COOPERATION, STABILITY AND SELF-ENFORCEMENT IN INTERNATIONAL ENVIRONMENTAL AGREEMENTS: A CONCEPTUAL DISCUSSION

by

Parkash CHANDER

National University of Singapore,
Singapore

Henry TULKENS

CORE, Université Catholique de Louvain
Louvain-la-Neuve, Belgium

July 2005
This version: January 2006

CORE DISCUSSION PAPER N° 2006/03

ABSTRACT[1]

In essence, any international environmental agreement (IEA) implies cooperation of a form or another. The paper seeks for logical foundations of this. It first deals with how the need for cooperation derives from the public good aspect of the externalities involved, as well as with where the source of cooperation lies in cooperative game theory. In either case, the quest for efficiency is claimed to be at the root of cooperation.

Next, cooperation is considered from the point of view of stability. After recalling the two competing concepts of stability in use in the IEA literature, new insights on the nature of the gamma core in general are given as well as of the Chander-Tulkens solution within the gamma core. Free riding is also evaluated in relation with the alternative forms of stability under scrutiny.

Finally, it is asked whether with the often mentioned virtue of "self enforcement" any conceptual gain is achieved, different from what is meant by efficiency and stability. A skeptical answer is offered, as a reply to Barrett's (2003) attempt at giving the notion a specific content.

[1] Paper prepared for the David Bradford Memorial Conference "The Design of Climate Policy" held at the Venice Summer Institute organized in San Servolo island by CESifo, Munich, July 22-23, 2005.

Outline

1. **INTRODUCTION**

This paper is not addressed to game theorists -- unless they are interested to learn something about how their products are being used. The paper is addressed, instead, to those economists who make use of game theory notions in the process of analyzing and advising on climate change negotiations.

In 1995 one of us presented a paper[2] entitled "Cooperation vs. Free Riding in International Affairs: Two Approaches" where the main question was: can a grand — worldwide — coalition prevail in climate decisions, or is the problem of such a logical structure that treaties involving only small groups of countries will ever be signed? The answer was in the form of an advocacy of the former thesis.

After 10years, that debate is, to say the least, not closed. Is the present exercise then just a remake? Or has progress been made ? While receiving and selecting papers for this conference, we felt that yes, there is progress, but further clarifications are still called for, let alone for ourselves. This motivates our present contribution, whose structure is clear enough from its title.

Let us introduce some notation for easier reference below. With N the set of all countries of the world, indexed $i = 1,\dots,n$, let p_i denote the amount (flow[3]) of pollutant emissions in country i, let the value of the increasing function (with an upper bound) $g_i(p_i)$, denote the level of country i's GDP, and let the function π_i (.) measure the total cost of damages caused in country i by the aggregate emissions $\Sigma\, p_i$.

In this setting, we shall call a "treaty" a joint choice by several countries of an abatement policy, that is, a level of p_i for each of them, as well as of possible transfers of resources among them. It is also, in general equilibrium terms, a

[2] Published thereafter as TULKENS 1998.

[3] The specific problems raised by stock externalities will not be considered in this discussion, although such are indeed the externalities generated by greenhouse gas emissions. Our immediate excuse is that they are not dealt with either in the literature we consider. More fundamentally, we think the issues at stake need to be clarified first within flow (static) models before being tackled in the dynamic context required by stock externalities.

state — or an "allocation" — of the simple international economy specified above.

In the absence of treaty, we assume that each country chooses the policy that suits it best, given the policies of the other countries, this resulting in a state of the international economy which is also a Nash equilibrium of the noncooperative game that can be associated with the above elements.

Efficiency for a group of countries, be it *N* or any subset *S* of it, is a joint policy of the members of the group that maximizes the group's aggregate welfare *W*. In the case of *N*, this objective reads

$$W_N = \Sigma_{\mathrm{N}} [\, g_i(p_i) — \pi_i\, (\Sigma_{\mathrm{N}} p_j)]$$

where all summation signs refer to indexes running from 1 to *n*. If the group is a subset *S* of *N* however, the maximand is denoted W_S with the first summation in the above expression including only the members of *S*, whereas the second one still bears on all *i*'s. This difference characteristically makes the IEA problem one of externalities.

2. COOPERATION

On the theme of cooperation concerning IEAs, we may distinguish two views. One is economic theoretic, the other is game theoretic.

2.1 The economic rationale for cooperation

The economic view finds its justification in the public good or diffuse characteristic of the externality generated by the emissions that cause climate change. Because the public good is global, that is, world wide, elementary public goods theory (SAMUELSON 1954) teaches us that efficiency (in the Pareto sense) can be reached only if all concerned parties are involved in the process of resource allocation required to master the externality in question. Thus, getting all parties involved — be it by sharing cost, or by revealing preferences, or both, or still by other means — is an essential requirement for efficiency. Economically, the social objective of efficiency entails the necessity of cooperation. Samuelson saw only the State as an appropriate actor for this purpose.

Independently of the public good characteristic just highlighted for the diffuse externalities here under discussion, another economic argument for cooperation in the sense of engaging in bargaining in the presence of externalities is provided by the Coase theorem. IEAs may be viewed as outcomes of voluntary negotiations between generators and recipients of externalities, as described by COASE 1961. This view concludes at an efficient outcome (under appropriate conditions). The Coasean view does not involve the State: quite to the contrary, it asserts that the efficient outcome will emerge from spontaneous negotiations between the parties involved.

2.2 Cooperative games

The game theoretic perspective is the one offered by the theory of cooperative games. This theory flourished in the 60s and 70s prominently within the Jerusalem school of game theory and produced a wealth of "solution concepts" meant to describe the outcome of games (social interactions, in a more recent and better adapted vocabulary) when coalitions of players are the object of analysis. These developments occurred quite independently of public goods and externality theory.

It happens to be difficult, though, to find in this literature arguments explaining and justifying the phenomenon of cooperation. Section 8.1 of MYERSON's 1991 book is entitled "Noncooperative foundations of cooperative game theory" should provide an answer to this query, but the author cautions the reader on how "subtle" the concept is. The attractive idea of giving a noncooperative foundation to cooperative game theory (the "Nash program") hits at a basic difficulty: the multiplicity of equilibria of the noncooperative games that might support cooperative solutions concepts. Criteria that are discussed at length for explaining how selection from among these equilibria might logically occur (focal arbitration of Schelling, institutions, contracts). Somehow, these criteria are one way or another inspired by the notion of efficiency: cooperation finds its *raison d'être* in the efficiency it allows to achieve.

It can be given its root in the outcome of some process of bargaining among the cooperating players[4].

These arguments hardly explain, however, *how* groups are formed, as admitted by the author. At any rate, all game theory textbooks, when they come to their cooperative games chapters (if any), consider that the theory bears on formed groups taken as given, without enquiring on how they got formed, and to which the joint objective of striving for efficiency within the group is attributed .

2.3 Games with externalities

Turning to cooperative games for analyzing IEAs is nevertheless quite justified. The theory has indeed provided very compelling arguments in support of competitive market exchanges, arguments pointing to the so-called strategic stability of market equilibria because they belong to the core of cooperative games associated with markets.

It is thus quite natural to ask whether the core concept, if applied to international economies with externalities, can offer similar properties for Coasean agreements between generators and recipients. This question was raised in the early 70s but it was not clearly dealt with all along the 70sand 80s, probably due to imprecise, unrealistic, or *ad hoc* representations of the externality phenomenon itself. Typically, the core theory was oscillating between results of non existence and problems of non convexities, and the cooperative games under consideration were in fact not really bearing on the multilateral and diffuse form that is commonly used nowadays and recalled in the above introduction.

In the early 90s, clarification on the externality formulation front allowed for the game theoretic apparatus to be called upon and its solution concepts to be applied in this field. The core was one of these concepts so adapted by CHANDER and TULKENS 1995 and 1997, under the name of "γ-core ". Thus,

[4] Pushing this view one step farther, some authors consider cooperative games as normative social science, as opposed to noncooperative games being positive science. This is an oversimplification.

one major concept of cooperative game theory was imported in the IEA literature. One may wonder why, and regret that, other such concepts from the Jerusalem school alluded to above — the bargaining set, the kernel, the nucleolus, the Shapley value or the von Neumann Morgenstern stable sets — have not been similarly more explored in the externalities context[5].

2.4 Coalition formation

At the beginning of the 90s however, there appeared in the early IEA literature of CARRARO and SINISCALCO 1993, 1995 as well as BARRETT 1994 arguments on *formation* of coalitions of countries, inspired from earlier cartel formation models, that some of authors later on called a noncooperative approach.

This theory is built around the idea that a group (coalition) S forms or does not form depending upon whether it passes the following test, called of internal and external stability :

$$S \text{ is internally stable if } \forall i \in S, \quad W_S^i > W_{\{i\}}^i$$

and

$$S \text{ is externally stable if } \forall i \notin S, \quad W_{\{i\}}^i > W_S^i .$$

(This is reminiscent of von Neumann and Morgenstern "stable sets", as expounded by OSBORNE and RUBINSTEIN 1994, p. 279, but is not identical however).

The 1995 paper mentioned above had several questions and remarks on the consistency of the developments made from this definition, especially on the point that it was not specified what happened with the other players.

This seems to have been corrected recently by some authors having recourse to the notion of *games in partition function form*. With this tool, the all

[5] A notable recent exception is to be found in the work of VAN STEENBERGHE 2004 who deals with the nucleolus and the Shapley value of our externality game, using the γ-characteristic function that allowed to define the core.

players set, *N*, is split into non overlapping and collectively exhaustive subsets, which define what is called a *coalition structure*: each partition is such a structure. To each element of a coalition structure, that is, to each coalition belonging to the structure, one can apply the above internal-external (I-E) stability test. If for a given structure the test is passed by all its coalitions, one may call the structure I-E stable. Other expressions are "multi-coalitional equilibrium", or still a "fragmented equilibrium".

This is definitely an improvement in the specification of what happens in the rest of the economy when the I-E stability test is put to any individual. Existence proof of such equilibrium structures for the standard IEA model have not been provided yet, to the best of our knowledge.

However, EYCKMANS and FINUS 2005 have explored the issue by means of numerical simulations with the specific CWS integrated assessment model[6]. They take all conceivable partitions of the set of six regions of the world that the model treats as "countries", they compute a multi-coalitional equilibrium for each structure (that is, a Nash equilibrium between the coalitions in the structure) and then they check which ones of these coalitions pass the I-E stability test. Similarly, BUCHNER and CARRARO 2005 examine with simulations on the FEEM-RICE model[7] the I-E stability of conceivable coalition structures.

In a multi coalitional equilibrium, each coalition *S* is assumed to achieve efficiency within itself along its members. While each coalition thus strives for internal efficiency, one could ask why is this quest limited to the members of *S*? Why do coalitions not strive for external efficiency, that is, contact other coalitions and adopt mutually beneficial and still more efficient strategies?

If it can be shown that the resulting merged coalition is not I-E stable, then the above argument applies for asserting that the merge will not takeplace. But if it happened to be I-E stable, then why should the merged coalition not form? We would have multiple equilibria, though. Any argument that would give precedence to the equilibrium with the merge over the one without it would be

[6] The CWS model was introduced by EYCKMANS and TULKENS 2003.

[7] The FEEM-RICE model was introduced in BUONANNO, CARRARO and GALEOTTI, 2003.

based on efficiency domination of the former. So, we are driven back to a reasoning on coalition formation essentially led by efficiency considerations[8].

Finally, it is probably clear that, no more than what cooperative game theory has to offer, I-E stability criteria do not teach much on the process of *how* stable coalitions are formed.

2.5 An axiomatic approach

On the theme of coalition formation in games with externalities, MASKIN 2003 has brought about a contribution based on other arguments. He (courageously) tackled the sequential process of discussions between players on whether or not they will act jointly. His analysis is grounded in an explicit axiomatics that bears (in part) on communication between the players. One of these axioms specifies that at some point any player is allowed to break communication lines between himself and (some of) the other players. On that basis, the conclusion is derived that the grand coalition will not form.

But doesn't that axiom contain the conclusion? Leaving aside this objection, it is to be praised that Maskin introduces the important factor of communication between the players as a determinant of cooperation. He acknowledges[9], however, that without this axiom, the grand coalition would form in the public good game of his paper (which is very close to the environmental model we deal with in IEA literature), and thus efficiency would prevail.

* * *

To summarize on the complementary themes of cooperation and coalition formation, we are left with the following: (1) the Nash program, which is incomplete; (2) the I-E stability argument that only explains the non formation of some coalitions (among which N); (3) the communication breakdown argument axiomatically called upon by Maskin to justify the non formation of

[8] Repeating this reasoning on further mergers might well end up with N as the only coalition!

[9] Private communication, after the Coalition Theory Network meeting in Paris, January 2005, where the paper was presented and discussed.

the grand coalition in games with externalities. Conceptual progress is there. General results are scarce, however.

3. STABILITY

Stability may be viewed from a positive point of view or, alternatively, from a normative viewpoint. The positive view considers stability as a logical property of some situations, whereas in the normative perspective, stability is taken as a desirable virtue, that is, something to be sought after.

The coalition formation literature, whose recent development was just evoked, belongs to the positive view.

A point to be clarified at the outset is the following: stability of what? Of coalitions or of allocations? In many of its formulations in the IEA literature the focus has been more on coalitions than on allocations. This is due, we surmise, to the systematic use of the symmetric players assumption[10] by the authors, which leads them to state their results in terms of a single number, namely the number of signatories, with no mention whatsoever of the ensuing state of the economy or of the environment. The oversimplification of the economic model has made one loose the object of interest. When it comes to derive policy (*i.e.* normative) statements it is not sure that such a slender basis provides a strong enough justification.

3.1 Alternative stability concepts: internal-external stability vs coalitional stability.

This being said, the main issue is: what kind of stability is at stake? In slightly more sophisticated terms, which concept of stability is being used? Internal-external stability which inspired the coalition formation literature is not the only stability notion offered by game theory; it is even a rather recent one. The expression of "strategic stability" has been known for decades to designate some of the solutions proposed by cooperative game theory.

[10] As well as the rudimentary description of environmental phenomena; but this is acceptable because no model will ever describe reality entirely.

3.2 Some further insights on the nature of γ-core solutions for the IEA game

Let us recall that for an IEA game, the γ-core consists of allocations (treaties[11]) that specify:

(i) a profile of emissions for each party, that are Pareto efficient at the world level;
(ii) transfers (some positive, some negative) amongst the parties[12], that cover the total cost of abatement for each of them, and are financed by contributions from each based on their relative (marginal) environmental damage costs[13].

The core property of an allocation is that if any individual or group of parties considers deviating from it, the best it can do is less attractive than what it gets in the allocation.

As we are dealing with a game with externalities, it is essential to be clear on one point: when an individual or a group of parties considers deviating, what should they think the *other* parties will do: punish them (for instance by polluting a lot), or let them go (and possibly adjusting their abatement policy to the absence of the defector)? From the alternative assumptions that may be made in this respect, the γ-core theory is based on assuming that in the case of defection by some subset *S* of parties, the other parties will abandon any form of cooperation and act to the best of their interest as singletons in the face of *S*.

[11] Incidentally, there is no a priori reason to believe that there is only one allocation (or treaty) that could belong to the core of the IEA game. In other words, the core is not a unique point solution concept, neither in general no in the particular case of IEAs.

[12] It is important always to recall (from CHANDER, TULKENS, VAN YPERSELE and WILLEMS 2003, section 5) that the same allocation can be achieved with the transfers being substituted by initial allowances of tradable emission permits, provided that the amounts of these allowances be such that the resulting competitive equilibrium on the permits market induces the γ-core allocation just defined. This point is of major importance when discussing the connection between the theories presently examined and actual treaties, such as the Kyoto Protocol for instance, where there are no explicit transfers specified. But the treaty's allowances play their role. For a further and thorough exploration of this substitution, see VAN STEENBERGHE 2004.

[13] In the notation of this paper, the CHANDER-TULKENS 1997 formula for these transfers T_i (>0 if received, <0 if paid) reads:

$$T_i = -(g_i(p_i^*) - g_i(p_i^-)) + \frac{\pi_i^{'*}}{\sum_{j\in N}\pi_j^{'*}}\left(\sum_{j\in N} g_j(p_j^*) - \sum_{j\in N} g_j(p_i^-)\right).$$

The outcome of this joint strategy being dubbed a "partial agreement Nash equilibrium with respect to *S* (PANE w.r.t. *S*).

When this form of cooperative game with externalities was presented, so much of the discussion was focused on this assumption that many people have lost sight of what the core allocation itself is. Therefore, we seize this opportunity to present, with the help of some diagrams, some apparently unnoticed properties of the solution we advocate.

In the standard multilateral externality model everybody uses nowadays to deal with IEAs, each player (country) *i* is at the same time a polluter and a pollutee[14]. In an effort to disentangle which roles each one of these two functions plays in the determination of the solution, let us consider successively the following elementary, actually unilateral, forms of the model, successively with two, and then three parties.

In the first instance, we have just one polluting country, indexed *r,* which is not polluted, and one polluted country indexed *e*, which is not a polluter. Think of a simple upstream-downstream river pollution situation. In the Edgeworth box - type of diagram appearing in Figure 1, that one of us introduced in 1974[15], the core of this two agents economy consists of all points on the segment A-B; it is also the locus of all allocations that may be reached by Coasean bargaining. Among these points, the Chander-Tulkens (CT) solution is seen in this case to be point A. Indeed, it is formulated as *(i)* a Pareto optimum; and *(ii)* with a transfer KL such that the polluter is compensated by the pollutee for his abatement cost, and nothing more. The segment A-B is reminiscent of the gains from trade in exchange interpretations of the Edgeworth box. The figure here illustrates vividly how the externality is somehow an object of exchange in this setting.

Let us now enlarge this economy with one more pollutee, with the two pollutees indexed e_1 and e_2 respectively. Figure 2 reproduces Figure 1 except that the second pollutee's indifference curve has been added horizontally to the

[14] This is why it is called multilateral.

[15] See TULKENS and SCHOUMAKER 1975, pp. 247 ff., probably independently redrawn in VARIAN 1990, pp. 539 and 542. This diagram can be deduced from the simplified version of the IEA model sketched out below the figure. Full details are given in the paper cited.

first pollutee's curve, so that at each point on the resulting curve MD the slope of the tangent measures the *sum* of the marginal rates of substitution between environmental pollution and the numeraire *y*.

Here, the Pareto efficient level of emission p^* is measured as O_rC, the core of the economy is the line DE and the CT solution is the core point D. At that allocation, the polluter is compensated just for the cost of his abatement, and no more. The bargaining gain (DE) is entirely appropriated by the pollutees as if, in process, they had acted as a single party[16].

What this illustration makes clear is threefold:

(i) it shows what the bargaining gain is made of with several pollutees, that is, with several recipients of the externality;
(ii) it identifies this gain with the core of the game;
(iii) it shows the particular nature of the CT solution within the core: it deprives the polluter of any pure bargaining gain; but all of his abatement cost is covered.

Note that other core points are conceivable and reachable, all more beneficial to the polluter. If they were reached, it would be out of pure bargaining power between *r* and the set of *e*'s; free riding on the part of *r* would play no role in this respect, for the simple reason that for *r*, there is nothing to free-ride about!

The relative positions of the players *qua* polluters *vs. qua* pollutees in the IEA game, in its core and at the CT solution are further highlighted with diagrams such as those appearing in Figures 3 and 4. They show how large the γ-core can be as well as the strongly pollutees-favoring character of the CT solution. All the ecological surplus goes to them. Yet, this is specific to the CT solution: other solutions in the core may benefit the polluters, as suggested by point R on Figure 4 where the two polluters succeed to reap part of the bargaining gain (they would reap it all if R was located on the c-d line).

[16] Unfortunately, the picture does not lend itself to show easily how, at the CT solution, the coverage of the polluter's abatement cost KL is shared between the two pollutees e_1 and e_2, and thus how the Coasean gain is shared amongst them.

3.3 The rationale for a game with a particular coalition structure

In extending the core concept to games of the IEA type, the key instrument is an appropriate specification of the characteristic function of the game. This function summarizes the ability of groups of players to object to, to block or to improve upon possible solutions (or, for that matter, to free ride on a solution). With externalities taking place between individual players as well as between groups, an essential element is that the value of function for a given group also depends on the actions of players not in the group.

The assumption made in this respect when formulating the γ-characteristic function in terms of PANE w.r.t.*S*es (specifying for each *S* that all non-members of *S* act as singletons) amounts to define the function on a particular partition of the set *N*. As already mentioned above, a particular partition of the set of players has been called[17] a *coalition structure.* We are thus dealing with a cooperative game with a coalition structure.

The question may be raised why limiting oneself to just that structure and not considering the family of all conceivable partitions that may contain *S* ? This would transform our IEA game, which thus far is one in characteristic function form, into a game in partition function form.

We have two reasons for not exploring this extension. One is the paucity of results, and even of treatment, of such games in the literature[18], that we could transpose to our IEA model. The other reason is one of substance: Not all coalition structures can be considered as rational ones, and equally likely to emerge. Some coalition structures are of little or no interest, such as $\left[N\setminus\{i\};\{i\}\right]$ for instance, when *i* is an unimportant country, both as emitter and as recipient of the externality. Also, the structure $\left[\{i\};N\setminus\{i\}\right]$, which is the one of optimistic individual free riding, is irrational for $N\setminus\{i\}$ as we shall argue below[19] in 3.4.2.

[17] By AUMANN and DRÈZE 1974.

[18] THRALL and LUCAS 1963 is an early source, limited to $n \leq 3$.

[19] These arguments also apply, and even more forcefully, to more asymmetric models where some players are either polluters only or pollutees only.

Thus, instead of considering mechanically all conceivable structures, taking into account the rationality of the collective behavior of the non-members of *S* leads one to rather select a well justified structure.

3.4 Free riding and stability

3.4.1 Two forms of free riding

Originally, the expression of free riding was used by SAMUELSON 1954 to describe the behavior of economic agents who conceal their preferences with respect to a public good[20] vis-à-vis a single producer — this producer being necessarily the State because of the impossibility of selling the good. On the public good production side, there was no question of leaving or joining coalitions, neither in that paper, nor in the following public goods literature — until the international environmental problems were taken up in the late sixties and early seventies.

Here, the necessarily voluntary character of the provision of the public good, that is, abatement of the environmental externality, together with the fact that the externality is multilateral, shifted the attention from the issue of individual consumers revealing preferences to a planning authority[21] to the problem of having several States participate or not in international voluntary agreements on a global externality. The expression of free riding reappeared here not as a preference revelation problem but instead as a way to behave in the face of such agreements[22].

[20] Correct revelation is necessary to be able to check whether efficiency is obtained; but if that information is used to determine the individuals' contributions to the financing of the public good, they will be tempted to understate their preferences and production will be suboptimal, whereas if no connection is made between what they reveal and what they have to pay, they will overstate their preferences and production will be larger than optimal.

[21] While Samuelson himself in 1954 wrote that only some smart game theorist could master the preference revelation problem raised by the free riding behavior he had identified, the challenge was successfully taken up by game theoretically minded economists fifteen years later in a series of papers written in the context of decentralized planning procedures, starting with DREZE and DE LA VALLEE POUSSIN 1971, pursued by ROBERTS 1979, HENRY 1979, GROVES and LEDYARD 1973, TIDEMAN 1974, and culminating with CHAMPSAUR and LAROQUE 1981. This literature may be seen as one of the main sources of the mechanism design stream of thought that developed subsequently.

[22] A third notion of free riding has been put forward by FINUS 2001 (p. ...), namely the behavior that consists in signing an agreement and then not complying with it. We are not sure this wording is appropriate. Non compliance is breaching an agreement that was

There are thus two forms of free riding, that we propose to call "preference revelation (PR) free riding" and "non participatory (NP) free riding", respectively. Notice that the two forms are not mutually exclusive, but we are not aware of any work that treats them simultaneously. We shall consider here essentially the latter, with occasional allusions to the former.

3.4.2 Free riding and I-E vs. γ-core stability

Free riding is a special form of instability of a group. Depending upon the stability concept one uses, what free riding designates will vary. Thus, if a core allocation is declared not to be I-E stable, that is, lets some *i* leave the grand coalition, it is because the non stability statement implicitly rests on the assumption that if *i* leaves N, $N\setminus\{i\}$ remains as a coalition, possibly re-optimizing its strategy, and tolerates *i*'s free riding, that is, it tolerates the global inefficiency induced by *i*'s defection. Now, this my be rational on the part of $N\setminus\{i\}$ only if the role of $\{i\}$ in the economic problem is small. But it is irrational whenever that role is important, either because *i* has a lot to abate [$(p^* - p^-)$ is large], or *i* is an important contributor to cost coverage (π_i is large), or *a fortiori* when both factors are at play: the severe efficiency loss in these three cases is such that it is more rational for the members of $N\setminus\{i\}$ to threaten to break apart into singletons, which would drive *i* back to the Nash equilibrium and thereby induce him not to defect in the first place[23].

Thus, the assumption behind internal-external stability of an (efficient) coalition N is that $N\setminus\{i\}$ *tolerates* free riding[24] and even adjusts to it. By contrast, the assumption behind the γ-core stability is that $N\setminus\{i\}$ *counters* free riding by reacting, not in an extreme punishing way (as with the α-characteristic function) but rather in a way just sufficient for making the free rider realize that he might be put in a situation in which he would definitely prefer what he gets with the grand coalition after all.

adhered to. In the two senses described above, free riding is either not signing the agreement, or being part of it under favourable conditions because of an information bias.

[23] This farsightedness argument was first introduced in CHANDER 2003.

[24] EYCKMANS and FINUS 2004 do even reward it, calling it "ideal". Note that to offer that compensation, it is needed to now the preferences of the free rider. Is there any reason to believe that he will reveal truthfully, while bargaining on a possible defection?

The strength of the γ-core concept in dealing with (actually, solving) the free rider problem thus lies in the farsighted rationality of the threat it assumes. The weakness of I-E stability is, instead, in that it legimates free riding.

3.4.3 Free riding and the particular CT core solution.

While the CT solution has all the core stability properties just outlined, in addition it allows one to see the effect of a party *i* joining but incorrectly revealing preferences, through the $\frac{\pi_i^{'*}}{\sum_{j\in N}\pi_j^{'*}}$ coefficients in the transfers formula. Understating π_I implies a lesser contribution of *i* to the coverage of the aggregate abatement cost. But that lower value of π_i also induces a less than optimal level of aggregate abatement since the optimality criterion is based on the sum of the π_i's. Thus, the CT solution to the IEA game is vulnerable to preference revelation free riding, at least away from the optimum[25].

To conclude, we are back again to the motivations for seeking stability: from a normative point of view, the reason for avoiding free riding is essentially that it prevents efficiency.

4. **SELF-ENFORCEMENT**

"Self-enforcement" is an intuitively quite attractive expression, when dealing with international agreements. It evokes the absence of an external authority, which is at the root of the problems raised by this type of agreements. It also contains an implicit reference to incentives. After its introduction by BARRETT 1994, the appearance of a book (BARRETT 2003) entirely devoted to that idea has positioned the author as its most articulate advocate.

[25] When the optimum is reached, there is an argument due to DRÈZE and DE LA VALLEE POUSSIN 1971 (section 3) establishing that it is a Nash equilibrium of a preference revelation game that all parties to reveal correctly their preferences. Away from the optimum, this is not the case anymore, but the bias in misrepresentation can be identified (see ROBERTS 1979).

For cooperative game-minded theorists like us, there is a bit of mystery with it: it is difficult to find in the standard literature a commonly received definition of self-enforcement. It does usually not appear in the index of game theory textbooks, and when it does (*e.g.* in MYERSON 1991), it is only to refer to a property of occasional interest. More importantly, in what sense is self-enforcement more than efficiency, or more than core or I-E stability? Is it an additional concept that we should add to our tool box for IEA analysis?

We feel that while the answer to the last question is definitely yes, the answers to the previous question are difficult to make precise.

Self-enforcement is a property of a treaty that "must satisfy three conditions: individual rationality, collective rationality and fairness" (BARRETT 2003, pp. xiii-xiv). Apart from the first one, which is used in its standard sense, these terms are given a special meaning. For instance, collective rationality is redefined successively in chapters 7 and 11 as a property of a treaty implying not only efficiency for the group under consideration, but in addition free riding deterrence (p. 213)[26], which is given on p.294 two possible forms (strong and weak collective rationality, respectively). A formal definition is offered in section 11.4, unfortunately with a model of identical players which is hardly convincing. On the other hand, fairness is not formally dealt with, but presented as a requirement that the treaty "be perceived by the parties as being legitimate" (p.xiv).

While potential readers, fond of precise definitions and rigorous developments of sufficiently rich and realistic models, are likely to be sometimes disappointed, the book offers nevertheless a remarkable intellectual challenge to theorists dealing with IEAs.

The one we like to highlight here is the theme of chapter 11, which describes a possible trade-off between the breadth of international cooperation (in terms of the number participants in a treaty) and its depth (in terms of the

[26] We have responded above to Barrett's criticism of the γ-core whereby he introduces his collective rationality concept: we claim that the threat he considers as non credible is in fact a farsighted rational one, as proved in CHANDER 2003.

size of the actions agreed upon by the parties): is a “broad but shallow” treaty better than a “narrow but deep” one?

A shallow treaty would be one that does not achieve full efficiency among the participating countries, *e.g.* by abating less than optimal; this would be the price, so to speak, for having it signed by many countries. The outcome is called by Barrett a “consensus treaty”, asserted elsewhere to be self-enforcing. That this is better than the opposite (deep and narrow) is claimed to be established (p.302) by means of an ingenious symmetric countries model. But we have already voiced the opinion that such a basis is itself quite shallow for transforming into scientific truth this conjecture.

Yet, the trade-off brought to light remains an important intellectual challenge: while it surely deserves scrutiny by means of better adapted, and therefore more elaborate game theoretic tools, it illustrates once more that before proving an idea to be true, it must be generated. This is a major merit of many ideas in Scott Barrett’s book.

Let me conclude with a perhaps timely question: would the David Bradford scheme presented at this conference be self-enforcing?

5. CONCLUSION

Neither stability nor cooperation are desirable *per se*. Both are there to achieve efficiency, because the welfare of people derives primarily from allocations, not from their stability or from cooperation. The virtues of Barrett's self-enforcement eventually point in the same direction, admitting that otherwise, no treaty would be signed at all.

At a less general level, the analysis revealed that there is much to gain in understanding if one distinguishes more explicitly between the involvement of countries as polluters from their involvement as pollutees.

In fact, this is already done, to some extent, within the Kyoto Protocol: the motivations behind the *aggregate* quotas that have been negotiated are essentially those of the pollutees: they result from their preferences; by contrast, the working of flexible mechanisms is of concern essentially for the polluters. What is less clear is how the bargaining gain turns out to be shared among these parties.

REFERENCES

AUMANN, R. and DRÈZE, J. 1974, "Cooperative Games with Coalition Structures", *International Journal of Game Theory* 3, 217-238.

BARRETT, S. 1994, "Self enforcing international environmental agreements", *Oxford Economic Papers* 46, 878-894.

BARRETT, S. 2003, *Environment and Statecraft: The strategy of Environmental Treaty-Making,* Oxford University Press, Oxford.

BUCHNER, B. and CARRARO, C. 2005, "regional and sub-global climate blocs: a game theoretic perspective on bottoçm-up climate regimes", FEEM, mimeo (presented at this Conference).

BUONANNO, P., CARRARO, C., GALEOTTI, M. 2003, "Endogenous induced technical change and the costs of Kyoto", *Resource and Energy Economics* 25 (2003) 11–34.

CARRARO, C. and SINISCALCO, D. 1993, "Strategies for the international protection of the environment", *Journal of Public Economics* 52, 309-328.

CARRARO, C. and SINISCALCO, D. 1995, "International coordination of environmental policies and stability of global environmental agreements", chapter 13 in Bovenberg, L. and Cnossen, S. (eds*), Public economics and the environment in an imperfect world,* Kluwer Academic Publishers, Boston, London, Dordrecht.

CHAMPSAUR and LAROQUE 1981, *Econometrica*

CHANDER P. 2003, "The _-Core and Coalition Formation", *CORE Discussion Paper* 2003/46; revised version: January 2005.

CHANDER, P. and TULKENS, H. 1995, "A core-theoretic solution for the design of cooperative agreements on transfrontier pollution", *International Tax and Public Finance* 2 (2), 279-294.

CHANDER, P. and TULKENS, H. 1997 "The Core Of An Economy With Multilateral Environmental Externalities", *International Journal of Game Theory* 26, 379-401.

CHANDER, P., TULKENS, H., VAN YPERSELE, J-P. and WILLEMS, S. 2002, "The Kyoto Protocol: An Economic and Game Theoretic Interpretation", chapter 6 (pp.98-117) in Kriström, B., Dasgupta P. and Löfgren K.-G. (eds), *Economic Theory for the Environment : Essays in Honor of Karl-Göran Mäler*, Edward Elgar, Cheltenham.

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

DRÈZE, J. et de la VALLÉE POUSSIN, D. 1971, "A Tâtonnement Process for Public Goods", *Review of Economic Studies* 38, 133-150.

EYCKMANS, J. and FINUS, M. 2005, "Measures to enhance the success of global climate treaties", mimeo.

EYCKMANS and FINUS 2004, "An almost ideal sharing scheme for coalition games with externalities", mimeo (September)

EYCKMANS, J. and TULKENS, H. 2003, "Simulating coalitionally stable burden sharing agreements for the climate change problem", *Resource and Energy Economics* 25, 299-327.

FINUS, M. 2001, *Game Theory and international Environmental Cooperation,* Edward Elgar, Cheltenham.

MASKIN, E. 2003, "Bargaining, Coalitions and Externalities", presidential address to the Econometric Society European Meeting, Stockholm, mimeo (August).

MYERSON, R. 1991, *Game Theory: Analysis of Conflict*, Harvard university Press, Cambridge, Mass.

OSBORNE and RUBINSTEIN 1994, *A Course in Game Theory*, The MIT Press, Cambridge, Mass.

ROBERTS, D.J. (1979), "Incentives in Planning Procedures for the Provision of Public Goods", *Review of Economic Studies* XLVI (2), 283-292.

SAMUELSON P.A. 1954, "The Pure Theory of Public Expenditure", *Review of Economics and Statistics* 36, 387-389.

THRALL and LUCAS 1963, "n-Person Games in Partition Function Form", *Naval Research Logistics Quarterly* 10, 281-298.

TULKENS, H. and SCHOUMAKER, F. 1975, "Stability Analysis of an Effluent Charge and the 'Polluters Pay' Principle", *Journal of Public Economics* 4, 245-269.

TULKENS, H. 1998, "Cooperation vs. free riding in international environmental affairs: two approaches", chapter 2 (pp. 330-44) in N. Hanley and H. Folmer (eds), *Game Theory and the Environment,* Elgar, Cheltenham. Translation in French published as chapter 2 (pp.47-72) in G. Rotillon, ed. *Régulation environnementale: jeux, coalitions, contrats*, Economica, Paris 2002.

VAN STEENBERGHE, V. 2004, "Core-stable and equitable allocations of greenhouse gas emission permits", *CORE Discussion Paper* 2004/75.

VARIAN, H.R. 1990, *Intermediate Microeconomics: A Modern Approach*, second edition, W.W.Norton & Company, New York and London.

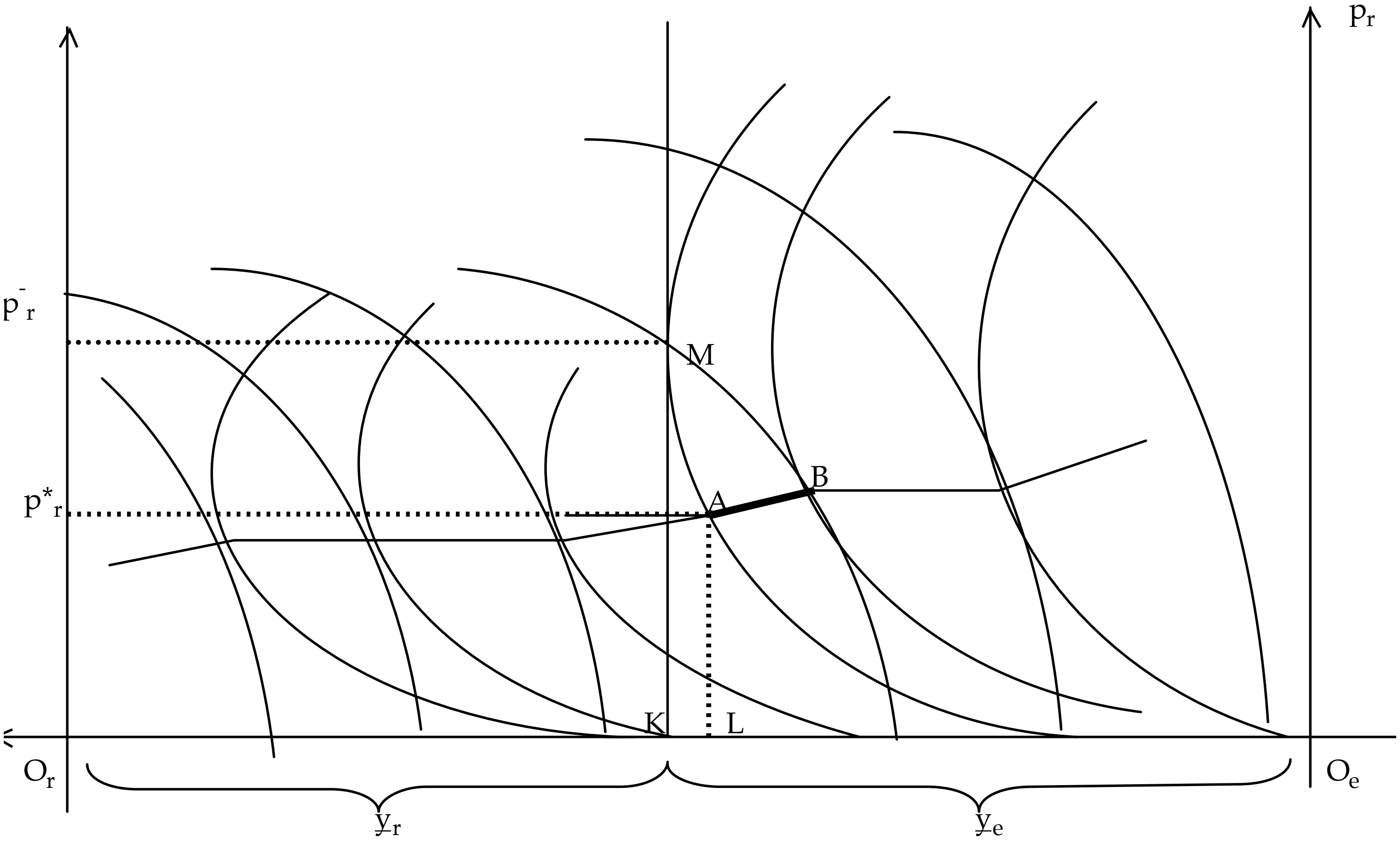


Figure 1 - A one polluter (*r*) **- one pollutee** (*e*) **economy**
Source: TULKENS and SCHOUMAKER 1975

Polluter : $u_r(p_r, y_r)$, $y_r \le \underline{y}_r$ (>0 : initial endowment)

$$with\ \ \partial u_r / \partial p_r \ge 0,\ \partial u_r / \partial y_r > 0,$$

Pollutee : $u_e(p_r, y_e)$, $y_e \le \underline{y}_e$ (>0 : initial endowment)

$$with\ \ \partial u_e / \partial p_r < 0,\ \partial u_e / \partial y_e > 0.$$

:

M : *Nash equilibrium*

A - B : *core* (with respect to M)

A : *CT solution* (where r receives from e a transfer KL)

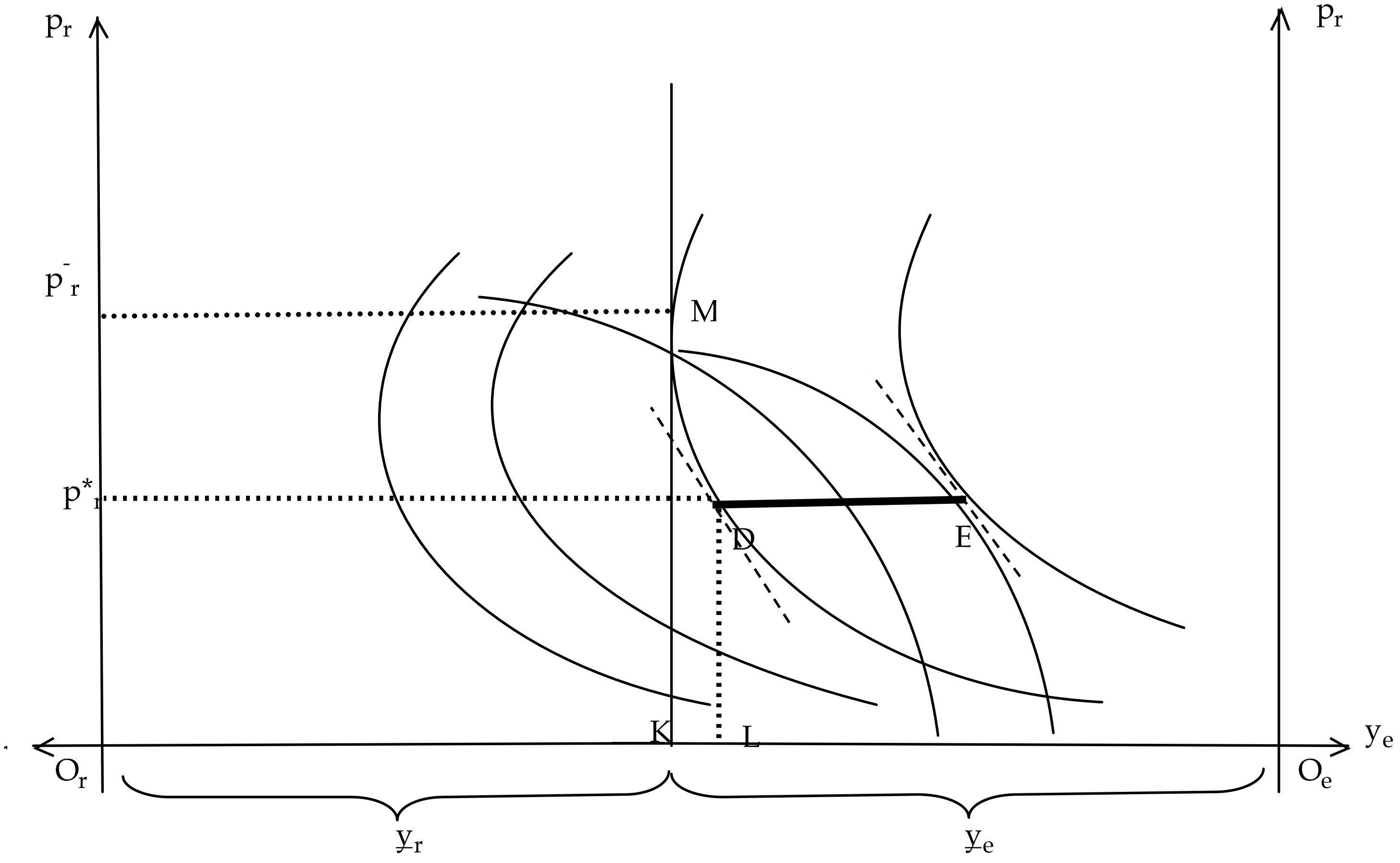


Figure 2 : One polluter (r) and two pollutees (e_1,e_2)

:

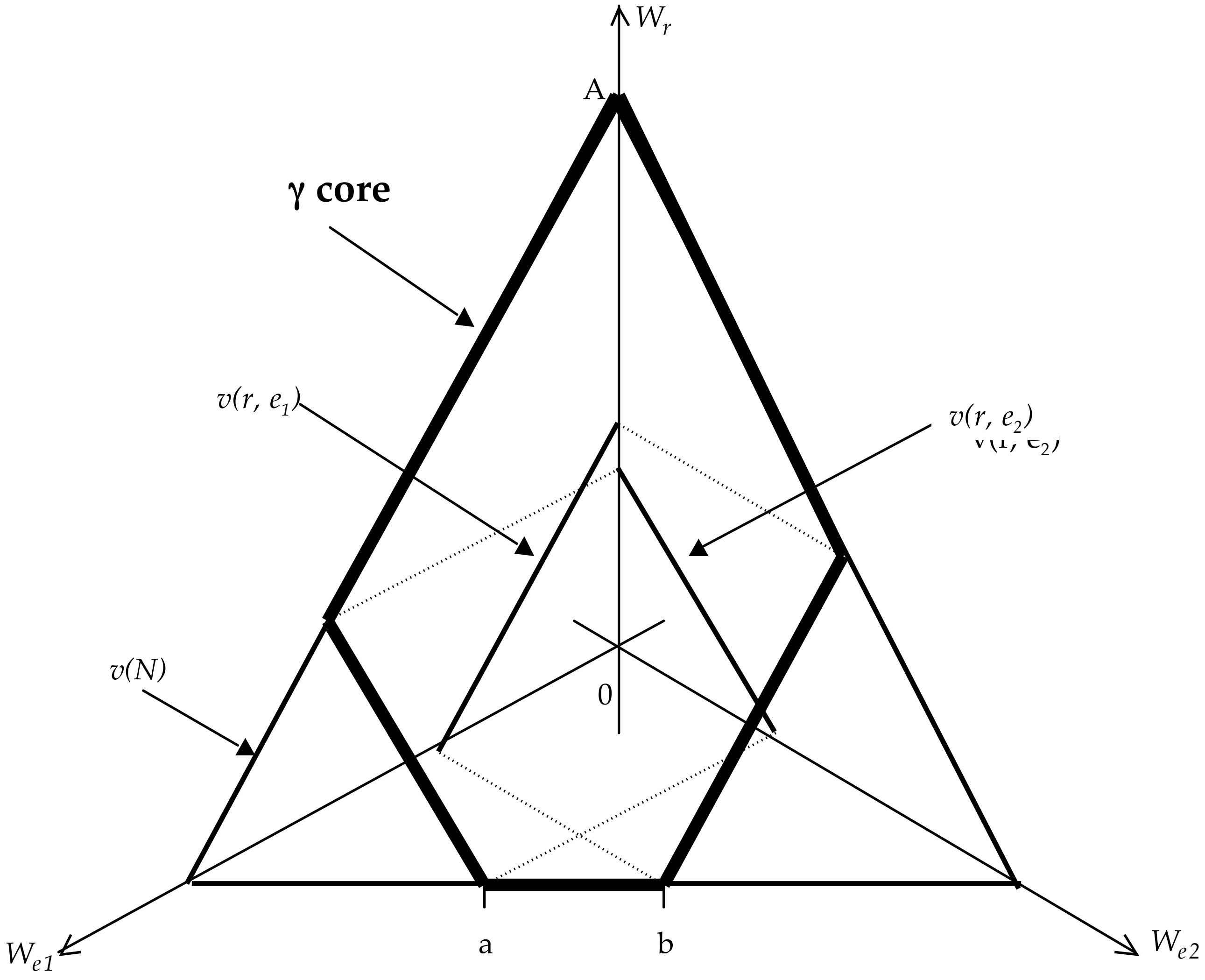
p_r^* is optimal because at for that value of p_r

slope at D (= MRS_r) = slope at E (= $\sum_i MRS_{e_i}$)

M : *Nash equilibrium*

D - E : *core* allocations

D : *CT solution* (r receives from e_1 and e_2 an aggragate transfer KL;
the respective shares of payment by e_1 and e_2 are not shown)



The game is defined by $N = \{r, e_1, e_2\}$ and the characteristic function $v(\,.\,)$.

Note that $v(e_1, e_2) = 0$.

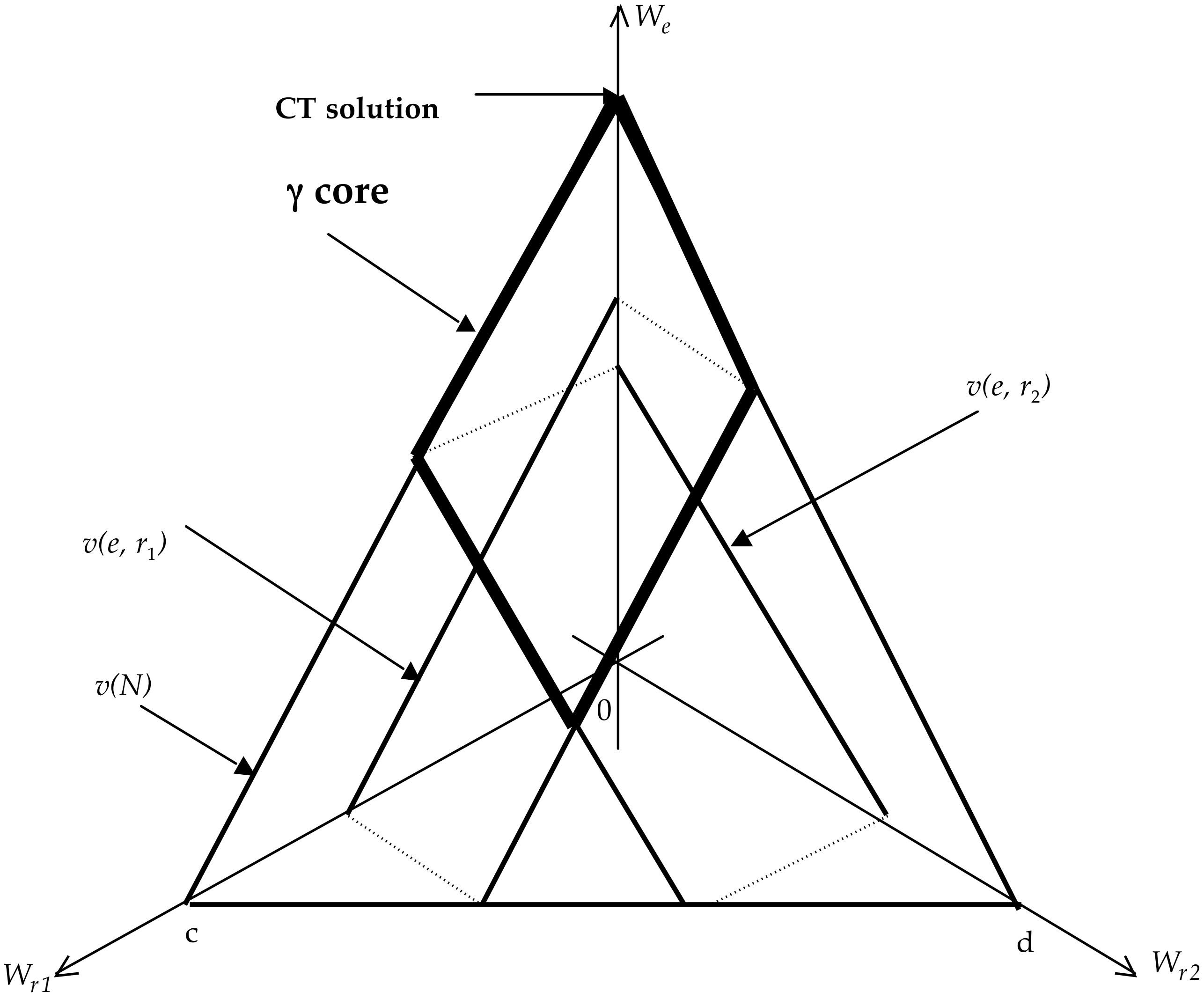
The origin is the welfare levels of players at the Nash equilibrium

The CT solution is one point along the segment [a, b].

: There, all the bargaining gain accrues to the pollutees.

That point A belongs to the core illustrates that the (single) polluter can reap all of the bargaining gain

Figure 3 : The γ core in payoffs space for any one polluter (r) and two pollutees (e_1,e_2) game



The game is defined by $N = \{e, r_1, r_2\}$ and the characteristic function $v(\,.\,)$.
Note that $v(r_1, r_2) = 0$.
The origin is the welfare levels of players at the Nash equilibrium

The CT solution is one point on the W_e axis.
There, all the bargaining gain (or ecological surplus) accrues
to the (single) pollutee.

Other core solutions give some of the gain to the polluters
but in this example, a solution where the two polluters would reap
all of the bargaining gain (*i.e.* a point along $[c, d]$) does not belong to the core.

In general, the stronger (weaker) the coalitions between the polluter and one pollutee,
the less (more) the pair of olluters can obtain from the bargaining gain.

**Figure 4 : The γ core in payoffs space
for any one pollutee (e) and two polluters (r_1,r_2) game**