \{C,P\}$,

(10) $(s_j)_{j\in N\setminus\{i\}} = (C,C)$ and $(s'_j)_{j\in N\setminus\{i\}} = (P,C) \Rightarrow u_i(s) > u_i(s')$,

(11) $(s_j)_{j\in N\setminus\{i\}} = (C,C)$ and $(s'_j)_{j\in N\setminus\{i\}} = (C,P) \Rightarrow u_i(s) > u_i(s')$,

(12) $(s_j)_{j\in N\setminus\{i\}} = (C,P)$ and $(s'_j)_{j\in N\setminus\{i\}} = (P,P) \Rightarrow u_i(s) > u_i(s')$,

(13) $(s_j)_{j\in N\setminus\{i\}} = (P,C)$ and $(s'_j)_{j\in N\setminus\{i\}} = (P,P) \Rightarrow u_i(s) > u_i(s')$.

Imagine the following procedure. At each stage of the process, a strategy profile $s \in \{C,P\} \times \{C,P\} \times \{C,P\}$ is the current satus quo. Then, any coalition of countries $S \subseteq N$ may form and propose (or threaten) to deviate to a different set of strategies, that is, a different status quo. If somebody else, either coalitions or an individual country, deviates, all member countries of S are free to propose to deviate further. The deviating coalition may be S itself: there is no permanent committment here, just a sequence of provisional status quo's on which the current agreements are conditioned, in the sense that agreements between member countries are no longer binding when the status quo changes. The countries are only interested in the payoff associated with a *permanent* status quo. So, the process continues in this way until there is a status quo from which nobody wishes to deviate. At this point the countries receive their payoffs.

Note that, although in order to fix ideas we are interpreting this as a model of negotiations (that is, all countries are only making proposals and counterproposals), our assumption that there are no 'interim' payoffs would allow one to interpret the model as a real sequence of moves and countermoves from a staus quo to the other.

We will define a notion of equilibrium called the Coalitional Equilibrium, which is a kind of subgame perfect equilibrium for a game where the players are all the possible coalition of countries (note that, strictly speaking, it is not possible to speak of subgame perfection here, since the strategic structure, although sequential and of perfect information, is not representable by means of a tree: for example, there may be cycles[3]). That is, we take coalitions, rather than individual countries, as the decision units (this is simply a way to formalize expressions such as 'coalition S gains from ect.'). We will have to be careful in distinguishing the strategies C and P, which refer to individual countries' behaviour, and the 'coalitional strategies', which express the behaviour of coalitions: for S $\subseteq$ N, a *coalitional strategy* is a map

(14) γ_S: $\{C,P\} \times \{C,P\} \times \{C,P\} \rightarrow \{C,P\}^S$,

where $\{C,P\}^S$ denotes the strategy space for coalition S, that is,

(15) $\{C,P\}^{\{1\}} = \{C,P\}$,

(16) $\{C,P\}^{\{1,2\}} = \{C,P\}^{\{2,3\}} = \{C,P\}^{\{2,3\}} = \{C,P\} \times \{C,P\}$,

(17) $\{C,P\}^N = \{C,P\} \times \{C,P\} \times \{C,P\}$.

A coalitional strategy for coalition S, then, specifies the proposal of coalition S at each possible status quo. For example, the expression $\gamma_{\{1,2\}}(C,C,P) = (P,P)$ means that if, at the current status quo, countries 1 and 2 are behaving cooperatively while country 3 is free-riding, then *if* countries 1 and 2 form a coalition, then they will jointly decide to pollute as well. Note carefully the *conditional* interpretation of a coalitional strategy, which specifies what a coalition plays conditional on its formation. In the example above, it could be the case that, say, $\gamma_{\{2,3\}}(C,C,P) = (C,C)$, meaning that *if* countries 2 and 3 form a coalition instead, then they will play cooperatively. What will happen depends on which coalition forms. At the beginning, no coalitions will have deviated yet, and the

[3] Moreover, unlike in a standard extensive form, nodes are not 'private property': at each node any coalition can form and move.

behaviour of the coalition will simply be the reaction to some initial proposal. It will turn out that, in order to specify the *set* of equilibrium proposals one does not need to clarify how the initial proposal comes about (clearly, which strategy profile happens to be proposed initially may be important to establish which of the equilibrium proposals will be agreed upon; however, in the present analysis we eschew the issue of selecting from the equilibrium set).

A coalitional strategy profile γ specifies a coalitional strategy for each coalition. We need some further definitions. Let Γ be the set of all possible coalitional strategy profiles. Given a status quo $s \in \{C,P\} \times \{C,P\} \times \{C,P\}$, a coalitional strategy profile γ determines the possible sequences of deviations and counterdeviations from s, such that two successive status quo's, s' and s'', are linked by the relation

(18) $(s''_i)_{i \in S} = \gamma_S(s')$ for some coalition $S \subseteq N$,

(19) $(s''_i) = s'_i$ for $i \in N \backslash S$.

That is, given γ, it possible to move from s' to s'' if there is a coalition S to which γ_S prescribes to play its part of s'', while all the other countries stay put at s'.

A status quo s'' is *reachable from s via* γ if there such a sequence of deviations leading from s to s''.

A status quo s'' is *terminal from s via* γ if it is reachable from s via γ and moreover γ prescribes to all coalitions to stay at s'' once it is reached. Let $\tau(s,\gamma)$ denote the set of terminal satus quo's from s via γ.

A status quo and a coalitional strategy, by determining a certain set of terminal points, determine a set of possible payoffs for each country. Let $f_i(s,\gamma)$ denote the set of payoffs for country $i \in N$ when the current status quo is s and the coalitional strategy profile is γ. So, $f_i(s,\gamma) = \{u_i(s') | s' \in \tau(s,\gamma)\}$.

If, given a status quo, a coalition of countries forms and deviates to a different coalitional strategy, it will determine a different set of possible terminal points, and therefore a different set of possible payoffs. Therefore we need a way to compare different sets of payoffs. We make the following behavioural asumption, which is a combination of 'optimism' and 'pessimism' as defined

by Greenberg (1990). A set of payoffs K is better than another set of payoffs K', written K > K', if there is at least one payoff in K which is better than any of the payoffs in K'.

Thus equipped, we now define an equilibrium concept:

A *Coalitional Equilibrium (CE)* is a pair (s^*,γ^*), of a strategy profile $s^* \in \{C,P\} \times \{C,P\} \times \{C,P\}$ and a coalitional strategy profile $\gamma^* \in \Gamma$ such that

(i) $(s^*_i)_{i \in S} = \gamma^*_S(s^*)$;

(ii) for all $s \in \{C,P\} \times \{C,P\} \times \{C,P\}$, for all $S \subseteq N$, there is no coalitional strategy γ_S such that, for all $i \in S$, $f_i(s,\gamma_S) > f_i(s,\gamma^*)$;

(iii) let s be any strategy profile; suppose that, for some $S \subseteq N$, $\gamma^*_S(s) \neq (s_i)_{i \in S}$; then, $f_i(s,\gamma^*) > u_i(s)$ for all $i \in S$.

If (s^*,γ^*) is a CE, we say that s^* is *an equilibrium point supported by* g^*.

In words, the idea behind this notion of equilibrium is very simple.

Condition (i) requires that, given a coalitional strategy profile, for a status quo to be an equilibrium point (with respect to the given coalitional behaviour), no coalition should want to deviate from it once it is reached.

Condition (ii) then imposes some equilibrium requirements on the coalitional strategies themselves, by introducing both a 'Nash' and a 'subgame perfection' feature. It requires that no coalition should be able to gain by switching to an alternative pattern of behaviour, that is, to an alternative coalitional strategy; and this must be true not only at the equilibrium status quo, s*, but at *any* status quo. When s = s*, this condition is a Nash equilibrium condition for coalitional strategies. For all other s, it applies at 'out of equilibrium' status quo's. In this sense, it is a requirement akin to subgame perfection, in that it rules out equilibria which are supported by countries making 'non-credible' threats or promises to pollute or to cooperate (a similar requirement appears in Harsanyi (1974), with whose model ours shares some features).

Finally, condition (iii) says that deviations from the current status quo should be 'motivated', in the sense that if a coalition departs from the current status quo, it must have some hope of ending up at a terminal point which is preferred to the status quo by all countries in the coalition (Mariotti

(1995)) shows that, without this condition, a 'folk-theorem' would apply: all strategy profiles could be supported as equilibrium points for some set of coalitional strategies).

With reference to the discussion of stability at the end of the previous section, it should be now clear how we intend to capture the 'farsightedness' of countries. By defining the coalitional equilibrium essentially as an equilibrium in coalitional *strategies*, we have implicitly assumed that when considering deviations from a current status-quo, any country of coalition of countries traces the ultimate consequences of such a deviation, without naively considering the new status quo as the final one.

Another important point is worth mentioning. Whenever a country considers its choice of an environmental strategy, the calculation of the consequences associated with each strategy must depend on the country's *expectations* concerning the reactions of the other countries. One of the main conceptual problems in the type of analysis we are conducting seems to be how to define such expectations. Our resolution has been to define expectations endogenously, as (Nash) *equilibrium expectations*. Indeed, as is usual in game-theoretical equilibrium analysis, a strategy has two roles here: one is to prescribe behaviour to a coalition, while the other is to define the expectations of the other countries concerning that coalition. This may be contrasted, for example, to Greenberg (1990) where expectations are defined *exogenously*, independently on the equilibrium concept (in this case what is an equilibrium situation will depend on these exogenous expectations); or to Chwe (1994), where the issue of expectation formation is by-passed by imposing an abstract and (presumably) intuitively appealing notion of *farsighed stability* (so, the spirit of his approach is close to that embodied in the literature on cartel stability discussed in the previous sections).

4. Symmetric Games: Identifying 'Chicken' versus 'Prisoner's Dilemma' Situations

Although we have imposed some natural restrictions on payoffs (equations (10) to (13)), there is still room in the model for different specifications, even assuming, as we shall do here, symmetry between countries. It will turn out that the exact specification of the strategic structure of the interaction between countries may be important -sometimes so in a non-obvious way- to determine what types of coalitional equilibria emerge. Our aim here is to provide some insights on how the possibility of environmental cooperation between countries depends on the relations between payoffs. In particular, we shall distinguish between two main specifications, which we dub *'Chicken'* and *'Prisoner's Dilemma'* (the terminology derives from the fact that the specifications we consider for our three-person games are obvious extensions of the well-known two-person games). Some

features are shared by the two cases: firstly, it always better for a country to change to free-riding from a status quo of full cooperation; secondly, the status quo of full cooperation Pareto-dominates the status quo of complete noncooperation. So, in both cases, there are myopic *individual* incentives to destabilize a grand agreement between all countries to limit pollution, but such an agreement is better, from the *collective* point of view, than no agreement whatsoever. Formally, we assume:

(20a) $s_i = P, s_j = C \text{ for } j \in N\backslash\{i\} \Rightarrow u_i(C,C,C) < u_i(s)$.

(20b) $u_i(C,C,C) > u_i(P,P,P)$.

The Prisoner's Dilemma case is characterized by the fact that to pollute is always a *dominant strategy* for each country. This is therefore the case where the myopic individual incentives to free-ride are the strongest:

Prisoner's Dilemma: $s_i = P, s'_i = C \Rightarrow u_i(s) > u_i(s')$

The Chicken case is characterized by the fact that the worst possible status quo is the one where all countries pollute. So, although the myopic individual incentives to pollute given in (20a) remain, there is some incentive to destabilize the status quo which is most damaging for the environment as well:

Chicken: $s \neq (P,P,P) \Rightarrow u_i(s) > u_i(P,P,P)$.

The kind of situations captured by the Chicken case are those where some sort of 'environmental' disaster is faced. The environmental damages faced in case of generalized noncooperation would be so grave that even at the individual level countries would prefer to unilaterally limit their polluting emissions. The Prisoner's Dilemma refers on the contrary to situations where, although the environmental damages caused by the emissions are recognized, they are not perceived to be so grave as to override individual, opportunistic considerations[4].

Unlike the two-player case, our three-player structure allows us to distinguish, for each case, two subcases. For the Prisoner's Dilemma, we define a *weak* subcase in which, fixed the strategy played by one country, the game played by the remaining two countries is a two-player Prisoner's Dilemma (and hence the cooperative outcome Pareto-dominates the fully noncooperative one); and a

[4] Carraro and Siniscalco (1993) have been the first to draw attention to the distinction between Chicken and Prisoner's Dilemma situations. They argue that most real situations are Chicken ones, and infer from this a greater scope for cooperation. Our analysis, though, will not support this inference.

strong subcase in which, fixed the strategy of one country, the cooperative outcome is Pareto-dominated, in the two-player subgame played by the other two countries, by the fully noncooperative one. The terminology is due to the fact that the basic flavour of the Prisoner's Dilemma situation seems to be enhanced in what we call the strong version, where the incentives to cooperate are truly minimal. While in the weak version there are at least *collective* incentives to cooperate within subcoalitions, not even these incentives are present in the strong version.

Strong Prisoner's Dilemma: $s_i = s_j = P$, $s'_i = s'_j = C \Rightarrow u_i(s) > u_i(s')$, $u_j(s) > u_j(s')$.

Weak Prisoner's Dilemma: $s_i = s_j = P$, $s'_i = s'_j = C \Rightarrow u_i(s) < u_i(s')$, $u_j(s) < u_j(s')$.

For the Chicken case we distinguish the following subcases. In the *strong* version, each country only has an (individual) incentive to free-ride when the status quo is (C,C,C) (condition 20a), but prefers to limit emissions when *at least* one other country is polluting. In the *weak* version, a country prefers the cooperative strategy only when *both* other countries are playing P, but if at least one other country behaves cooperatively, then it prefers to pollute. Again, the terminology is due to the fact that the basic feature of the game under consideration seems to be enhanced in its strong version. Here, the 'disaster' character of a pollution situation occurs when even only one country pollutes in the strong version, while in the weak version this 'disaster' character only emerges if at least two countries pollute.

Strong Chicken: $s_j = P$ for some $j \in N\backslash\{i\}$, $s_i = P$, $s'_i = C \Rightarrow u_i(s') > u_i(s)$.

Weak Chicken: $s_j = C$ for some $j \in N\backslash\{i\}$, $s_i = P$, $s'_i = C \Rightarrow u_i(s') < u_i(s)$.

It would perhaps appear that the task to create an emission reducing agreement is most difficult in the Strong Prisoner's Dilemma, while it is easiest in the Strong Chicken game, and that consequently we should expect the Coalitional Equilibria to display 'less cooperation' (only partial agreements or even no agreement at all) in the first case than in the second case, where the scope for full cooperation seems to be widest. In general, it would also appear that one should expect more cooperation in the Prisoner's Dilemma cases than in the Chicken cases. In the next section, we shall see that the logic underpinning agreements between 'farsighted' countries is a great deal more subtle than that.

5. 'Chicken' versus 'Prisoner's Dilemma' Situations: Analysis

In this section we study the Coalitional Equilibria of the four situations defined in the previous section. In the interest of readability, instead of using symbolic notation we will consider examples with specific numerical values for the payoffs. There is no loss of generality whatsoever in doing so: our solution concept and the definition of our cases only depend on the *ordinal* relations between the payoffs, not on the actual numbers employed.

As is customary, we use a matrix representation where one country chooses rows, the other country chooses columns and the third country chooses boxes. In this and the following examples, we will accordingly refer to the countries as Row, Column or Box as is the case.

One additional warning: again in the interest of readability, we will be quite informal in our derivation of the Coalitional Equilibria. Formal analyses for this type of games are not difficult, but can get extremely tedious and do not add much insight.

The Weak Prisoner's Dilemma

To analyze the game of figure 1 (where, remember, P is a dominant strategy), we use a general result proved in Mariotti (1995, proposition 4.1):

Fact 1: at a Coalitional Equilibrium point, no country can be forced below the payoff it can guarantee itself without the collaboration of any other country.

	C	P	C	P
C	6,6,6	4,8,4	4,4,8	1,5,5
P	8,4,4	5,5,1	5,1,5	2,2,2

C P

Figure 1

In this case, each country can guarantee itself a payoff of 2 by choosing to pollute. This means that it can never be the case that a status quo where one country cooperates while the others free-ride can be an equilibrium point: for in that case the cooperating country would be getting only 1. So, (C,P,P), (P,C,P) and (P,P,C) can be excluded from the equilibrium set. Now consider any point such as (P,C,C), where two countries cooperate while the other pollutes. Can there be profitable

deviations from this status quo? Clearly not by country Row, which is getting its maximum payoff. Countries Column and Box can only hope to end up at (C,C,C) if they must successfully deviate (since we have excluded (C,P,P), (P,C,P) and (P,P,C)). The status quo's (C,C,C) and (P,C,C) cannot both be in the equilibrium set, because country Row would deviate from the first to the second. This means that one set of equilibrium points is:

{(P,C,C), (C,P,C), (C,C,P)}.

The status quo (C,C,C) can be profitably improved upon, in one step, by any individual country, as shown above, and (P,P,P) can be improved upon by the coalition of all countries moving to one of the equilibrium points.

The analysis so far leaves open the possibility that (C,C,C) is an equilibrium point and (P,C,C), (C,P,C) and (C,C,P) are not. And, indeed,

{(C,C,C)}

is another (singleton) set of Equilibrium Points, supported by the following coalitional strategies:

- every coalition stays at (C,C,C) when there;
- the grand coalition moves to (C,C,C) whenever at least two countries are polluting: this generates the terminal point (C,C,C) immediately, with gains made by all countries;
- when two countries are cooperating and the other is polluting, the two cooperating countries both deviate to playing P: this will generate the 'interim' status quo (P,P,P) and then the terminal status quo (C,C,C) where the deviating countries are better off than at the initial status quo;
- in all other cases, coalitions stay put.

To summarize:

Fact 2: in the Weak Prisoner's Dilemma, both full cooperation and partial cooperation (between two countries) are possible in equilibrium. Not more than one country pollutes in equilibrium.

It may be at first sight surprising that in a situation with few incentives to cooperate such as this one, cooperation is possible in equilibrium to such a large extent; and even more surprising that, moreover, full or almost full cooperation is *necessary:* lack of cooperation is not an equilibrium status quo. But the intuition for the result, although nontrivial, is not difficult to follow. The basic

point to note is that, in this game, noncooperation is a 'credible threat' when both the other countries behave noncooperatively. Then, nobody believes that situations where a country is the 'odd one out', being free-rided on by the others, can persist. This leaves only two possibilities, depending on what beliefs the countries have regarding the behaviour of coalitions. If it is a common belief that the grand coalition will form at a status quo like (P,P,C) or at (P,P,P) and and treat all countries *symmetrically*, then full cooperation will be the equilibrium status quo[5] . But if is commonly believed that each country will stubbornly refuse the symmetric Pareto optimal status quo, and each of them will vie for its own preferred outcome, then only partial cooperation can be supported. Once a country has 'won the battle' and managed to make its preferred outcome the status quo, any threat on the part of the other countries (to move, say, to (P,P,P)) will be noncredible, since they would be getting less at the threatened outcome than at the status quo.

The main lesson we learn from this first analysis is that when farsightedness is taken into account, the ability of countries to *threat* becomes paramount. The situation is essentially resolved as a balance of threats. Because of this, *an apparent abundance of incentives to free-ride may ultimately support cooperation*, by acting as a credible threat of retaliation against any deviation.

The Strong Prisoner's Dilemma

	C	P	C	P
C	3,3,3	1,5,1	1,1,5	0,4,4
P	5,1,1	4,4,0	4,0,4	2,2,2

[5] Carraro and Botteon (1995) have rightly emphasized the importance of the sharing rule in coalitions. Although they make their case in asymmetric games, the point is valid in symmetric games as well: the internal game within a coalition, although symmetric, may well have nonsymmetric outcomes!

C P

Figure 2

Here, as discussed in the previous section, free-riding is always collectively better for two-country coalitions. For example, suppose that country Box plays C: then by jointly playing (P,P) countries Row and Column are better off than by playing (C,C). A similar observation applies when country Box plays P. As before, P is a dominant strategy.

By using Fact 1, we can see that, similarly to the case of the Weak Prisoner's Dilemma, no country will allow to get stuck in a status quo where the other two countries pollute while it plays cooperatively, which would yield it a payoff of 0, while by playing P it can get at worst 2: so we can exclude (C,P,P), (P,C,P) and (P,P,C) from the equilibrium set. In this case, unlike the previous one, we can also use Fact 1 to eliminate the status quo's where two countries behave cooperatively while the third one pollutes, which yields the cooperating countries only 1. This leaves only full cooperation and full noncooperation as possible equilibrium outcomes. But since one Pareto dominates the other, only (C,C,C) can survive, since once at (P,P,P), all countries will agree on a grand agreement to reduce emissions:

Fact 3: in the Strong Prisoner's Dilemma, only the fully cooperative outcome is possible in equilibrium.

This is the most *prima facie* surprising result of the paper. However, the discussion of the weak version of the game provides an explanation. Here, the incentives to free-ride are so strong that any such threat is credible. At the fully cooperative status quo, any country can credibly threaten any other country that a deviation from full cooperation will be retaliated against. This will lead to a sequence of retaliations, until the worst possible status quo is reached. At this point, all countries will agree to go back to full cooperation: but then, there is no incentive to deviate from full cooperation in the first place!

The Strong Chicken Game

In this game, consider a status quo such as as (P,P,P), with payoffs of 0 for everybody: each country would rather unilaterally reduce emissions and get a payoff of 1. If it did so, then both other countries would prefer to unilaterally reduce emissions, rather than free-ride on the third country.

C P C P

C	4,4,4	3,5,3	3,3,5	1,2,2
P	5,3,3	2,2,1	1,2,1	0,0,0
	C		P	

Figure 3

Fact 1 is once again of help, bacuse it eliminates the full noncooperation status quo (P,P,P). It is easy to see that full cooperation is possible in equilibrium. That is, one set of Equilibrium Points is

{(C,C,C)},

supported by coalitional strategies defined as in the Weak Prisoner's Dilemma Game. All outcomes with two-country cooperation are also possible in equilibrium. That is, another set of Equilibrium Points is

{(P,C,C), (C,P,C), (C,C,P)},

supported by the following coalitional strategies:

- every coalition stays at any equilibrium status quo when there;
- the grand coalition moves to any status quo in the equilibrium set from any status quo where at least two countries pollute: all countries gain by doing so;
- individual countries switch to P at the (C,C,C) status quo;
- in all other cases, coalitions stay put.

Can a status quo like (C,P,P) be an Equilibrium Point? Only if (C,C,C), (P,C,C), (C,P,C), and (C,C,P) are not. Then some coalition must deviate once at one of these status quo's. But clearly in this case there is no terminal status quo where any coalition can gain from moving: the highest payoff a country can hope for at an equilibrium point is 2, which is lower than any payoff at (C,C,C), (P,C,C), (C,P,C), and (C,C,P). This would then violate condition (iii) of the definition of a Coalitional Equilibrium. We have then shown:

Fact 4: in the Strong Chicken Game, both full cooperation and partial cooperation (between two countries) are possible in equilibrium. Not more than one country pollutes in equilibrium.

Thus, we have the same result as in the Weak prisoner's Dilemma, but for very different reasons. In that case, everything hinged on the direct credibility of each *individual* country's threat to switch to P when both other countries were behaving noncooperatively: that is, no matter what the reaction of the other countries was, that country could not be worse off. Here, switching to P in the same situation can potentially be dangerous: the other two countries can threaten to stay at (P,P,P), giving everybody a payoff of zero. But this *counter-threat is not credible*, and this is what makes the original threat credible. Beside being a confirmation of our view of equilibrium agreements as a balance of threats, this example gives us some clear insight into the importance of *indirect deviations* when countries are farsighted.

The Weak Chicken Game

In this game, consider a status quo such as as (P,P,P), with payoffs of 0 for everybody: each country would rather unilaterally reduce emissions and get a payoff of 1. If it does so, then both other countries would prefer to free-ride on the third country, rather than unilaterally reduce emissions. Every country would, as usual, prefer to be a lone free-rider

	C	P	C	P
C	4,4,4	2,5,2	2,2,5	1,3,3
P	5,2,2	3,3,1	1,3,1	0,0,0
	C		P	

Figure 4

Both sets {(C,C,C)} and {(P,C,C), (C,P,C), (C,C,P)} can be Equilibrium Points sets, and can be supported by strategies analogous to the previous case. Now, however, the argument used before to rule out Equilibrium Points such as (P,P,C) is no longer valid: for example, (P,C,C) no longer Pareto dominates (P,P,C). However, now the grand coalition could still move from (P,P,C) to (C,C,C,) with everybody gaining. Therefore, if status quo's like (P,P,C) are Equilibrium Points, then (C,C,C) cannot be one. This can only be the case if status quo's like (P,C,C) are Equilibrium Points. Now, consider the case for country Box to move from (P,P,C) to (P,P,P): whatever terminal point is reached from (P,P,P), country Box is better off compared to (P,P,C), so it will implement the move. But this contradicts (P,P,C) and (P,C,C) both being Equilibrium Points. Thus we have shown:

Fact 5: in the Weak Chicken Game, both full cooperation and partial cooperation (between two countries) are possible in equilibrium. Not more than one country pollutes in equilibrium.

Once again, it may be noted that, although the possible outcomes are the same as in two out of the three previous cases, the 'balance of threats' that supports the possible agreements in each case is quite different.

6. Concluding Remarks

We have shown that, in a reasonably wide range of strategic situations, a high degree of international cooperation in the reduction of polluting emissions can be supported between nonmyopic countries. Notably, in no case could a majority of 'free-riders' be observed in equilibrium, and full cooperation is always a possible equilibrium outcome.

We have argued that the nature of international agreements between non-myopic countries depends on the *balance of credible threats* that can support those agreements. In particular, we have seen that in situations that initially look unpromising for cooperation (Strong Prisoner's Dilemma), the very strength of the individual incentives to free-ride may indeed persuade all countries to cooperate for fear of a chain of retaliations. Clearly, in order for this outcome to be taken seriously, countries must be farsighted enough to understand that the ultimate consequence of free-riding and pursuit of myopic individual interests is the worst possible outcome.

These results should be contrasted with those emerging from models where a more myopic behaviour of countries is assumed. For example, both Carraro and Siniscalco (1993) and Barrett (1994a) have argued that full cooperation is in general not possible (Barrett also argues that if a high degree of cooperation is possible, then the benefits of cooperation compared to noncooperation are small). One could argue that such models are more realistic descriptions of reality, as they do not make high demands on the rationality of the countries involved. Two observations should be made in this respect.

Firstly, even if this argument was correct, it would still be a matter of interest to know that international cooperation for the protection of the environment could be enhanced by promoting more 'farsighted' behaviour on the part of countries. Note that this is different from the trivial claim that, in order to properly appreciate the magnitude of the damages caused to the environment one must take 'a long term view'. The type of farsightedness we are arguing for is *strategic*, not

temporal. It is not obvious a priori that countries which are more farsighted in this sense should be more inclined to cooperate[6].

Secondly, we believe that the sophistication of real decision-makers in understanding the ultimate consequences of certain sequences of actions and counteractions should not be underestimated, especially when decisions are made at the level of countries. It is not difficult to think of situations where cooperation is sustained beyond the point that would be justified by myopic self-interest. For example, one can observe plenty of cartel-like behaviour between competing firms in an industry, one does not observe tariff wars all the time, and so on. As these examples and the environmental negotiations which have taken place so far suggest, it certainly takes time to reach an agreement: our point is merely that large international self-enforcing agreements, which are 'collectively rational', are also compatible with the individual rationality of the countries.

References

d'Aspremont, C.A. and J.J. Gabsewitcz, On the Stability of Collusion, in: G.F Matthewson and J.E. Stiglitz, eds., *New Developments in the Analysis of Market Structure*, MacMiullan Press, New York, 1986.

S. Barrett, Self-Enforcing International Environmental Agreements, *Oxford Economic Papers* **46** (1994a), 878-94.

S. Barrett, The Biodiversity Supergame, *Environmental and Resource Economics* **4** (1994b), 878-94.

S. Barret, Toward a Theory of International Environmental Cooperation, Nota di Lavoro 60.95, Fndazione ENI-Enrico Mattei (1995).

D. Bernheim, B. Peleg and M. Whinston, Coalition-Proof Nash Equilibria I. Concepts, *Journal of Economic Theory* **47** (1987), 1-12.

M. Botteon and C. Carraro, Burden-Sharing and Coalition Stability in Environmental Negotiations with Asymmetric Countries (1995), mimeo.

S. Brams, "Theory of Moves", Cambridge University Press, Cambridge, 1994.

[6] In a different context, Brams (1994) has also studied the implications of 'nonmyopic equilibria' for a wide range of political and social situations. Although our approach is much in the spirit of his, there is the important difference that Brams focuses on individual deviations and does not address in detail the issue of coalition formation.

C. Carraro and D. Siniscalco, Strategies for the International protection of the Environment, *Journal of Public Economics* **52** (1993), 309-28.

P. Chandler and H. Tulkens, Strategically Stable Cost-Sharing in an Economic-Ecological Negotiation Process, 1993, forthcoming in: K.G. Mahler, ed., *International Environmental problems: an Economic Perspective*, Kluwer Academic Publishers, Dordrecht.

M.S.Y Chwe, Farsighted Coalitional Stability, *Journal of Economic Theory* **54** (1994), 299-325.

M.P. Donsimoni, N.P. Economides and H.M. Polemarchakis, Stable Cartels, *International Economic Review* **27** (1986), 317-27.

J. Greenberg, "The Theory of Social Situations: An Alternative Game-Theoretic Approach", Cambridge University Press, Cambridge, 1990.

J. Harsanyi, An Equilibrium-Point Interpretation of Stable Sets and a Proposed Alternative Definition, *Management Science* **20** (1974), 1472-95.

M. Hoel, The Formation of Environmental Coalitions, in: C. Carraro, ed., *Trade, Innovation, Environment*, Kluwer Academic Publishers, Dordrecht, 1994.

M. Mariotti, "A Model of Agreements in Strategic Form Games", University of Manchester working paper in Economics 9528 (previous version available as Fondazione ENI Nota di lavoro)

D. Ray and R. Vohra, Equilibrium Binding Agreements, Brown University, Department of Economics Working Paper 92-8 (1992).

F. Stahler, Some Reflections on Multilateral Environmental Agreements, Nota di lavoro 76.94, Fondazione ENI-Enrico Mattei.

L. Xue, Farsighted Optimistic and Conservative Coalitional Stability, Mc Gill University Working Paper in Economics 9/94 (1994) (1995 mimeo version available under the title "Coalitional Stability under Perfect Foresight").