3. Stability and design of international environmental agreements: the case of transboundary pollution

Michael Finus*

1. INTRODUCTION

Concern about international environmental problems[1] has grown immensely over the last four decades. This led to the signature of several international environmental agreements (IEAs), as for instance the Helsinki and Oslo Protocols on the reduction of sulphur signed in 1985 and 1994, respectively, the Montreal Protocol on the reduction of chlorofluorocarbons (CFCs) that deplete the ozone layer signed in 1987 and the Kyoto Protocol on the reduction of greenhouse gases causing global warming signed in 1997.[2,3] This concern is also reflected in numerous papers on the economics of international environmental problems. In this chapter I survey the game-theoretical literature on coalitions analysing the formation and stability of IEAs. The fundamental result motivating all analyses is that as long as environmental problems are not of purely local nature, global welfare can be raised through cooperation. The fundamental assumption of all models is that there is no international agency that can establish binding agreements.[4] Consequently, cooperation faces three fundamental constraints (see section 2.1 for details): (1) IEAs have to be profitable for all potential participants; (2) the parties must agree on the particular design of an IEA by consensus; and (3) the treaty must be enforced by the parties themselves. The main feature according to which models can be structured is the type of free-riding they capture. Two types of free-riding can be distinguished. The first type implies that a country is either not a member of an IEA or is a member of an agreement

* I have benefited from grant No. 2136001059748 (WIMEK), University of Wageningen, The Netherlands, from discussions with Carlo Carraro, Alfred Endres, Michaela Hedyas, Pierre van Mouche, Carsten Mössig, Bianca Rundshagen and Ekko van Ierland and would like to acknowledge research assistance by Frank Brockmeier and Eva Schreiner. I also would like to thank the editors of this volume and two anonymous referees for many constructive comments. I am particularly indebted to Henk Folmer, who provided me with many suggestions that substantially improved the exposition of the material.

that contributes less to the improvement of environmental quality than members of other agreements. This type of free-riding is captured by models that I call 'membership models', where the first aspect is modelled in 'traditional' single-coalition games and the second aspect in 'new' multiple-coalition games. The second type of free-riding implies that a country is a member of an IEA but does not comply with the terms of the agreement. This type of free-riding is analysed in models that I call 'compliance models'. Since the bulk of the theoretical literature that I review is related to pollution problems, I restrict myself, by and large, to this variety, though most of the qualitative results also apply to other problems, as for instance the depletion of fish stocks and the deforestation of tropical rain forests.

In what follows, I present empirical evidence on the problems of cooperation (section 2.1) and on important issues of treaty design (section 2.2). I introduce a basic framework for the analysis of international pollution problems (section 2.3) and give an overview of the features of possible extensions (section 2.4). Subsequently, I provide a summary of important results obtained with membership models (section 3) and with compliance models (section 4), critically review the models with respect to their theoretical consistency, their ability to contribute to the understanding of real-world phenomena and the extent to which they capture the three fundamental constraints of cooperation. Finally, I point out topics for future research in section 5.

2. BACKGROUND INFORMATION AND FUNDAMENTALS

2.1 Problems of Cooperation

Profitable agreements

Profitability implies that countries must find it beneficial to participate in an IEA. For instance, in spring 2001 President Bush announced that the USA would withdraw from the Kyoto Protocol since abatement costs from the 7 per cent emission reduction, as agreed in 1997, were expected to exceed the benefits from reduced global warming. Also many developing countries did not sign this protocol, given their priority for economic development over environmental issues. Generally, although cooperation raises global welfare, individual countries may be worse off. This may happen for ambitious and/or efficient abatement policies if countries have heterogeneous welfare functions, and has been confirmed by many empirical studies, for instance on global warming (IPCC, 2001) or acid rain (Mäler, 1994). In these cases some countries' abatement obligations are too high compared

to perceived benefits from abatement, rendering an IEA unprofitable for them. For instance, an effiient allocation of abatement burdens requires that developing countries with low marginal abatement cost contribute more than industrialized countries to the reduction of greenhouse gases, although developing countries value associated benefits on average less than industrialized countries. Thus, if differences are pronounced enough, developing countries may be worse off from joining an IEA as long as they receive no compensation.

Consensus agreements

Since there are several options in designing a treaty that is profitable for all participants, countries usually find it hard to agree on a particular design. Critical issues are the level of abatement, the allocation of abatement burdens, and the level, kind, as well as the net donors and recipients, of compensation payments. The struggle for consensus is evident from considering how long it takes from the recognition of an environmental problem to the start of negotiations, the signature of an IEA, and its ratification and enforcement. Generally, it seems relatively easy for countries agreeing on 'framework conventions', which are mainly declarations of intention, but far more difficult agreeing on 'protocols' with explicit and serious emission reductions.[5] For instance, the problem of protecting the ozone layer was first discussed at a meeting of the United Nations Environmental Program in 1976. Preparation for a treaty started no earlier than 1981, and concluded with the adoption of the Framework Convention in Vienna in 1985. First reduction targets for ozone-depleting substances were agreed upon in the Montreal Protocol in 1987, which entered into force in 1989. For greenhouse gases the negotiation time was even longer: the Framework Convention on Climate Change (FCCC) was signed in Rio de Janeiro in 1992, but preliminary emission ceilings were agreed no earlier than 1997 under the Kyoto Protocol, modified and relaxed ceilings (without participation of the USA) were only accepted in 2001 at the meeting in Marrakesh. As of August 2002 this treaty had not yet come into force. In addition, acidification of water and soil was first noticed in 1972 at the UN conference in Sweden, and the Framework Convention on Long-Range Transboundary Air Pollution (LRTAP) was signed in 1979 in Geneva, but serious action was taken only in 1985 when the Helsinki Protocol on sulphur reduction was signed. However, not only the agreements on protocols but also on their amendments frequently reflect only the lowest common denominator. For instance, according to Article 20 of the Kyoto Protocol, amendments can only be passed by unanimity. If no consensus can be reached, the changes are only binding for those participants that accepted the amendments. Similar articles are part of almost all protocols.

Therefore, it is not surprising that amendment protocols, which successively tighten emission standards, are signed by substantially fewer countries than the original protocols (see evidence below).

Self-enforcing agreements

Even if countries can agree on the design of a treaty that is profitable for all participants, free-riding jeopardizes the success of IEAs. A country is usually better off either by remaining a non-participant (first type of free-riding) or by acceding to an IEA but violating its terms (second type of free-riding). The first type of free-riding is obvious when it is seen that in most IEAs the number of signatories falls short of the total number of countries involved in the externality problem. This is true at least for those IEAs with explicit and ambitious abatement targets. For instance, the pollutants CFCs and greenhouse gases affect all countries, a total of roughly 200, but only 38 industrialized countries have accepted emission ceilings under the Kyoto Protocol. Also only 26 countries signed the Montreal Protocol in 1987, though participation has risen substantially over recent years to 180 parties at present. However, the more ambitious amendment protocols number fewer participants (London 1990: 153, Copenhagen 1992: 128, Montreal 1997: 63, Beijing 1999: 11; for details see Appendix 3.1). Moreover, though sulphur is a major air pollutant, the 1985 Helsinki Protocol counts currently only 22 parties, of which 16 are EU countries. In contrast, participation in the framework conventions without specific abatement obligations preceding these protocols is very high (FCCC: 186 parties, Vienna Convention: 180 parties and LRTAP: 48 parties).

There is also ample evidence that the second type of free-riding jeopardizes the success of IEAs. Keohane (1995, p. 217) writes: 'compliance is not very adequate. I believe that every study that has looked hard to compliance [of all major IEAs] has concluded . . . that compliance is spotty.' Also Brown Weiss and Jacobson (1997, p. 87ff.) found instances of violations of all IEAs covered by their extensive study. For instance, no less than over 300 infractions of CITES[6] have been counted per year (Sand, 1997, p.25). Moreover, all important parties breached the International Convention for the Regulation of Whaling (Heister, 1997, p. 68).[7]

Effective agreements

In the light of the three fundamental constraints it is evident that as a general conclusion it would be wrong to claim that small IEAs are inferior to large IEAs.[8] Among a small group of countries it might be easier to agree on ambitious abatement targets and compliance might be easier to enforce. Also an inefficient may be superior to an effiient allocation of abatement burdens if it leads to a more symmetrical distribution of the gains from cooperation.

This may ensure a higher rate of participation and compliance and may put less strain on critical countries so that they agree on higher abatement targets. From the discussion it is also evident that success of a treaty cannot be inferred from a high participation rate and degree of compliance. This is not only obvious when considering framework conventions but may also be true for other protocols. If an IEA sets only low abatement targets and/or targets that are close to non-cooperative levels, participation and compliance will be no problem. Thus, success can only be measured if abatement targets under an IEA are compared with estimated abatement levels in the absence of a treaty and, ideally, are evaluated in terms of costs and benefits. For instance, two econometric studies by Murdoch and Sandler (1997a, b) suggest that agreed sulphur reduction under the Helsinki Protocol signed in 1985, and agreed CFC reductions under the Montreal Protocol signed in 1987, though they may seem large, are more in line with non-cooperative than with cooperative behaviour of governments. For the Helsinki Protocol their conclusion can be supported by noting that some members and even some non-members had already achieved the reduction target in 1985 when the treaty was signed, and that not only all members but also most non-members met and even overfulfilled the 30 per cent sulphur reduction target in 1993. This conclusion is also supported by the game-theoretical analysis of Finus and Tjotta (forthcoming), which evaluates sulphur targets under the successor agreement, the Oslo Protocol, signed in 1994 (see section 4).

2.2 Treaty Design

Abatement targets

The level and allocation of abatement targets affect welfare of countries and thus also participation and compliance with treaty obligations. Under many 'old' IEAs uniform emission reduction quotas have been negotiated, which implies that countries have to reduce their emissions by the same percentage for some base year. The list of examples is long and includes several protocols under the umbrella of the framework convention LRTAP. For instance, the Helsinki Protocol suggested a 30 per cent reduction of sulphur emissions from 1980 levels by 1993. Moreover, the Protocol Concerning the Control of Emissions of Nitrogen Oxides or Their Transboundary Fluxes, signed in Sofia in 1988, called on countries uniformly to freeze their emissions at 1987 levels by 1995 and the Protocol Concerning the Control of Emissions of Volatile Organic Compounds or Their Fluxes, signed in Geneva in 1991, required parties to reduce 1988 emissions by 30 per cent by 1999. Only 'modern' IEAs apply the 'principle of different responsibilities', including the Oslo, Kyoto and Montreal Protocols. However, even though the Montreal Protocol allows developing countries to be exempted from certain regulations, to claim a transition period until full compliance is required and to draw on support from various financial mechanisms to meet their targets (see evidence below), it calls on uniform reductions of various CFC pollutants in the different amendments.[9] Also in the original draft of the Kyoto Protocol greenhouse gas emission reductions of the major global players are very similar (USA: 7 per cent, Japan and Canada: 6 per cent and EU: 8 per cent).[10]

Barrett (1992a, b) and Hoel (1992) suggest that uniform abatement obligations constitute some kind of focal point on which bargaining partners can agree relatively easily. However, their models provide little evidence that helps to explain the prominence of uniform quotas. Endres (1996, 1997), Endres and Finus (1998, 1999, 2002), Eyckmans (1999) and Finus and Rundshagen (1998b) compare the outcome and stability of negotiations under different policy regimes, assuming that countries agree on the lowest common denominator. Their main finding is that although uniform quotas are inefficient, the negotiation outcome may be superior in terms of global emission and welfare as well as stability compared to efficient policy regimes since the interests of the blocking country (the country that makes the smallest proposal) are better accounted for in the negotiations. This is also confirmed in the coalition model of Finus and Rundshagen (1998a), where the choice of the policy regime is endogenized (see section 4).

Compensation measures

Transfers are an obvious instrument to compensate the losers from co-operation, to increase participation in an IEA and to encourage compliance. Possible compensation measures are monetary and in-kind transfers, which comprise for instance technical assistance to developing countries from industrialized countries. Whereas monetary transfers directly target compensation, in-kind transfers do so only indirectly and hence the aim of compensation is often blurred and overlapped by other aims. Therefore, theoretically, the efficiency of in-kind transfers is lower than that of monetary transfers. However, the order of frequency of the application of these instruments is reversed in practice. Almost all IEAs have no provisions for monetary transfers. One prominent exception is the Montreal Protocol, under which a multilateral fund has been established to which industrialized countries are supposed to contribute and from which developing countries and countries in transition can receive support. However, recipients can only claim compensation for their incremental costs of abatement (Jordan and Werksman 1996, pp. 247ff. and Kummer, 1994, p. 260).[11] Moreover, payment started only in 1991, but has risen constantly ever since. Outstanding contributions amount to roughly 12 to 16 per cent per year, transfers are often delayed, some donors only issued promissory notes and

some have fulfilled their obligations only in the form of in-kind transfers.[12] A second prominent exception is the Convention of Biological Diversity signed in 1992 in Rio de Janeiro, where developing countries can receive support from the 'Global Environmental Facility'. However, this fund also covers only incremental costs, and the backlog of transfers is very large.[13] Another exception, though different, is the Kyoto Protocol. Among Annex 1 countries transfers are paid indirectly under joint implementation (Articles 3 and 4), where countries can jointly meet their targets in the form of a bubble and paid directly under the emission trading system (Article 17). Under the clean development mechanism (CDM; Article 12) Annex 1 countries can reduce their abatement burdens by financing 'project activities resulting in certified emission reductions' in countries not included in Annex 1 of the protocol.[14] In contrast to monetary transfers, the number of IEAs including provisions for technical exchange and assistance between industrialized and developing countries is larger, though a closer reading reveals that obligations are usually very vague.

Until now, the literature on IEAs has presented little evidence that helps to explain the resistance of governments to pay monetary transfers. Two intuitive arguments are due to Mäler (1990): first, transfers provide an incentive for governments to strategically misrepresent their preferences in order to extract larger compensation payments or to pay low transfers. For instance, under CDM, developing countries may certify emission reductions that they would have undertaken anyway. Also, under the Montreal Protocol, if a developing country indicates non-compliance despite 'best intentions' to the Implementation Council, it may receive additional financial assistance. Second, governments may fear that if they pay transfers they are judged as weak bargainers, which may weaken their position in future negotiations. Further arguments have been developed in three models. First, paying transfers to non-participants for additional abatement efforts may provide a disincentive to join an IEA (Hoel and Schneider, 1997; see section 3.3.4). Second, there is a compliance problem between donor and recipient (Finus, 2002a; see section 4.2). Either the recipient may take the money but does not fulfil its promised abatement obligation or the recipient fulfils its part of the deal but the donor does not pay the promised transfers. Third, there is a compliance problem within the group of donor countries (Barrett, 1994a; see section 4.2). Individual donors are better off if they free-ride, though the group of donors as a whole benefits from transfers through higher participation and compliance.

Issue linkage

An alternative compensation measure is issue linkage, where concessions in one agreement are exchanged for concessions in another agreement. Since package deals are sometimes secretly negotiated, it is not that easy to gather empirical evidence. Most reported examples include bilateral links (Ragland, 1995 and Bennett et al., 1998). For instance, Krutilla (1975) suggests that the Columbia River Treaty of 1961 between the USA and Canada which – viewed as a single issue was to the disadvantage of the USA – was built on concessions by Canada involving North American defence. In the context of multilateral agreements only a wider interpretation allows us to detect issue linkage. One example is the Montreal Protocol, where the import and export of controlled substances with non-parties is banned (Article 4) or the efforts to include environmental issues in the World Trade Organization (WTO), which may be interpreted as a link between an IEA and a trade agreement. Also the provision of technical assistance and exchange under many protocols may be interpreted as a link between an IEA and an agreement to share the cost of R&D. Moreover, considering that the various transboundary pollutants have initially been regulated in separate agreements (sulphur: Helsinki and Oslo Protocols, nitrogen oxides: Sofia, and volatile organic compounds: Geneva) but are now treated together in the Gothenburg Protocol signed in 1999, and that the Kyoto Protocol deals with several global pollutants in one agreement, suggests some kind of issue linkage under the last two mentioned protocols. In the literature, it has been suggested that issue linkage can raise participation (section 3.3.5) and compliance (section 4.5) in an IEA.

Sanctions

Obvious measures to control free-riding are sanctions. However, empirical evidence tells us that either most IEAs have no provision for sanctions or they have hardly been used in the past. Probably, the only exception of sanctioning non-participation is the above-mentioned Article 4 under the Montreal Protocol. For sanctioning non-compliance most IEAs have only a provision for the establishment of an arbitration and dispute settlement committee if one party accuses another of violating the spirit of an agreement (Marauhn, 1996, pp. 696ff., Széll, 1995, pp. 97ff.; and Werksman, 1997, pp. 85ff.). Due to the voluntary character of the arbitration scheme and since the provision contains no threat of punishment, it is not surprising that there are no reported instances of application (Sand, 1996, p. 777). Again, the ozone regime is an exception, where the parties first agreed on an indicative list of measures (Annex V) at their fourth meeting in Copenhagen in 1992 and then defined non-compliance at their sixth meeting in Nairobi in 1994.[15] The measures include (a) assistance in the collection and the reporting of data, technical assistance, technology transfers and financial assistance, (b) issuing cautions and (c) suspension of specific rights and privileges, including transfers of technology, financial mechanism and institutional arrangements. It is

evident that only item (c) can be regarded as sanctions. Moreover, these sanctions can only be used against developing countries since only these can claim assistance and enjoy specific rights and privileges (for example, they are allowed a longer transition period until they have to meet the targets of the various protocols) under Article 5 of the ozone regime.[16] However, any formal statement of non-compliance by the Implementation Committee has to be passed by unanimity.[17] Another exemption is the Kyoto Protocol, where the parties agreed at the meeting in Marrakesh in 2001 on 'Consequences Applied by the Enforcement Branch' (Annex XV).[18] Similar to the Montreal Protocol, most measures include assistance to meet the targets rather than tough sanctions, and complicated voting procedures precede any formal statement of non-compliance. However, two tough punishment options have been decided: a party (a) may be excluded from the emission trading system and (b) must reduce 30 per cent more of its assigned emissions in the second commitment period (2013–17). It remains to be seen whether these sanctions will be used in the future.

In contrast to Chayes and Chayes (1993, 1995), I interpret the empirical evidence on sanctions not to imply that free-riding is not a problem, but to suggest that the design of effective sanctions faces credibility, institutional and technical problems in reality (Finus, 2002a):[19]

1. Sanctioning countries for not acceding to an IEA is at odds with the notion of voluntary participation.
2. Sanctions often also have a negative effect on those countries carrying out the punishment. Thus harsh sanctions are not always credible and constitute themselves as a public good that is subject to free-riding.
3. Sanctioning non-compliance is flawed by the fact that under most treaties signatories can withdraw from the agreement after giving notice three (Kyoto Protocol, Article 27) or four years (Montreal Protocol, Article 19) in advance.
4. Sanctions may be in conflict with the regulations of other treaties (for example, trade sanctions and WTO).
5. Coordination of sanctions among signatories is often time-consuming and costly.

2.3 Basic Framework[20]

Let there be N countries, $i \in I = \{1, \ldots, N\}$, and the welfare of country i, π_i, be given by

$$\pi_i = \beta_i(e_i) - \phi_i\left(\sum_{j=1}^{N} a_{ij} e_j\right) \qquad (3.1)$$

Country i benefits from its own emissions, e_i, where it is usually assumed that benefits increase ($\beta_i' > 0$) at a decreasing rate ($\beta_i'' \leq 0$). Thus emissions can be viewed as an input in the production and consumption of goods where the law of diminishing returns applies. Country i also suffers damages from its own (e_i) and foreign ($e_j, j \neq i$) emissions. The transportation coefficient a_{ij}, $0 \leq a_{ij} \leq 1$, indicates the portion of emissions of country j, which is deposited in country i. Whereas for local pollutants, $a_{ii} = 1$ and $a_{ij} = a_{ji} = 0$, the transportation coefficients will be between zero and one for transboundary pollutants, as for instance the acid rain pollutant sulphur. For an upwind country, like the UK, a_{ij} and a_{ji} will be small and for a downwind country, like Norway, these coefficients will be large. For global pollutants, like CFCs and greenhouse gases, all coefficients are one since emissions disperse uniformly in the atmosphere.[21] The standard assumption is that damages increase in depositions ($\phi_i' > 0$) at an increasing rate ($\phi_i'' \geq 0$). Hence, due the limited absorption and regeneration capacity of most environmental systems, environmental damages increase more than proportionally with increasing depositions.

Since benefits from abatement correspond to reduced damages from depositions and cost of abatement correspond to a loss of benefits from reduced emissions (opportunity cost of abatement), a country's welfare function (also called payoff function in the game-theoretical terminology) has also been modelled in terms of abatement in the literature. Since qualitative results are not affected by such a change, I will relate all subsequent models to (3.1) in order to use a uniform terminology.

If each of the N countries pursues its own interest, that is, all countries behave non-cooperatively, each country maximizes (3.1) with respect to its own emissions ($\max_{e_i} \pi_i$), taking emissions from other countries as given. The simultaneous solution of the N first-order conditions $\beta_i' = a_{ii}\phi_i'$[22] delivers the non-cooperative Nash equilibrium emission vector $e^N = (e_1^N, \ldots, e_N^N)$. Since this equilibrium *de facto* implies that countries form singleton coalitions, it seems plausible to assume that it represents the *status quo* before an IEA is signed. In contrast, if governments were to pursue the common interest, that is, they behaved *fully cooperatively*, they would maximize the aggregate payoff over all countries ($\max_{e_1, \ldots, e_N} \sum_{i=1}^{N} \pi_i$). Again, the simultaneous solution of the N first-order conditions $\beta_i' = \sum_{j=1}^{N} a_{ji}\phi_j'$ delivers the fully cooperative (also called globally or socially optimal) emission vector $e^S = (e_1^S, \ldots, e_N^S)$. This may be interpreted as if all countries form a grand coalition and jointly maximize the aggregate welfare of their coalition. Since $e^S \neq e^N$ as long as there is some transboundary pollution ($a_{ij} \neq 0$ for some $j \neq i$), global welfare could be raised through cooperation, that is, $\sum_{i=1}^{N} \pi_i(e^N) < \sum_{i=1}^{N} \pi_i(e^S)$. This is also true for more pragmatic solutions

(which most IEAs are), where either the grand coalition chooses more moderate abatement targets than in the social optimum or only a subgroup of countries forms a coalition (coalitions), implying a partially cooperative emission vector $e^n = (e_1^n, ..., e_N^n)$. However, in the basic framework, no form of cooperation can be enforced. In a static game, any strategy of conditional cooperation ('I will cooperate provided you also cooperate') would be irrational since it cannot be rewarded at later stages. Thus any other emission vector different from the (static) Nash equilibrium would imply that at least one country has an incentive to revise its decision. Thus, to explain any form of cooperation requires extending the basic framework (see section 2.4).

In order to study the free-riding behaviour of countries in the context of coalition formation, it is helpful to note that the first-order conditions derived from the maximization behaviour of single countries or coalitions for any coalition structure[23] different from the grand coalition can be interpreted as best-reply functions. A best-reply function describes the optimal choice of emissions of a country (coalition) for a given level of emissions of outsiders (and given transportation coefficients). Total differentiation of the first-order conditions delivers the slopes of the reaction functions that approximate the direction and the extent of change of emissions of countries (coalitions) to an external change of emissions. Usually, these functions are negatively sloped since an increase (decrease) of external emissions increases (reduces) marginal damages and a best reply calls on a country (coalition) to reduce (increase) emissions, which increases (reduces) marginal benefits in order to equalize marginal benefits and damages. Only under special conditions ((a) $a_{ij}=0$, (b) $a_{ij}=0$ and (c) linear damage cost functions) is the optimal choice of a country (coalition) independent of external emissions (dominant strategy) and the slope of a country's (coalition's) reaction function zero. The literature refers to the standard case as non-orthogonal and to the special case as orthogonal best-reply functions.[24]

2.4 Extended Framework

Table 3.1 provides an overview of important features (column 1), subfeatures (column 2) and characteristics (columns 3 and 4) according to which various coalition models can be structured. Those characteristics that can be related to the basic model are indicated in italic. All other entries are related to extended frameworks, which suggests that the number of possible extensions is large. Therefore, in order not to lose track at this stage of the discussion, I will only briefly sketch some important issues of Table 3.1 and encourage the reader to return to this subsection after reading

Table 3.1 Structure of coalition models

Main features	Sub-features	Characteristics	
Time	Framework	Implicit dynamic	Explicit dynamic
	Horizon		Finite or infinite
	Interval		Discrete or continuous
Payoff	Structural relation	*Independent* (flow pollution)	Dependent (stock pollution)
	Arguments	*Only material payoffs*	Also non-material payoffs
	Transfers	*No*	Yes
Equilibrium concepts	Strategic relation	*Independent*	Dependent
	Sanctions	Different degrees of harshness and credibility of sanctions	
	Deviations	*Single*	Multiple
Number of issues		*Single*	Multiple
Rules of coalition formation	Sequence of coalition formation	Simultaneous	Sequential
	Number of coalitions	Single	Multiple
	Membership	Open	Exclusive
	Consensus	Different degrees of consensus with respect to membership	

sections 3 and 4 to gain a full understanding of the driving forces of coalition models and their classification.

The first main feature and an important prerequisite for cooperation is '*Time*'. Whereas non-cooperative behaviour is the only equilibrium strategy in the static basic model (conditional cooperation is not possible), cooperative behaviour is possible in a dynamic extended model since countries can condition their strategies on previous behaviour and/or can react to deviations from agreed strategies through some form of punishment.[25] However, in some models the dynamic aspect is not immediately obvious. I call this an 'implicit dynamic framework', which means that time is not explicitly modelled and the dynamic story is exogenous to the model. In contrast, an 'explicit dynamic framework' implies that 'real' time is captured and modelled. In the case of an explicit dynamic time framework, the time horizon can be either finite or infinite and the time intervals can be either discrete or continuous. An infinite time horizon does not necessarily imply an infinite life of agents but only that the end of the game is not

known with certainty. Discrete time implies that strategies can only be revised at certain points in time whereas strategies can immediately be revised if time is continuous.

The second main feature is '*Payoff*'. The first sub-feature, 'Structural relation', is closely related to the dynamics of a model (main feature 'Time'). 'Structural independence' means that payoffs at time t depend only on strategies (that is, emissions) at time t whereas structural dependence implies that they also depend on previous strategies. Since in the context of IEAs most coalition models capture only structural dependence with respect to damages from emissions, the line of distinction can also be drawn between the assumption of flow and stock pollutants. This sub-feature can also be related to three important games. Repeated games assume the same payoff function at each point in time and usually discrete time intervals, though we will encounter a version with continuous time in section 3.2. In contrast, difference and differential games capture structural dependence of payoffs where the former assume discrete and the latter continuous time (Dockner et al., 2000 and de Zeeuw and van der Ploeg, 1991). Of course, trivially, since the basic model is static, its payoff structure can be classified as independent. The second sub-feature concerns the arguments in countries' payoff functions. Whereas 'material payoffs' refer to benefits and costs from emissions as captured in the basic model in equation (3.1), all other dimensions such as reputation and fairness that usually favour more cooperation are captured by the term 'non-material payoffs'. The same positive effect usually applies to the third sub-feature, 'Transfers', which may be seen as an additional strategy to emissions to achieve cooperation.

The third main feature is captured by the term 'Equilibrium concepts', where the first two sub-features, 'Strategic relation' and 'Sanctions' also have a close connection to the dynamics of a model. Strategic independence implies that strategies are chosen once and for all and cannot be revised, whereas strategic dependence implies that strategies at time t are conditioned on previous actions and can be revised if new information becomes available.[76] Consequently and trivially, there is no strategic dependence in the basic model due to its static nature. The second sub-feature, 'Different degrees of harshness and credibility of sanctions', is related to two facts. First, in a dynamic setting the free-rider incentives (of type 1 and 2) do not vanish but may be controlled through either implicit or explicit threats of sanctions. Second, threats of punishment have to be credible to be deterrent, which corresponds to different notions of equilibrium concepts discussed in subsequent sections. Those notions are also related to the third sub-feature, 'Deviations'. Whereas we defined stability in the basic model as a state that is immune to single deviations (Nash equilibrium), some coalition models define stability in terms of multiple deviations. Of course, in the basic model, this simple definition was sufficient since cooperative agreements were not stable anyway, but is less obvious in extended models where full or partial cooperation is possible.

The fourth main feature is the '*Number of issues*'. Whereas the basic model restricted attention to one issue, that is, one pollutant, some coalition models also consider multiple issues, as for instance additional pollutants, trade flows, investment in R&D and so on. Multiple issues can improve upon the possibilities of establishing cooperation between countries if issues are cleverly and strategically linked. The success of issue linkage depends on a number of factors, which are discussed in subsequent sections, but the main reason is that issue linkage, like transfers, increases the number of policy options (strategies) to achieve cooperation.

The fifth main feature is the 'Rules of coalition formation'. The rules may be interpreted as the institutional setting in which countries strike informal or formal cooperative agreements with other countries. At this stage it suffices to point out that the rules of coalition formation have a crucial impact on the outcome, but its role has only recently been analysed in a strand of literature that I call 'new coalition theory', discussed in subsection 3.3.6.

In summarizing the preliminary discussion, five conclusions seem important. First, a dynamic time framework is the most important ingredient and extension compared to the basic model in order to capture the phenomenon of cooperation. Second, it will become apparent from sections 3 and 4 that the extensions non-material payoffs, transfers and issue linkage will usually have a positive effect on the possibility of cooperation. In terms of the rules of coalition formation it will be evident that the possibility of forming multiple coalitions instead of only one coalition, restricting membership to an IEA (exclusive membership) instead of allowing any country to join and requiring a high instead of a low degree of consensus with respect to membership in an IEA will lead to superior outcomes in terms of global welfare and emissions. Third, no clear-cut conclusions about the effect of characteristics of other sub-features that constitute an extension to the basic model can be drawn. This will depend on the specifics of models. Fourth, roughly speaking, the right-hand-side characteristics (fourth column) in Table 3.1 imply a higher degree of sophistication than the left-hand-side characteristics (third column). However, sophistication comes at the cost of complexity. Therefore, it will become apparent that all models make some exogenous assumptions and solve for the remaining endogenous variable in order to keep the analysis tractable. For instance, all coalition models assume certain rules of coalition formation when determining equilibrium coalition structures but do not derive these rules from the negotiation process between the potential

participants to an IEA. Moreover, all models focus either on the first or the second type of free-riding and capture the other type of free-riding only deficiently. I take this phenomenon as the fundamental feature to structure the following discussion. I call models that focus on the first type of free-riding 'membership models' (M-models: section 3) and those that focus on the second type of free-riding 'compliance models' (C-models: section 4). M-models are concerned with the coalition formation process and stability of membership. They analyse whether a country remains a non-participant or participates in a coalition and, if it participates, with which countries it will form a coalition. However, M-models are not concerned with whether and how agreed emission ceilings within a coalition are enforced. This is the focus of C-models, which emphasize the role of sanctions in enforcing compliance. However, C-models usually start their analysis from a given membership and give less attention to the process of coalition formation and issue of membership.[27] Fifth, it will be apparent that if structural dependence is modelled (which is only the case in some M-models, applying the stability concept of the core; subsection 3.2), this is only done in terms of emissions. The reason is simple: in all models payoffs are a function of emissions and transfers, which are only indirectly a function of membership. Moreover, whereas for emissions a strategic dependence is interesting because of stock pollutants, it is less interesting for transfers as long as it is assumed that transfers at time t are paid out of the gains from cooperation at time t. Also, if strategic dependence is assumed, it is usually only modelled in terms of emissions and transfers (which is the case in some M-models applying the stability concept of the core; subsection 3.2, and in all C-models; section 4), though it would generally be possible (and very useful) to capture strategic dependence in terms of membership.

3 MEMBERSHIP MODELS

3.1 Introduction

Membership has been analysed within cooperative and non-cooperative game theory. The classical distinction is that cooperative game theory assumes the possibility of binding agreements whereas non-cooperative game theory neglects this possibility. However, it will become apparent that this distinction is not very helpful since all M-models share some fundamental features.[28] First, not only cooperative but also non-cooperative game-theoretical M-models assume some form of commitment within coalitions. That is, all M-models assume that countries comply with their

emission reduction and transfer obligations if they form a coalition and therefore free-rider problems of real IEAs are underestimated. Second, not only M-models belonging to non-cooperative game theory but also those belonging to cooperative game theory assume some form of punishment if countries leave an agreement. Third, all M-models check stability of membership in an implicit dynamic framework. That is, they analyse whether a country or group of countries have an incentive to move from a particular coalition structure (state 1) to another coalition structure (states 2, 3, ...), where the time path to switch from one to another state is not modelled. Therefore, I propose to distinguish both theories in terms of their tools and foci.

The first attempts to study coalition formation are rooted in cooperative game theory.[29] The analytical tool is the characteristic function (see Definition 1, below) that assigns to each coalition a worth, which is the aggregate payoff a coalition can get irrespective of the behaviour of outsiders. What irrespective means depends on the specific assumptions associated with this function and will be discussed in subsection 3.2.1. The focus of the analysis is on the allocation of the gains from cooperation, but not that players may choose inefficient strategies. Therefore, in games with externalities, stability of the grand coalition implementing the socially optimal strategy vector is analysed. The central question of the analysis is: which transfer scheme or bargaining rule enables the grand coalition to be sustained?

Proponents of non-cooperative game theory criticize three features of cooperative coalition theory: (1) for rational actors it seems natural to assume that they base their decision about membership on individual rather than on aggregate payoffs; (2) some assumptions of cooperative game theory about the behaviour of countries outside a coalition are difficult to justify since they require irrational behaviour of countries (see section 3.2.1 for details); (3) cooperative game theory cannot explain why most IEAs are inefficient in terms of participation and emission reductions. Therefore, scholars of non-cooperative game theory propose analysing coalition formation based on a valuation function (see Definition 1) that assigns an individual payoff to each player for any possible coalition structure, assuming that each coalition pursues its own interests, and not restricting coalition formation to the grand coalition. Hence the behaviour of insiders and outsiders is guided by self-interest and is based on the same assumption of rationality. That is, countries cooperate within their coalition but behave non-cooperatively against outsiders. Therefore, higher than globally optimal emissions and inefficient coalition structures (different from the grand coalition) typically emerge in equilibrium. The central question of the analysis is: which coalition structure can be sustained as an equilibrium for a given transfer scheme or bargaining rule?