CICERO Working Paper 2002:06

Enforcing the Climate Regime:
Game Theory and the Marrakesh Accords

Jon Hovi

November 2002

CICERO
Center for International Climate
and Environmental Research
P.O. Box 1129 Blindern
N-0318 Oslo, Norway
Phone: +47 22 85 87 50
Fax: +47 22 85 87 51
E-mail: admin@cicero.uio.no
Web: www.cicero.uio.no

CICERO Senter for klimaforskning
P.B. 1129 Blindern, 0318 Oslo
Telefon: 22 85 87 50
Faks: 22 85 87 51
E-post: admin@cicero.uio.no
Nett: www.cicero.uio.no

Tittel: Enforcing the Climate Regime: Game Theory and the Marrakesh Accords

Forfatter(e): Jon Hovi

CICERO Working Paper 2002:06
23 sider

Finansieringskilde: Norges forskningsråd

Prosjekt: Håndheving av Kyoto-avtalen

Prosjektleder: Jon Hovi

Kvalitetsansvarlig: Hege Westskog

Nøkkelord: Kyotoprotokollen, håndheving, forhandlinger

Sammendrag: Forfatteren gjennomgår et antall innsikter om håndheving av internasjonale avtaler som kan avledes fra ulike ikke-kooperative likevektsbegreper. Samtidig evalueres Marrakesh-avtalens bestemmelser for håndheving av Kyoto-protokollen i lys av disse innsiktene. Fem ulike likevektsbegreper diskuteres – Nash-likevekt, delspillperfekt likevekt, reforhandlingssikker likevekt, koalisjonssikker likevekt og perfekt Bayesiansk likevekt. Disse likevektsbegrepene har en rekke implikasjoner vedrørende betingelsene for effektiv håndheving, herunder: (1) Eventuelle reaksjoner på registrerte overtredelser bør være mer enn proporsjonale med bruddets omfang; (2) For at slike reaksjoner skal være troverdige, bør de være av en slik art at de medfører en bevegelse langs Pareto-grensen – ikke innføring av en sub-optimal tilstand. (3) Et effektivt håndhevingssystem må kunne tøyle såvel kollektive som individuelle incentiver til å jukse. (4) Et regime som åpner for fullt innsyn i alle deler av håndhevingsprosessen er ikke ubetinget av det gode. Det konkluderes med at det å konstruere et effektivt system for håndheving av Kyoto-protokollen er en formidabel oppgave, som Marrakesh-avtalen bare delvis evner å løse. En skal imidlertid huske at partene under utformingen av håndhevelsesregimet også hadde andre viktige hensyn å ivareta, som måtte avveies mot ønsket om å lage et mest mulig effektivt håndhevingssystem.

Språk: Engelsk

Rapporten kan bestilles fra:
CICERO Senter for klimaforskning
P.B. 1129 Blindern
0318 Oslo

Eller lastes ned fra:
http://www.cicero.uio.no

Title: Enforcing the Climate Regime: Game Theory and the Marrakesh Accords

Author(s): Jon Hovi

CICERO Working Paper 2002:06
23 pages

Financed by: Research Council of Norway

Project: Enforcing the Kyoto Protocol

Project manager: Jon Hovi

Quality manager: Hege Westskog

Keywords: Kyoto Protocol, enforcement, negotiations

Abstract: This article reviews basic insights about compliance and "hard" enforcement that can be derived from various non-cooperative equilibrium concepts, and evaluates the Marrakesh Accords in light of these insights. Five different notions of equilibrium are considered – the Nash equilibrium, the subgame perfect equilibrium, the renegotiation proof equilibrium, the coalition proof equilibrium, and the perfect Bayesian equilibrium. These various types of equilibrium have a number of implications for effective enforcement: (1) Consequences of non-compliance should be more than proportionate. (2) To be credible punishment needs to take place on the Pareto frontier, rather than by reversion to some suboptimal state. (3) An effective enforcement system must be able to curb collective as well as individual incentives to cheat. (4) A fully transparent enforcement regime could in fact turn out to be detrimental for compliance levels. It is concluded that constructing an effective system for "hard" enforcement of the Kyoto Protocol is a formidable task that has only partially been accomplished by the Marrakesh Accords. A possible explanation is that the design of a compliance system for the climate regime involved a careful balancing of the desire to minimize non-compliance against other important considerations.

Language of report: English

The report may be ordered from:
CICERO (Center for International Climate and Environmental Research – Oslo)
PO Box 1129 Blindern
0318 Oslo, NORWAY

Or be downloaded from:
http://www.cicero.uio.no

Contents

1 Introduction

The process of constructing an international regime on climate change, initiated by the Rio Conference in 1992, has recently suffered two serious blows and celebrated two moderate triumphs.[1] The first blow was the failure in November 2000 to reach agreement at the sixth Conference of the Parties (COP-6) in the Hague. The second blow was the announcement by President George W. Bush in March 2001 that the United States was not going to ratify the Kyoto Protocol. These two events led some observers to doubt the realism of developing a viable and effective regime to control emissions of greenhouse gases.

The two triumphs were that the Parties proved able to reach agreement at the Bonn and Marrakesh conferences in July and November 2001. While watering down considerably some of the provisions of the original treaty, the Bonn and Marrakesh deals make it likely that by the end of 2002 the Kyoto Protocol will have been ratified by a sufficient number of countries, and will thus have entered into force. It therefore seems that the Dutch environment minister, Jan Pronk, was right when after the Hague meeting he said that "we did not succeed [at the COP-6, but] looking back, I think it is better to say that perhaps we did not yet succeed".[2]

One of the main challenges of constructing a climate regime has been the design of an effective system for monitoring, verification and enforcement. Article 18 of the Kyoto Protocol deferred this important task, including "the development of an indicative list of consequences, taking into account the cause, type, degree and frequency of non-compliance", to the subsequent Conferences of the Parties. Compliance was one of the most contentious issues in Bonn, and the specifics of the compliance system were again deferred. Further progress was made in Marrakesh, where the parties reached agreement on major procedures and consequences of non-compliance. However, whether or not the compliance mechanism is going to be legally binding will not be decided until the first Meeting of the Parties (MOP) following the entering into force of the Kyoto Protocol.

A considerable body of work has emerged on the question of compliance in relation to the Kyoto Protocol. A number of writers have sought to characterize the protocol's overall approach to compliance and commented on the provisions that are particularly relevant for this question (Bodansky 2001, Grubb 1999, Oberthür & Ott 1999). Others have discussed how a climate regime can achieve clarity with regard to obligations, party performance and regime response (Mitchell 2001). Yet others have proposed possible responses that the regime might invoke against a party that is found to be in non-compliance (e.g., Hargrave et. al 1999, Heister et. al. 1997).

However, no systematic attempt has so far been made to identify the conditions under which a compliance system for the climate regime is likely to be effective. In this article, we use noncooperative game theory in an attempt to fill this gap. One of the core elements in this theory is the notion of equilibrium.[3] Since an equilibrium concept is basically a notion of stability, this is a natural place to look for conditions for effective enforcement. Accordingly, we basically do two things. First, we survey a number of equilibrium concepts, and ask what

1 This introduction was finalized in August 2002.

2 Cited in The Economist, December 2nd 2000:22.

3 Other core concepts include information and credibility (Guk 1997).

these concepts tell us about the conditions for effective enforcement. Second, we ask to what extent these conditions are satisfied by the design of the compliance mechanism in the Marrakesh Accords.

The article is organized as follows. Having made some introductory remarks in section two, we focus in section three on the notion of Nash equilibrium, the mother of all the other solution concepts to be explored in this article. Subsequent sections discuss four so-called refinements of the Nash equilibrium.[4] Section four considers the subgame perfect equilibrium, section five the renegotiation proof equilibrium, section six the coalition proof equilibrium, and, finally, section seven the perfect Bayesian equilibrium. In each section we offer a simple explanation of the refinement in question, and show how each solution concept can be motivated by the need to curb certain types of incentive for non-compliance. This is followed by a discussion of possible implications of each equilibrium concept for the design of an effective enforcement system for the climate regime. Some of these implications will likely seem immediately appealing to most readers, while others are less intuitive and therefore perhaps also more interesting. Throughout, we seek to make the presentation accessible even to readers with minimal training in game theory.

2 Preliminaries

It is a commonplace in the theory of international politics that the anarchic (or self-help) character of the international system creates potential difficulties for the enforcement of international agreements.[5] Absent an authoritative international judicial system, empowered with the mandate and means to judge and sanction states that violate treaties, other instruments must be relied upon to ensure compliance. In practice, enforcement must either be carried out by the parties themselves, or by some institution erected, accepted and empowered by the parties. We shall call these two types of enforcement decentralized and centralized, respectively.[6]

It is often claimed that because of the anarchy condition, international treaties must be *self-enforcing* (e.g., Finus 2001:4, Heister et. al. 1997:23).[7] However, it is not entirely clear what this means in practice. At least two definitional issues need to be adressed. A first question is what kinds of responses to non-compliance are consistent with a *self*-enforcing agreement. Some authors (e.g., Barrett 1994) seem to denote as self-enforcing any agreement that is policed by the parties themselves, rather than by a third party (such as a court). Others (e.g., Hovi 1998) restrict the term further, demanding that responses to non-compliance must also be confined to issues covered by the agreement itself. According to the latter view, so-called external enforcement mechanisms (such as the use of trade sanctions to enforce a treaty on climate change) are incompatible with the notion of a self-enforcing agreement. The idea is that it is misleading to call an agreement self-enforcing if elements outside the agreement are needed to police it. To be *self*-enforcing, an agreement must both be policed by the parties

4 The equilibrium concepts considered in this paper does not represent an exhaustive list of all existing refinements of the Nash equilibrium See van Damme (1991: ch. 1) for an introduction to other proposals.

5 Key references are Axelrod (1984), Keohane (1984), Oye (ed., 1986) and Taylor (1976, 1989).

6 We believe that both centralized and decentralized means of enforcement has a role to play in the climate regime. Accordingly, we use models and illustrations from both categories.

7 A related but not identical concept is that of „stable" agreements (or coalitions), which is sometimes said to exist if no free-riding takes place (e.g., Carraro & Siniscalco 1993)..

themselves *and* be enforceable by internal responses alone. In a self-enforcing agreement, therefore, violations are typically deterred by threats of exclusion, suspension of member rights, or termination of the treaty.[8] In the context of global warming, more specific examples of internal punishments are strictures on the right to emissions trading, and reduced national allowances in subsequent commitment periods.[9]

A second and perhaps more important definitional issue concerns the conditions for an agreement to be *enforced*. Enforcement is the process of ensuring compliance with some behavioural or outcome standard, as laid down in an agreement, a rule, a law, a norm or in some other way.[10] The concept is used in at least two related meanings. A wide definition includes the use of "soft" instruments (such as supervision and capacity-building) as well as "hard" consequences (e.g., financial penalties or suspension of privileges). A more narrow definition says that only hard measures should be subsumed under the concept.[11] The latter way of using the term seems to underlie the language of the Marrakesh Accords, insofar as the Compliance Committee has been separated in two sections, a "facilitative" branch and an "enforcement" branch, where only the latter is authorized to invoke hard consequences. This narrow definition will also be used in this paper. By implication, the Kyoto protocol is "enforced" to the extent that hard consequences are being used to deter non-compliance or as a response to observed non-compliance.

Before we proceed, a final comment is in order. Chayes and Chayes (1993, 1995) argue that international compliance owes little to enforcement (i.e., to the "hard" approach). Instead, they insist, it is the management (or "soft") approach that holds the key to compliance in the international system. Somewhat in the same spirit, Stranlund (1995) applies a formal model to show that under certain conditions, a system where voluntary compliance is encouraged by a subsidy, can dominate a regime in which mandatory compliance is sustained by a fine. This rather optimistic view is apparently not fully shared by the parties to the Marrakesh Accords.[12] Following proposals by the United States and the European Union, the parties have divided the Compliance Committee in two components. The facilitative branch will be responsible for providing advice and facilitation to the parties, while the enforcement branch will focus on quantified emissions limitations or reduction commitments and on eligibility for the flexible mechanisms. The underlying premise seems to be that there is a need for both "soft" facilitation and "hard" enforcement. It is the view of the present authors that this was a wise decision. True, international non-compliance is sometimes the result of insufficient capacity by states to fulfil their obligations, despite the best of intentions. But regimes must also be prepared for occasional or even more frequent instances of deliberate violations. Thus, although we certainly acknowledge the importance of facilitative procedures, we believe that there is a need for a system of hard enforcement as well. In the

8 In trivial cases, such as in pure coordination games, sanctions are superfluous, because even in the short run nothing can be gained by defecting.

9 Cf. Kerr (1998).

10 For example, by a threat or a promise.

11 Note that "hard" enforcement includes both attempts to deter or prevent non-compliance and measures designed to punish a perpetrator or re-establish compliance after deterrence has failed.

12 Nor is it universally accepted in the literature on compliance. For example, Downs et. al. (1996) argue that a high rate of compliance often stems from treaties being formulated in a manner that requires them to do little more than they would do without the treaty. In other words, a high rate of compliance can be an indication that the relevant treaty is relatively unambitious.

following we examine the conditions under which a system of hard enforcement is likely to prevent intentional non-compliance.

3 Nash Equilibrium

It is no exaggeration to say that the Nash equilibrium is one of the core concepts of game theory. There are two main ways of defining a Nash equilibrium.[13] The first and most common definition says that a Nash equilibrium is a set of strategies, one for each player, which are best responses vis-à-vis each other. The second definition starts with the assumption that each party chooses its strategy in ignorance of the strategies pursued by the other players. The outcome is then a Nash equilibrium if neither side has a reason to regret its own choice of strategy when it learns about the strategies of the other parties. Notice that these two formulations are not rival definitions. The words differ, but the meaning is the same.

What does the Nash equilibrium tell us about the conditions for regime compliance? A very general answer is that the Nash equilibrium provides a minimal condition for stability. If an agreement contains provisions that establish compliance as a Nash equilibrium, then, *given* the strategies of the other parties, no signatory can benefit from non-compliance.

A more specific lesson from the Nash equilibrium is that non-compliance is best deterred by threatening *severe* punishment. In fact, the more severe the penalty, the more likely it is that compliance is sustained as a Nash equilibrium To see why, consider the simple model in figure 1.

Figure 1

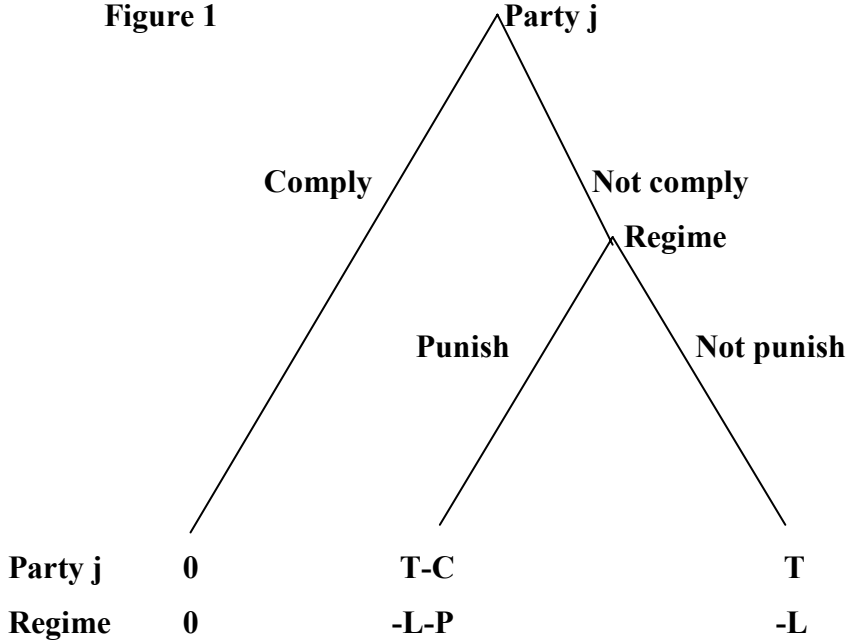


Figure 1 depicts the compliance game as one between a particular party (here referred to as Party j) on the one hand, and the climate regime on the other. It may be useful to think of "the regime" more specifically as the enforcement branch of the Compliance Committee. Thus, we are here dealing with a case of what we have previously called centralized enforcement. The

13 The concept was originally introduced by John F. Nash, hence the term "Nash euilibrium" (Nash 1951).

payoffs may be interpreted as costs and benefits compared to the situation where Party j complies. T is the additional payoff obtained by Party j if it chooses not to comply, and C is the cost suffered by Party j if punished. Finally, L is the cost imposed on the regime if Party j does not comply, while P is the cost incurred by the regime if it punishes Party j.

For our purposes, it is useful to consider the game on normal form, as shown in Figure 2.

Figure 2

		Regime	
		Punish if Party j does not comply	**Not punish if Party j does not comply**
Party j	**Comply**	**0 , 0**	**0 , 0**
	Not comply	**T-C , -L-P**	**T , -L**

Note that if Party j complies, then *both* of the regime's strategies are best replies. The reason is that if Party j complies, then the regime has no move to make, meaning that its choice of strategy has no impact on the outcome. The condition for the top left cell to be a NE, therefore, is simply that $T \leq C$. At first sight, this result might seem to contradict the claim made by Heister et al. (1995) that a system of strict proportionality between the punishment and the gains from cheating (i.e., a system where T=C) would be unable to deter non-compliance. However, if there is the slightest delay between a breach and the imposition of the punishment, and future payoffs are discounted, then the claim made by Heister et al. remains correct even in cases where T equals C. In other words, the Nash equilibrium suggests that the penalty level should be more than proportional. In fact, other things being equal, then the more severe the punishment (the larger C is), the more likely it is that compliance is sustained as a Nash equilibrium

It is hardly surprising that these simple lessons from the Nash equilibrium are reflected in the Parties' own thinking about the compliance problem in the climate regime. For example, a note from the President of the COP-6 expressed a general concern about the need for high penalty levels:

> [The] Parties recognize that penalty rates will be an essential element of the compliance system. Although they will partly serve as an interest rate for the delays in the achievement of emission commitments, they should also serve as an incentive to comply and they should, therefore, be set at a relatively high level (COP-6 2000:12).

The note then goes on to propose a system with more than proportionate penalty levels: [14]

> [I]f a Party has been determined as being in non compliance with its commitments under Article 3.1, the enforcement branch should apply the following consequences:
>
> - Subtraction of excess emissions from the assigned amount of the subsequent commitment period.
> - [The] Penalty rate should be set at 1.5 and be increased by 0.25 after the subsequent commitment period if the Party is not in compliance at the end of the subsequent commitment period. (COP-6 2000:12)

The Marrakesh Accords set the penalty rate (or restoration rate, as it is now called) at 1.3 rather than 1.5, but the principle of more than proportionate responses was retained.

Few would probably deny that the severity of consequences is a significant determinant of compliance. However, since the Nash equilibrium *only* focuses on the severity of consequences, one may reasonably question whether all Nash equilibria are plausible candidates for a self-enforcing agreement. Indeed, the equilibrium concepts to be discussed in subsequent sections illustrate the following simple, but important point: Even if a future enforcement regime of the Kyoto Protocol should manage to sustain compliance as a Nash equilibrium, a Party may still have an incentive to cheat. The requirements of a Nash equilibrium are simply too weak to ensure compliance by rational decision makers. We therefore need to consider stronger solution concepts.

4 Subgame perfection

A problem with the Nash equilibrium is that it is possible for a strategy to prescribe irrational behaviour for some parts of a game (some subgames) and still belong to a Nash equilibrium for that game. For example, a Nash equilibrium can be sustained by a threat that is empty in the sense that it would not be individually rational for the threatener to carry it out, should the relevant transgression take place. Worse, this is true even if the game is one of complete information, *so that it is common knowledge that the threat is empty*. Clearly, if a threat is known with certainty to be empty, there is no reason for the threatened party to let it influence its behaviour.

Consider for a moment the fictitious (and unthinkable) proposal that any instance of non-compliance with the Kyoto Protocol, however small, be punished with a major strike with nuclear weapons. Assume, furthermore, that the probability is zero that such a response would actually be carried out. Finally, suppose that all of this is common knowledge. Under these circumstances such a threat could safely be ignored by the relevant states. The fact remains, however, that this threat would establish compliance as a Nash equilibrium. The explanation is that the Nash equilibrium does not even ask whether or not a threat is credible. It merely asks if compliance is a best response, *given* that non-compliance causes the threat to be carried out. It goes without saying that to comply is a best response *if* one would otherwise suffer a nuclear attack. Conversely, *given* that all countries comply, it is costless to threaten nuclear bombardment, because a need to carry the threat out never arises. This explains why,

14 This is true even though proportionality is in fact proposed by UNFCCC (2000:section II, article 2a) as a basic principle of the compliance entity.

given that all parties comply, it qualifies as a best reply to threaten nuclear extinction of any defecting state.[15]

The Nash equilibrium requires that the players' strategies prescribe rational behaviour on the equilibrium path.[16] However, it ignores whatever happens *off* this path. The above example makes it clear that this is unreasonable. In the case of nuclear deterrence of non-compliance, it is not enough to establish that it is rational to comply, *given* that non-compliance is punished in this way. One also needs to ask if that punishment will actually be implemented, should a transgression occur. This is done by the subgame perfect equilibrium, which requires that the players' strategies prescribe rational behaviour in *all* parts of the game, not only on the equilibrium path.[17] More specifically, the subgame perfect equilibrium requires that the players' strategies must be best replies to each other for all subgames.[18] The game in Figure 1 has two subgames. In addition to the entire game, there is also a subgame that consists only of the regime's move. The subgame perfect equilibrium requires that the parties' strategies prescribe rational behaviour for both of these subgames. What this means in practice, is that a second condition for effective deterrence of non-compliance is added to the one generated by the Nash equilibrium. It is, in other words, not enough that the threatened punishment is *severe*. In addition, it must also be *credible*. In fact, the above example reminds us that a very harsh punishment may be less than ideal, precisely because it is unlikely to be credible. The Nash equilibrium where compliance is assured by way of nuclear deterrence does not qualify as a subgame perfect equilibrium, since presumbly no-one would find it in their best interest to respond to a violation of the Kyoto Protocol with a nuclear strike.[19] It therefore goes without saying that less severe means of punishment must be relied upon to deter non-compliance. Experience from other international treaties suggests, however, that far more modest reactions may not be credible either. For example, international commodity agreements regularly provide for, but very rarely invoke, suspension of the right to vote as a potential response to non-compliance (Werksman 1998:27). Similarly, Chayes and Chayes (1995) point out that loss of membership or other privileges, although explicitly included as an option in many international treaties, are typically used only in extraordinary circumstances.[20] This seems to indicate that lack of credibility is a quite widespread problem for international enforcement.

It is not impossible that this will be the case for the climate regime as well. For example, according to the Marrakesh Accords, one of the consequences available to the enforcement branch is suspension of the eligibility to emissions trading. However, if a party's right to sell quotas is suspended, this is likely to affect third parties via the markets for quotas and energy (Hagem, Mæstad & Westskog, forthcoming). The impact will be particularly large if the relevant country is a major actor in the market for quotas. For example, if Russia – a major supplier of quotas – were to have its right to trade suspended, a significant price

15 It is, however, not the only best reply. In fact, given that all signatories comply, so that no response actually needs to be made by the regime, any strategy is equally good and hence a best reply.

16 The equilibrium path is the sequence of moves prescribed by a set of equilibrium strategies in equilibrium.

17 The subgame perfect equilibrium is usually attributed to Reinhard Selten (1965, 1975).

18 A subgame starts with an information set that is a singleton (i.e., consists of a single decision node), includes all the following nodes, and does not cut across any information set (Gibbons 1992:122-123).

19 However, the game also contains a second Nash equilibrium, which is subgame perfect. In this equilibrium, Party j chooses not to comply, while the regime uses its right-hand strategy, meaning that it does not punish Party j for its non-compliance.

20 See also OECD (1998:66).

increase would likely follow. By implication, there would be a conflict of interest between net buyers or net sellers of quotas over whether or not Russia should have its right to trade suspended. Hence, it is far from obvious that the threat of suspension will be credible.[21]

The conditions for compliance to be sustained as a subgame perfect equilibrium in the game in Figure 1 are easily identified by so-called backward induction. The first condition is that it must be rational for the regime to punish Party j, should the latter fail to comply. All things considered, therefore, the regime must derive a net gain from punishing a party that does not comply. In formal terms, this requires that -L-P>-L, i.e., that P is negative. The second condition is that $0 \geq$ T-C, or that $C \geq T$. As we have already seen, the latter requirement is the condition for compliance to be sustained as a Nash equilibrium.

What kinds of punishment are likely to satisfy the criterion of subgame perfection? No attempt to answer this question in full will be made here. Instead, we offer a simple example (involving decentralized enforcement). Suppose the parties were to agree that the Kyoto Protocol be terminated immediately if any one of them commits a (sufficiently serious) violation. In the theory of repeated games, reactions of this kind are typically modelled by the so-called Grim Trigger strategy. Grim Trigger instructs a player to comply at every stage of the game until a defection occurs, and to defect indefinitely from then on.[22] In other words, cooperation stops once and for all if a defection takes place. It is well known that, in the infinitely repeated Prisoners' Dilemma[23] – a model that has been extensively used to study enforcement problems – use of Grim Trigger by both (or all) players is a Nash equilibrium.[24] To see that this Nash equilibrium is subgame perfect, note that after a defection (which never actually occurs in equilibrium), the parties would play the unique Nash equilibrium of the stage game in all subsequent periods. This means that Grim Trigger is a best response against itself (and hence a Nash equilibrium) for the punishment phase, as well as for the entire game. Thus, the threatened punishment is credible in the sense implied by the notion of SPE.

The model of the infinitely repeated Prisoners' Dilemma, then, suggests that an enforcement scheme prescribing termination of the Kyoto Protocol if a (sufficiently serious) violation takes place, thereby leaving each country free to choose its own level of greenhouse gases, is likely to be subgame perfect. If treaty termination means that the parties play a Nash equilibrium in the stage game from then on,[25] then the parties' strategies would be best responses to each other even after a violation has taken place.

---

21 Granted, the Marrakesh accords clearly states that the members of the enforcement branch shall be appointed according to professional merits, not on political grounds. In principle, however, this is also true for a number of other international bodies, even if in practice political considerations play a major role. It remains to be seen what happens in the case of the enforcement branch.

22 An example of an international treaty where use of the Grim Trigger has been formalized is the Convention on Conservation of North Pacific Fur Seals. This treaty stipulates that each of the four parties can give notice of the treaty's termination if consultations on a violation prove unsuccessful (Heister et. al. 1997:28).

23 In a Prisoners' Dilemma each player must choose whether to cooperate (comply) or to defect. Each player maximizes its individual payoff by defecting, regardless of the choices of other players. Yet, all players are better off if all cooperate than if all defect. Hence, individual rationality leads to collective irrationality.

24 This result presupposes that future payoffs are not too heavily discounted. Early references are Friedman (1971) and Shubik (1970).

25 It is a common assumption in the game-theoretic literature on bargaining and international regimes that if an agreement is not reached, or if cooperation breaks down, then the "non-cooperative equilibrium" will be played (e.g., Sprinz & Helm 1999).

Yet, to terminate an entire treaty in response to a single violation hardly sounds plausible. One obvious reason for this is that treaty termination would threaten the global environment. In addition, no party can reasonably be expected to voluntarily join an agreement unless – all things considered – it sees itself better off by doing so. But if a treaty makes all parties better off, then they incur a collective loss if the treaty is brought to an end. Although a threat of treaty termination is likely to satisfy the demands of *individual* rationality implied by the notion of subgame perfection, it would not be *collectively* rational to implement it.

How could this problem be resolved? One idea with some intuitive appeal might be to suspend the agreement temporarily, rather than to terminate it permanently.[26] This would clearly do less damage to the environment. It would also entail a smaller loss to the parties themselves. Still, as punishment it might well be severe enough to outweigh any benefit derived from cheating. A threat of treaty suspension is also credible in the sense implied by the subgame perfect equilibrium, for largely the same reasons that make treaty termination consistent with subgame perfection.

Yet, this proposal does not solve the above problem. Even if a violation were only supposed to cause the treaty to be suspended temporarily, it would still be true that the parties are collectively better off by restarting cooperation at once, than by suffering the delay caused by suspension. Hence, although treaty suspension diminishes the problem created by treaty termination, it does not eliminate it. To do that, we have to look elsewhere. This leads to the notion of renegotiation proofness.

5 Renegotiation proofness[27]

The notion of subgame perfection guarantees that equilibrium behaviour is consistent with the imperatives of individual rationality. In the context of enforcement, however, there is at least one important objection to the notion of subgame perfection. Even if compliance is consistent with subgame perfection, it may be undermined by a collective incentive to renegotiate after a violation has taken place. It may simply be in the interest of all parties to let bygones be bygones, and resume cooperation without further delay. But if a party can foresee this, then a deterrent will not be credible even if it satisfies the criterion of subgame perfection. In turn, this undermines the incentive to comply in the first place. A party may cheat simply because it anticipates that after the fact, an invitation to renegotiate is likely to be accepted. As we saw in the previous section, this is precisely the problem with a system that tries to deter non-compliance via a threat of treaty termination or suspension.

26 A third idea could be to use ”semi-permanent” punishment, where sanctions remain in place until the incumbent leadership leaves office (McGillingray and Smith 2000). However, although this kind of arrangement has considerable intuitive appeal in some contexts (Iraq, Jugoslavia), it seems less attractive in the case of global warming.

27 The literature offers several different notions of renegotiation proofness. The one used here is due to Farrell and Maskin (1989).

Figure 3

		State B				
		B1	**B2**	**B3**	**B4**	**B5**
State A	**A1**	1,1	5,0	0,0	0,0	0,0
	A2	0,5	4,4	0,0	0,0	0,0
	A3	0,0	0,0	3,3	0,0	0,0
	A4	0,0	0,0	0,0	$4, \frac{1}{2}$	0,0
	A5	0,0	0,0	0,0	0,0	$\frac{1}{2}, 4$

To get a formal sense of the problem, consider figure 3.[28] It may be noted, in passing, that the four upper left cells in the matrix form a Prisoners' Dilemma. However, the full game differs from the Prisoners' Dilemma in that it has four Nash equilibria – (A1, B1), (A3, B3), (A4, B4), and (A5, B5).

Now, suppose that this is the stage game for a larger, repeated game, in which the game in Figure 3 is played twice. Call this larger game G. A subgame perfect equilibrium for G is as follows:

1. In the first stage, play (A2,B2)
2. In the second stage, play
   - (A3,B3) if (A2,B2) is played in the first stage;
   - (A1,B1) if at least one state defects (i.e., if any other outcome than (A2,B2) is played) at the first stage.

To see that this is a Nash equilibrium for G, note that, given state B's strategy, playing A2 at the first stage and A3 at the second stage gives A a total payoff of 7. By contrast, any other combination of actions for the two stages gives 6 at best. By symmetry, the same applies to B. To see that this Nash equilibrium is subgame perfect, note that not only are the players' strategies best replies to each other for the second stage if both parties comply at the first stage. They are also best replies to each other for the second stage if one or both of them fail to comply at the first stage. In either case, the parties play a Nash equilibrium for the stage game in the second period.

However, it may be objected to this equilibrium that after a defection, the players would be foolish to settle for the (1,1) outcome in the second stage. Instead, they ought to renegotiate, i.e., go for the (3,3) outcome, as this would mean a higher payoff to both sides. But if a party anticipates that after a violation it will be possible to renegotiate, and thus avoid punishment, then the deterrence effect of the threat to play (1,1) in the second stage is destroyed. Is there any way around this problem?

28 This model has been drawn from Gibbons (1992:87).

The problem is caused by the fact that playing (1,1) in the second stage hurts *both* sides, not only the party that has violated the agreement. To eliminate the problem one needs to design a punishment that the other side has an incentive to implement. Consider the following agreement:

1. At the first stage, play (A2, B2)
2. At the second stage, play
   - (A3, B3) if neither party defects at the first stage;
   - (A3, B3) if both parties defect at the first stage;
   - (A4, B4) if only state B defects at the first stage;
   - (A5, B5) if only state A defects at the first stage;

We now see that it will be impossible for a defector to renegotiate. The reason is that with this enforcement regime, punishment of a non-compliant state is combined with a reward to the other side, meaning that the latter has a positive incentive to resist renegotiation after a violation has taken place.

How can renegotiation proofness be accounted for in a regime on climate change? The short answer is that one need to make it clear to would-be defectors that renegotiation is out of the question. Decision makers must therefore have an incentive to insist that transgressions be punished. Hence, a main lesson for enforcement from the notion of renegotiation proofness is that one ought to avoid deterrents that, if carried out, would be detrimental to all parties. In other words, threatening reversion to a Pareto suboptimal state is unlikely to successfully deter violations. Instead, punishment should take place *on the Pareto frontier*. The question, then, is how this can be achieved in practice.

Barrett (1999) shows that in the infinitely repeated N-person Prisoners' Dilemma (with linear utility functions), renegotiation proofness may be achieved if the parties use a strategy that Barrett refers to as "Getting-even". This strategy instructs player j to cooperate unless j has defected less often in the past than any of the other players. Barrett shows, however, that renegotiation proofness can only be achieved in "small" groups. What counts as "small" depends on the parameters of the model, so that in one context 3 members may be too many, whereas in another context, 150 members may count as small. Barrett also demonstrates that the larger the gains from cooperation, the less likely it is that full cooperation may be sustained as a renegotiation proof equilibrium in his model. To the extent that the gains from an international treaty on climate change count as "large", therefore, the prospects of establishing renegotiation proofness might seem dismal.

However, Barrett only considers decentralized enforcement. Moreover, his model admits only one possible way of punishing a defection, namely through further defections. If transgressors also run the risk of other potential reactions, possibly implemented in a centralized way, the prospects for ensuring renegotiation proofness seem less gloomy. We shall here limit the discussion to two options.

One possibility is to introduce some kind of compensation scheme. For example, if a party to the climate regime is found to have exceeded its allowance for year t, then part of its allowance for the years t+1, t+2,…., t+T might be transferred to other countries. This would

be a penalty for the non-compliant party, but at the same time it would represent a reward to compliant countries. The latter would thus have a positive incentive to insist that the compensation is carried through. Needless to say, the same would be true if compensation is made in monetary terms rather than via transfers of emission allowances.

While compensation schemes can be implemented in a decentralized as well as in a centralized framework, a second possibility requires centralized enforcement. What we have in mind is to decouple entirely the decision to punish defectors from the incentives faced by other parties. In practice, this would probably require a highly legalized enforcement system, where decisions are made independently of political processes. This would minimize the potential impact of political motives to back down on previously issued threats to penalize non-compliant parties. It is worth noting that this solution would not only solve the problem of renegotiation. It would also solve the credibility question raised by the notion of subgame perfection. However, a system of this kind would also provide the regime with supra-national authority. It is far from clear that such a system is politically feasible. UNFCCC (2000: section III, article 42) proposed, as one of two different options, an appeal system that gives the final word to an "appelate body consisting of three members who are recognized authorities in relevant fields". By contrast, the second option was that "an appeal may be made to [the COP/MOP]". Needless to say, these two alternatives could well have very different implications for the credibility of sanctions. From the point of view of renegotiation proofness it was therefore slightly discouraging that in Marrakesh the final word of the appeal system was given to the Meeting of the Parties, which may – with a three-fourths majority decide to override the decision of the enforcement branch.[29] On the other hand, any consequences applied by the enforcement branch will stand pending the decision by the MOP. And since the requirement of a three-fourths majority will probably be fulfilled only in exceptional cases, the influence of political organs seems to be fairly limited, at least on paper.[30]

6 Coalition proofness

While the notions of Nash equilibrium and subgame perfect equilibrium direct our attention to conditions for compliance that follow from individual rationality, the RPE reminds us that effective enforcement also needs to be consistent with collective rationality. Otherwise the parties may want to deviate collectively from previously agreed-upon procedures for punishment, once a defection has taken place.

> A second idea that involves collective rationality is incorporated in the notion of coalition-proof Nash equilibrium. The basic idea underlying this concept is that non-compliance may sometimes be caused by countries acting collectively (i.e., as a coalition), rather than individually. In other words, two or more parties may conspire to take advantage of the others. Clearly, an effective enforcement regime needs to deter conspirations of this sort.

Although the concept of coalition-proof Nash equilibrium is relatively new, the basic idea can be traced back to the fifties, when Robert Aumann (1959) introduced the notion of strong Nash equilibrium. A strong Nash equilibrium is a set of strategies such that no individual player and *no group of players* can get a higher payoff by deviating. Thus, it is not

29 Procedures and Mechanisms relating to compliance under the Kyoto Protocol, section XI.3.

30 Still, it remains to be seen to what extent the selection of members to the enforcement branch will in practice (as well as on paper) be made strictly on the basis of professional merits, rather than for political motives.

enough that each player's strategy is a best reply to the other players' strategies. In addition, coordinated deviations by two or more players acting jointly must be unprofitable as well. It follows that any strong Nash equilibrium must be Pareto efficient, since in any (strictly) Pareto dominated equilibrium the coalition of all players is able to benefit by jointly altering their strategies (Finus 2001:289).

Consider a situation where a large number of countries must decide whether or not to comply with a given treaty, such as the Kyoto Protocol. If no single party can gain by defecting unilaterally, there exists a Nash equilibrium where all countries comply. But this Nash equilibrium is not necessarily strong, since this would also require that no subset of countries is able to achieve a better outcome by defecting jointly, given that all countries not in the subset comply.

The idea that two or more countries might conspire to cheat jointly is certainly an important one. However, the strong Nash equilibrium suffers from a serious weakness. Suppose countries 1,2,…,k would all benefit by defecting jointly from an agreement with N parties (k<N). Would they necessarily join forces? The answer is clearly no, because the resulting conspiracy may not itself be an equilibrium. Given that other members of the subset defect, it may be better still for a given member (or a group of members) to deviate from the conspiracy. Because the strong Nash equilibrium does not distinguish between those deviations that are themselves equilibria and those that are not, the strong Nash equilibrium is not a very plausible solution concept.

To meet this problem, Bernheim et. al. (1987) proposed the concept of coalition-proof Nash equilibrium. Compared to Aumann's concept, the coalition-proof Nash equilibrium is more in line with individual rationality, since only joint deviations that are themselves self-enforcing are seen as potential threats to a treaty's stability.[31] What is new is that any coalition or sub-group is evaluated by the same standards as the full set of actors. This makes the coalition-proof Nash equilibrium a more convincing solution concept when dealing with the possibility of conspiring sub-coalitions. When forming a coalition to cheat jointly, it would be naive not to consider the possibility of further in-group cheating.

One can get a formal sense of the problem by considering the following model, for which a simplified payoff matrix is provided in Figure 4.

31 But there are weaknesses with the coalition-proof Nash equilibrium too, as pointed out by Ehud Kalai (Greenberg 1989:200), among others. The coalition-proof Nash equilibrium only directs attention to the coalition. There must not exist any common deviation within the coalition. But it is not difficult to imagine one or more coalition members joining forces with someone outside the first coalition, and that this new coalition could make all its members better off than they would be if the first coalition was formed. Under these circumstances, it seems counterintuitive to talk about the first coalition as an equilibrium.

Figure 4

		State II	
		Comply	Not comply
State I	Comply	0 0 NE1	$A_1 - P$ $-B$
	Not comply	$-B$ $A_1 - P$	A_2 A_2 NE2

In this model, two states first decide individually and simultaneously whether or not to comply with a climate treaty. Having observed the two states' behaviour, a regime (which is able to perfectly monitor the two states' behaviour) then decides whether to punish one or both of the parties. If both states comply, no punishment is imposed, and each receives a zero payoff.[32] By contrast, if one state defects unilaterally, we assume that the regime will punish it. Hence, the cheating state's payoff is $A_1 - P$, where $A_1 > 0$ is the additional benefit achieved through non-compliance, and $P>0$ is the cost of being punished by the regime. The compliant state then gets a payoff of $-B<0$. Finally, we assume that if neither state complies, then the regime is incapable of imposing any punishment.[33] In this case, each state gets a payoff of A_2.

What is the solution of this game? Provided that $P > A_1$, the threatened punishment is sufficiently severe to deter individual non-compliance, and we have a Nash equilibrium where both states comply. But this is not necessarily the only Nash equilibrium of the game. If $A_2 > -B$, there is also a second Nash equilibrium where neither state complies: Given that state II fails to comply, it is a best reply for state I to reciprocate, and vice versa. Given certain conditions, therefore, there are two Nash equilibria in the game.

Are any of these two Nash equilibria coalition proof? First, consider the Nash equilibrium where both parties comply (NE1). Because this is a Nash equilibrium, then by definition no state can benefit by switching to 'not comply', given that the second state complies. But if the parties *jointly* switch strategies, they both get a payoff of A_2. If $A_2 >0$, it follows that NE1 is not coalition proof. Given the regime's strategy, both states are better off if they agree on NE2 rather than NE1. Conversely, we find that NE2 must be coalition proof, since in this case no joint change of strategies can make both states better off, given the regime's strategy.

From this concept of coalition-proof Nash equilibrium, we arrive at one somewhat pessimistic and one more optimistic implication for the enforcement of climate agreements. On the negative side, the extent to which we can expect an international regime on climate change to satisfy the requirement of coalition proofness is likely to be limited. Most

32 This means that all other payoffs may be interpreted as cost or benefits relative to a situation where both states comply (and the regime does not punish).

33 In short, it is assumed that the following is a dominant strategy for the regime: (1) If both countries comply, do not punish. (2) If one (and only one) party fails to comply, punish that party, but not the other one. (3) If both countries fail to comply, do not punish.

obviously, should the grand coalition of all member states decide to defect collectively from the Kyoto protocol, there is probably little that any realistic enforcement regime would be able to do about it. On the other hand, it is difficult to see what the rationale might be for such a coalition to form in the first place. After all, if the parties consider the treaty so unreasonable as to warrant collective defection, they could simply decide to change it.[34]

On a more positive note, the emerging climate regime has some features that might (unintentionally) help promote coalition proofness. Consider the Clean Development Mechanism (CDM). A potential problem with this mechanism is that two countries might conspire to design a fictitious project, or to design a real project based on false premises (e.g., regarding base lines). At first glance, one might suspect that conspiracies of this sort might be next to impossible to deter, since the procedures and consequences available to the enforcement branch do not apply to developing countries. However, it is enough if the regime is able to deter Annex I countries from conspiracies of this kind. The simple reason is that no CDM project – real or fictitious – is possible without involving an Annex I party.[35]

One might add that the separation of member countries into two categories (Annex I and non-Annex I) is helpful in a second respect as well. In a group of N countries, the number of possible bilateral coalitions is $\frac{N(N-1)}{2}$. By contrast, assume that the countries are divided into two subgroups, industrial countries and developing countries. Let *a* be the proportion of industrial countries, so that the proportion of developing countries is 1-*a*. If only cross-group coalitions are allowed, then the number of potential bilateral coalitions is reduced to $N^2(a-a^2)$. This means that a substantial number of potential coalitions is ruled out when N is large.[36] If at least some of these coalitions would have been able to benefit from collective (equilibrium) defections, it follows that the rule helps promote coalition proofness.

In the game in figure 4, the regime was assumed to be able to perfectly verify the parties' actions. If verification is imperfect, the regime is likely to be even more vulnerable to joint cheating. For example, this can enable a coalition of countries to use smoke-screen tactics to cover up its members' actions. Even if it is verifiable that cheating is taking place, it may be difficult for the regime to identify the guilty party or parties. In turn, this could make implementation of punishment difficult or even impossible. This leads us to enforcement under uncertainty, the topic of the next section.

7 Perfect Bayesian Equilibrium

A main characteristic of the problem of global warming is often said to be the element of uncertainty. In game theory, uncertainty is usually dealt with by introducing asymmetric information.[37] In a game of this type, the players' information sets differ in ways relevant to

34However, this requires that the parties are able to agree on a new design for the treaty, which may not be obvious.

35 Of course, a similar conclusion holds more generally as well: To deter a conspiracy, it suffices to deter a single (pivotal) member of the coalition.

36 To be specific, a total of $\frac{N^2(1-2a+2a^2)-N}{2}$ coalitions are ruled out. This number is maximized for a=0 and a=1, when all coalitions are ruled out. It is minimized for a= $\frac{1}{2}$, when $\frac{N^2-2N}{4}$ coalitions are ruled out. Although a minimum, this is still a substantial number when N is large.

37 There are exceptions. For example, in the rationalist theory of war, it is not uncommon to assume that nature makes the final move, deciding with probability p (0<p<1) that one side prevails, and with probability 1-p that the other side prevails (e.g., Fearon 1994, Gartzke 1999). If p is assumed to be common knowledge (meaning that there is no uncertainty about the parties'

their behaviour. This may be due to uncertainty about a player's *type* (e.g., strategy set and payoff function)[38] or about previous moves in the game. The most commonly used solution concept for this category of games is the perfect Bayesian equilibrium.

A perfect Bayesian equilibrium has two main characteristics. First, the players' strategies are best replies vis-à-vis each other for all subgames, given the parties' beliefs. The reason why beliefs are now important is that in a game of asymmetric information, at least one player has relevant information that is not shared by the other(s). In a game of symmetric information, by contrast, all information relevant to the parties' behaviour is common knowledge. The second characteristic of a perfect Bayesian equilibrium is that, wherever possible, beliefs are updated during the course of play, using Bayes' rule under the assumption of equilibrium behaviour. Using Bayes' rule has been acknowledged by game theorists as the rational way to update beliefs in the light of new information ever since the pioneering work by Harsanyi in the late 1960s (Harsanyi 1967-68).

It is probably fair to say that the lessons implied by the notion of perfect Bayesian equilibrium about compliance and enforcement are generally less clear-cut and therefore less easy to summarize than the lessons referred to in previous sections. Still, some of the insights derived from other equilibrium concepts carry over to games of asymmetric information in a fairly straightforward manner. For example, in a game of asymmetric information a threat need not necessarily be absolutely credible in order to deter non-compliance. In short, Party j need not be convinced that a violation will be detected and punished with *certainty*. It may be enough if it holds this outcome as sufficiently *probable*. If the regime's type is assumed to be common knowledge, by contrast, this simple but important point disappears, since the probability of a threat being carried out is then either zero or one.[39]

While in this simple example, the introduction of asymmetric information adds only marginally to the insights provided by a model of symmetric information, there are a number of phenomena that can only be analysed in a satisfactory way by introducing asymmetric information. Reputation effects provide a good example. Consider the game in Figure 5, which is identical to the one in Figure 1, except that the payoffs have been slightly modified. For the moment, assume that the regime's utilities are common knowledge, while there are two types of parties – "resilient" and "compliant". A resilient party will *never* comply, regardless of the regime's reaction. A compliant party, by contrast, has preferences as shown in Figure 5, meaning that it prefers to comply if it expects the regime to punish non-compliance, but not otherwise. Each party's type is private information. However, it is common knowledge that each party is resilient with probability q, independent of the others. If a compliant party complies, it gets a payoff of 0, while the regime gets a payoff of a>0. If the regime does not punish, a compliant Party j gets a payoff of b>0, while the regime obtains a zero payoff. By contrast, if the regime punishes, both the compliant Party j and the regime get a payoff of -1.

Note that even if Party j does not comply, the regime prefers (in the short run) not to punish. Thus, if the game is played only once, the threat to punish is not credible. The solution of the game, therefore, is that Party j fails to comply (regardless of its type), while the regime does not punish.

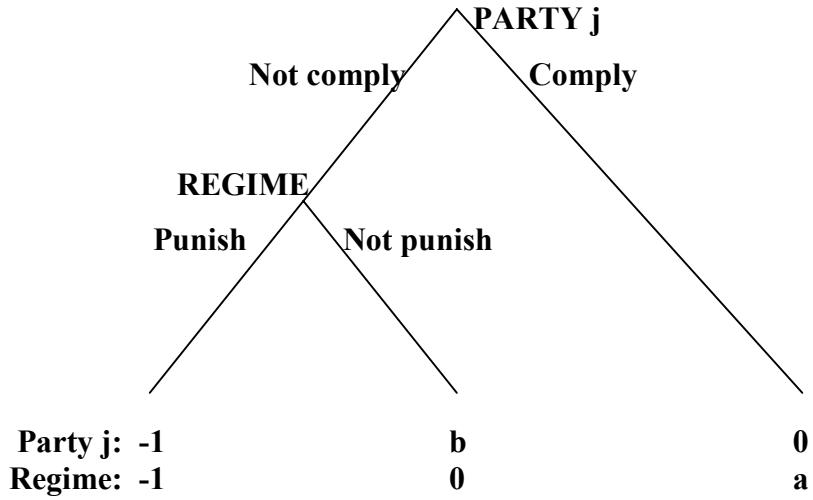
---

relative strength, but still uncertainty about the outcome of a war), one has a model that includes uncertainty, but not asymmetrical information. In Rasmusen's (1989) terms, this is a model with 'uncertain' information.

38 For a more detailed account of the notion of "type", see Fudenberg & Tirole (1994:213-214).

39 Strictly speaking, this conclusion presupposes that there is no instance of indifference between end nodes.

Figure 5



A similar conclusion holds even if the game is repeated a finite number of times. Assume that the regime must play the game in Figure 5 against a total of 100 Parties, one at the time.[40] Each Party participates only once, but is able to observe the outcomes of all previous play. Backward induction reveals that the regime does not punish in the final period, since 0>-1. Given that b>0, this means that Party 1 will not comply, regardless of its type. Since the Parties can foresee the outcome in the final period, the next last period becomes – in a sense – the last. Thus, the same reasoning applies to that period, and to every single period up to the first. The conclusion is that every Party fails to comply, yet the regime does not punish in any period.

This result may seem surprising (to readers without training in game theory). Intuitively, one might expect the regime to punish non-compliance early in the game, in order to deter further non-compliance later on. Why is it that the above model gives a different prediction? The answer lies in a combination of two vital assumptions – that *both* the regime's type *and* the number of iterations is common knowledge. We shall now see that if either of these assumptions is relaxed,[41] then the reputation effect one intuitively expects can prevail.[42]

7.1 Infinite number of iterations, no uncertainty about the regime's type

First, consider the infinite horizon case. Note that regardless of how often the regime punishes, it cannot deter non-compliance by resilient countries. Thus, the expected, periodic payoff of punishing equals (1-q)a+q(-1). For the regime to punish, this

40 A possible objection to this model in the present context is that the parties are assumed to confront the regime in a predetermined order. In a climate change context, it may be more realistic to assume that each party can choose to defect at any point in time.

41 Note that relaxing these assumptions make the model more empirically relevant.

42 The following draws heavily on Fudenberg & Tirole (1994:369-374). See also Kreps & Wilson (1982), Milgrom & Roberts (1982) and Selten (1978).

payoff must exceed the zero payoff which the regime gets if it does not punish. In turn, this requires that the probability q of a given party being resilient satisfies $q < \frac{a}{a+1}$. Note that the threshold value for q is larger, the greater a is. Thus, as one would intuitively expect, the higher the regime values compliance, the more likely it becomes that it punishes non-compliance.

A second condition for effective deterrence is that the regime's threat to punish is credible, meaning that the short-term cost of carrying the threat out must be outweighed by the long-term gain that follows from other countries being deterred from defecting in the future. The latter condition is fulfilled if $-1+\delta\frac{(1-q)a-q}{1-\delta} > 0$, i.e., if (1-q)a-q> $\frac{1-\delta}{\delta}$. In words, this means that for the threat to be credible, the regime's discount factor δ must be sufficiently close to 1.[43]

If both of these conditions are fulfilled, there is a perfect Bayesian equilibrium where all resilient Parties are deterred from non-compliance. In this equilibrium, the regime punishes every instance of non-compliance, as long as it has never refrained from doing so in the past. On the other hand, the regime never punishes again if in the past it has refrained from punishment at least once. A compliant party complies if the regime has never previously declined to punish non-compliance, and does not comply if the regime has failed to punish at least once.

7.2 Finite horizon, uncertainty about the regime's type

Let us now turn to the case with a finite horizon and uncertainty about the regime's type. Suppose that when the game begins, the regime is "strong" with probability p and "weak" with probability 1-p. If strong, the regime *always* punishes a Party that does not comply. If weak, the regime's preferences are as shown in Figure 5. Thus, if the regime ever declines to punish a case of non-compliance, it has revealed that it is weak.

Since a strong regime punishes every instance of non-compliance, the interesting problem is to identify the conditions under which a *weak* regime will punish. The short answer is that this requires a sufficiently high prior probability that the regime is strong (a sufficiently high p). If this condition is fulfilled, compliant countries are deterred from non-compliance. When a defection occurs – this happens (at the latest) the first time the regime confronts a resilient party – it then pays for a weak regime to punish. The reason is simply that this is the only way the regime can uphold its reputation for being strong. Bayes' rule tells us that when a violation occurs, and the regime punishes, the posterior probability that the regime is strong is at least p, given that the regime has never failed to punish non-compliance in the past.

If the horizon is long, even a very small amount of uncertainty (a small p) may suffice to deter non-compliance early in the game. The reason is that the threshold value for p depends on the remaining horizon. It can be shown that with k periods remaining, a compliant party complies as long as $p > \left[\frac{b}{b+1}\right]^k$.[44] Thus, the threshold value for p increases geometrically

43 Note that as δ tends to 1, the right-hand side in this expression tends to zero. Thus, the more emphasis the regime places on future payoffs (other things being equal), the more likely it is that the condition holds.

44 For details, see Fudenberg & Tirole (1991:371-374).

at rate $\frac{b}{b+1}$ as the number of remaining periods shrinks. With a long horizon, therefore, even a (very) small p can have a dramatic influence on the outcome. For example, with b=1 and 10 periods remaining, the threshold value for p is smaller than 0,001.[45]

What can we learn from this model about the enforcement of climate agreements? The model shows that in a repeated game, an element of uncertainty about a player's type can ensure that reputation effects will dominate behaviour. This suggests that, to the extent that one wants to minimize non-compliance, there might be limits to the extent to which the decision making process of the enforcement branch ought to be open and transparent. Again, this is an argument for leaving the decision-making capacity of the enforcement branch to independent experts, rather than to a body of politically appointed representatives of the member countries. The latter variant probably makes it more difficult to preserve an element of uncertainty regarding the willingness of the enforcement branch to punish.

This conclusion is supported by a second implication of the model. Assuming that member countries can observe past play, then under no circumstance can a reputation for being strong make a weak regime worse off. Thus, to the extent that transparency entails a risk that this kind of reputation may be undermined, it is a potential problem for the regime. Note that other recent contributions in the compliance literature hold that maximum clarity about all aspects of the compliance system is a clear-cut advantage (e.g., Mitchell 2001). In contrast, the present model suggests that maximum clarity is not *necessarily* desirable.[46]

In fact, even more extreme secrecy can be advantageous. Compare the game above with a second scenario which is identical in all respects, except that each party is unable to observe the outcome in previous periods. The reason for this could be, for example, that the regime's decisions about non-compliance and penalties are kept secret. It goes without saying that this scenario prevents the regime from building a reputation. Yet, it is possible that the regime might get a higher payoff here than in the first scenario. The explanation is that if all Parties can observe past play, the regime must punish violations by resilient countries in order to uphold its reputation for toughness and thus deter violations by compliant countries. If previous play cannot be observed, by contrast, there is no need to do this. In the latter case, compliant Parties comply as long as $p > \left[\frac{b}{b+1}\right]^k$. Thus, a weak regime is here able to sustain its reputation without having to punish in *any* period. If past play is observable, this kind of behaviour would destroy the regime's reputation the first time it confronts a resilient Party. From then on, not even compliant Parties would comply.

Note that, from an environmental point of view, the two scenarios are equivalent. In either case, the number of countries that fail to comply in equilibrium equals the number of resilient Parties (assuming that $p > \left[\frac{b}{b+1}\right]^k$). The difference is that if past play is observable, then the regime must incur the cost of punishing resilient Parties, even though these countries cannot be deterred. If past play in unobservable, there is no need to do this.

However, there are also circumstances under which the first scenario is preferable from the point of view of the regime. Suppose the regime is strong, in the sense that it prefers

45 To be exact, the threshold value is $\left(\frac{1}{2}\right)^{10} = \frac{1}{1024}$.

46 According to the model, it might be advantageous to retain some uncertainty both about the regime's type (e.g., it's desire to always act strictly according to the rules), and about the process. However, as correctly pointed out by one of the anonymous reviewers of this paper, some transparency reasoning and behaviour of legal institutions appears to be at the heart of their legitimacy. What the above model suggests is that the added legitimacy may come at a cost.

(even in the short term) to punish if a party fails to comply. In this case, punishing a resilient party would imply a net benefit, rather than a net cost for the regime. Moreover, there is no risk of destroying the regime's reputation. In fact, any reputation can only be a disadvantage for the regime, since if the regime is strong, then any uncertainty must involve a positive probability that it might be weak. Thus, secrecy is advantageous only to the extent that the regime incurs a short-term cost when punishing a member party.

In previous sections we noted that there are certain advantages connected to a regime where enforcement is taken care of by legal experts rather than by political delegates. In light of the discussion in the present section we may rephrase these advantages as follows. First, the legal-experts-model makes it more likely that the regime will actually be strong, so that problems of subgame perfection and renegotiation are overcome. Second, if the regime is weak, the legal expert model tends to make it easier to exploit uncertainty to build a reputation for being strong, thus creating the necessary deterrence effect.

However, to take advantage of uncertainty requires a minimum of secrecy. Needless to say, this is not unproblematic. In particular, secrecy is not usually considered to be consistent with the requirements of due process. Thus, it comes as no big surprise that the procedures for the enforcement branch laid down in the Marrakesh Accords emphasize that hearings should normally be conducted in public, and that other parties as well as the public should be informed about findings and decisions.[47] It may also be noted that the adoption of a clear-cut legal expert model would mean that the parties would surrender a considerable amount of autonomy to the regime. It is therefore in many ways remarkable that the Marrakesh accords retain so little influence in matters of non-compliance at the political level. In principle, at least, this happens primarily via the appeal system, and even there it is arguably not very important.

8 Conclusions

This paper has explored a number of lessons from noncooperative game theory that is relevant for effective enforcement of existing and future climate agreements. Together, these lessons suggest that constructing a system for "hard" enforcement that effectively deters non-compliance is a formidable task, which has only partially been accomplished by the compliance mechanism agreed upon in Marrakesh. The main lessons identified in the previous discussion may be summarized as follows:

- The notion of Nash equilibrium reminds us that relatively severe consequences may be needed to deter non-compliance. In particular, the penalty level should be set at a more than proportional rate.
- The subgame perfect equilibrium teaches us that consequences not only need to be severe, but also individually rational to implement, should a transgression take place.
- The renegotiation proof equilibrium tells us that it is must be collectively rational to impose punishment on a party found to be in non-compliance. Such punishment should therefore take place on the Pareto frontier, rather than by reversion to some suboptimal state.
- The coalition proof equilibrium suggests that a regime on climate change needs to curb not only individual, but also collective (subgroup) incentives to cheat. The flexible

47 For example, see Procedures and Mechanisms relating to compliance under the Kyoto Protocol, sections IX.2 and IX.6.

mechanisms provide some possibilities for, but also some constraints on, potentially profitable collective cheating.

- The notion of perfect Bayesian equilibrium implies that private information may be exploited by the regime to deter non-compliance. In particular, a fully transparent enforcement regime could turn out to be detrimental for compliance levels.

A compliance mechanism that takes into account all of the above lessons would likely come close to minimizing (intentional) non-compliance. However, the compliance mechanism of the climate regime is the endproduct of a complex process, where the need to deter non-compliance has been carefully balanced against a number of other considerations, including a desire to satisfy the requirements of due process. Thus, the fact that the Marrakesh Accords do not incorporate all of the above lessons from non-cooperative game theory is not unconditionally a weakness.

References

Adger, W.M. 1995. "Compliance with the Climate Change Convention", Atmospheric Environment 29:1905-1915.

Aumann, R. (1959), "Acceptable Points in General Cooperative N-Person Games", in A.W. Tucker et al., Contributions to the Theory of Games IV. Princeton: Princeton University Press.

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

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

Barrett, S. (1999), "A Theory of Full International Cooperation", Journal of Theoretical Politics 11, 519-541.

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

Bodansky, D. (2001), "International Law and the Design of a Climate Change Regime", in U. Luterbacher and D.F. Sprinz (eds.), International Relations and Global Climate Change. Cambridge, Mass.: MIT Press.

Carraro, C. & D. Siniscalco (1993), "Strategies for the International Protection of the Environment", Journal of Public Economics 52:309-328.

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

Chayes, A & A.H. Chayes (1995), The New Sovereignty. Compliance with International Regulatory Agreements. Cambridge, Mass.: Harvard University Press.

COP-6 (2000), "Note by the President of COP6, 23 November 2000".

Downs, G.W., D.M. Rocke & P.N. Barsoom (1996), "Is the Good News about Compliance Good News about Cooperation?", International Organization 50, 379ff.

Farrell, J. & E. Maskin (1989). "Renegotiation in Repeated Games", Games and Economic Behaviour 1, 327-360.

Fearon, J. (1995), "Rationalist Explanations for War", International Organization 49:379-414.

Finus, M. 2001. Game Theory and International Environmental Cooperation. Cheltenham: Edward Elgar.

Friedman, J. 1971. "A Non-Cooperative Equilibrium for Supergames", Review of Economic Studies 38:1-12.

Fudenberg, D. & J. Tirole 1994. Game Theory. Cambridge, Mass.: MIT Press.

Gartzke, E. 1999. "War is in the Error Term", International Organization 53:567-587.

Gibbons, R. 1992. A Primer in Game Theory. New York: Harvester Wheatsheaf.

Greenberg, J. 1989. "Deriving Strong and Coalition-Proof Nash Equilibria from Abstract system", Journal of Economic Theory 49:195-202.

Grubb, M., C. Vrolijk & D. Brack 1999. The Kyoto Protocol. A Guide and Assessment. London: The Royal Institute of International Affairs.

Gul, Faruk (1997), "A Nobel Prize for Game Theorists: The Contributions of Harsanyi, Nash and Selten", Journal of Economic Perspectives, 11(3), 159–174.

Hagem,C., O. Mæstad & H. Westskog (forthcoming). "Effective Enforcement and Imprecise Deterrents: Impacts of Punishment on Punishers via the Markets for Permits and Energy", in J. Hovi, O.S. Stokke & G. Ulfstein, eds., Compliance with Climate Commitments: Conditions and Mechanisms.

Hargrave, T., N. Helme, S. Kerr, and T. Denne (1999), Defining Kyoto Protocol Non-Compliance Procedures and Mechanisms. Leiden: Center for Clean Air Policy.

Harsanyi, J.C. (1967-68), "Games With Incomplete Information Played by 'Bayesian' Players. Part I: The Basic Model", Management Science 14, 159-182; "Part II: Bayesian Equilibrium Points", Management Science 14, 320-334; "Part III: The Basic Probability Distribution of the Game", Management Science 14, 486-502.

Heister, J., E. Mohr, F. Stähler, P.-T. Stoll and R. Wolfrum (1997), "Strategies to Enforce Compliance with an International CO Treaty", International Environmental Affairs 9, 22-53.

Hovi, J. (1998), Games, Threats and Treaties. Understanding Commitments in International Relations. London: Pinter.

IISD (2000), "Summary of the Workshop on Compliance Under the Kyoto Protocol", Earth Negotiation Bulletin 12, No. 124. http://www.iisd.ca/vol12/enb12124e.html

Keohane, R.O. (1984), After Hegemony. Princeton, N.J.: Princeton University Press.

Kerr, S. (1998), "Enforcing Compliance: The Allocation of liability in International GHG Emissions Trading and the Clean Development Mechanism", RFF Climate Issue Brief #15 (Internet edition). http://www.rff.org/issue_briefs/PDF_files/ccbrf15.pdf

Kreps, D. & R. Wilson (1982), "Reputation and Imperfect Information", Journal of Economic Theory 27, 253-279.

McGillingray, F. & A Smith (2000), "Trust and Cooperation Through Agent-specific Punishments", International Organization 54, 809-824.

Milgrom, P. & J. Roberts (1982), "Predation, Reputation and Entry Deterrence", Journal of Economic Theory 27, 280-312.

Mitchell, R. (2001). "Institutional Aspects of Implementation, Compliance, and Effectiveness", ch. 11 in U. Luterbacher and D.F. Sprinz (eds.), International Relations and Global Climate Change. Cambridge, Mass.: MIT Press.

Nash, J. F. (1951), "Non-cooperative Games", Annals of Mathematics 54, 286-95.

Oberthür, S. & H.E. Ott (1999), The Kyoto Protocol. International Climate Policy for the 21st Century. Berlin: Springer Verlag.

OECD (1998), Ensuring Compliance with a Global Climate Change Agreement. Paris: OECD Information Paper.

Oye, K.A., ed. (1986), Cooperation under Anarchy. Princeton, N.J.: Princeton University Press.

Pew Center on Global Climate Change (2001), "Climate Talks in Bonn: News and Information". http://www.pewclimate.or/bonn/daily.cfm?m

Rasmusen, E. (1989), Games and Information. An Introduction to Game Theory. Oxford: Basil Blackwell.

Selten, R. (1965), "Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit", Zeitschrift für die gesamte Staatswissenschaft 12, 301-24.

Selten, R. (1975), "Re-Examination of the Perfectness Concept for Equilibrium Points in Extensive Games", International Journal of Game Theory 4, 25-55.

Shubik, M. (1970), "Game Theory, Behavior, and the Paradox of the Prisoner's Dilemma", Journal of Conflict Resolution 14, 181-193.

Sprinz, D. & C. Helm (1999), "The Effect of Global Environmental Regimes: A Measurement Concept", International Political Science Review 20, 359-369.

Stranlund, J.K. (1995), "Public Mechanisms to Support Compliance to an Environmental Norm", Journal of Environmental Economics and Management 28, 205-222.

Taylor, M. (1976), Anarchy and Cooperation. London: John Wiley.

Taylor, M. (1987), The Possibility of Cooperation. Cambridge: Cambridge University Press.

UNFCCC (2000), "Procedures and Mechanisms Relating to Compliance under the Kyoto Protocol." Text proposed by the Co-Chairmen of the Joint Working Group on Compliance to the COP-6.

van Damme, E. (1991), Stability and Perfection of Nash Equilibria. Berlin: Springer-Verlag.

Werksman, J. (1998), Responding to Non-Compliance under the Climate Change Regime. Paris: OECD Information Paper.