ISSN 1835-9728

Environmental Economics Research Hub

Research Reports

Climate Change and Game Theory

Peter John Wood

Research Report No. 62

May 2010

About the authors

Peter Wood is a Postdoctoral Fellow in Resource Management in Asia-Pacific Program Research School of Pacific and Asian Studies, College of Asia and Pacific, Australian National University

Email: Peter.J.Wood@anu.edu.au +61 2 6125 6284

Environmental Economics Research Hub Research Reports are published by The Crawford School of Economics and Government, Australian National University, Canberra 0200 Australia.

These Reports present work in progress being undertaken by project teams within the Environmental Economics Research Hub (EERH). The EERH is funded by the Department of Environment and Water Heritage and the Arts under the Commonwealth Environment Research Facility.

The views and interpretations expressed in these Reports are those of the author(s) and should not be attributed to any organisation associated with the EERH.

Because these reports present the results of work in progress, they should not be reproduced in part or in whole without the authorisation of the EERH Director, Professor Jeff Bennett (jeff.bennett@anu.edu.au)

Crawford School of Economics and Government
THE AUSTRALIAN NATIONAL UNIVERSITY

http://www.crawford.anu.edu.au

Climate Change and Game Theory: a Mathematical Survey

Peter John Wood [*]

May 5, 2010

Abstract

This survey paper examines the problem of achieving global cooperation to reduce greenhouse gas emissions. Contributions to this problem are reviewed from non-cooperative game theory, cooperative game theory, and implementation theory.

Solutions to games where players have a continuous choice about how much to pollute, games where players make decisions about treaty participation, and games where players make decisions about treaty ratification, are examined. The implications of linking cooperation on climate change with cooperation on other issues, such as trade, is examined. Cooperative and non-cooperative approaches to coalition formation are investigated in order to examine the behaviour of coalitions cooperating on climate change.

One way to achieve cooperation is to design a game, known as a mechanism, whose equilibrium corresponds to an optimal outcome. This paper examines some mechanisms that are based on conditional commitments, and could lead to substantial cooperation.

Key Words and Phrases. Climate change negotiations; game theory; implementation theory; coalition formation; subgame perfect equilibrium.

[*] Dr Peter John Wood, Resource Management in Asia-Pacific Program, The Crawford School of Economics and Government, The Australian National University, Canberra, ACT 0200, Australia, Email: Peter.J.Wood@anu.edu.au

Contents

1 Introduction

A key reason why achieving international cooperation to address climate change is difficult is that there are strong free-rider incentives. These incentives arise because climate change mitigation is a global public good – everyone benefits from there being less global warming, and everyone has an incentive for someone else to take on the burden of emission reductions. This is compounded by the fact that because of sovereignty issues, international institutions are weak compared to national ones. Game theory, which analyses the mathematics of strategic behaviour, can help us obtain a better understanding of how the incentive to free-ride works, identify the potential barriers to cooperation, and find approaches to facilitate a cooperative outcome. This paper surveys the game theoretic literature that relates to climate change, with an emphasis on approaches that try to find ways to facilitate cooperation. Our work complements previous surveys (Finus, 2001; Barrett, 2003) because it is more recent, and is also shorter.

Game theory is often applied by assuming that the game is given, and used to predict the behaviour of participants. But an area of game theory known as implementation theory treats the desired outcome as given, and asks how to design a process that leads to this outcome (Jackson, 2001). An example of such as process could be the negotiations for an international environmental agreement. This approach may help us design processes that are more likely to lead to cooperative outcomes.

Addressing the free-rider incentives associated with climate change mitigation requires that we find mechanisms to facilitate cooperation between states. One such approach is international treaty-making. In 1992, countries negotiated the *United Nations Framework Convention on Climate Change* (UNFCCC). Since then, countries have negotiated the *Kyoto Protocol* to the UNFCCC, which included emissions reduction commitments for some developed countries, known as Annex I countries. Countries have been engaging in further negotiations at *conferences of parties* to the UNFCCC and *meetings of parties* to the Kyoto Protocol.

The difficulties with finding cooperation were illustrated in the 2009 Copenhagen climate negotiations, which resulted in a political accord, but where after years of negotiations there remained too much disagreement between nations to arrive at a binding international treaty. There has been an ongoing political debate about the role of the United Nations, and whether more could be achieved in negotiations involving smaller groups of countries. The Copenhagen negotiations may have made the latter more likely. The lead US climate change negotiator, Todd Stern, stated: “You cant negotiate in a group of 192 countries. Its ridiculous to think that you could” (Little, 2010). Nicholas Stern has offered a different perspective, stating that “The fact of Copenhagen and the setting of the deadline two years previously at Bali did concentrate minds, and it did lead…to quite specific plans from countries that hadn’t set them out before”, and that it was vital to stick with the UN process, whatever its frustrations (Black, 2010).

Game theory can provide useful insights when considering debates such as these. In fact, there has been a parallel debate in the game theory literature (see Section 3) on whether cooperation is more likely to arise from a 'grand coalition' of all countries, or from smaller coalitions. Game theory provides insight both into the stability of coalitions, and the implications of different processes for forming coalitions.

Game theory is relevant to questions involving participation, compliance, and enforcement of international climate agreements. Because international institutions are weak, it is difficult to enforce an agreed outcome. One issue that has been raised is the role that the threat of trade restrictions could play in prevention of free-riding.

When using a model to help understand a problem, it is important to be aware of the limitations of the model. Many applications of game theory require that decision makers are rational. That is, they have clear preferences, form expectations about unknowns, and make decisions that are consistent with these preferences and expectations. These assumptions may not be consistent with experimental psychology. Occasionally it is possible to devise experiments which can test these assumptions in the context of a game. This is done with the ultimatum game, which we describe in Example 2.3. The results suggest that human actions depend on concepts of fairness and reciprocity as well as purely rational and strategic considerations.

Ostrom (2009) has considered the the role that human behaviour considerations relate to cooperation problems, and applied this to climate change. She found that a 'surprisingly large number of individuals facing collective action problems do cooperate'. She also found that cooperation is more likely if people gain reputations for being trustworthy reciprocators; reliable information is available about costs and benefits of action; individuals have a long-term time horizon; and are not in a highly competitive environment.

In many of the situations that we describe here, countries are assumed to be the players in the game. That is, they are assumed to have clear preferences, usually based on the aggregate welfare of the countries citizens. In reality, different citizens have greatly different preferences, and the decision making is based on a political process. An example of this is the process of treaty ratification, which we will discuss in Example 2.4.

Despite the above limitations of our methodology, many of the mechanisms described here are important because their game theoretic solutions are cooperative. If humans are more cooperative than predicted by game theory, these mechanisms may still lead to cooperative outcomes. Mechanisms that are expected to lead to cooperation can also be further tested using behavioural experiments. The game theoretic investigation of cooperative mechanisms may ultimately facilitate cooperation.

In Section 2, we introduce games where players make decisions independently. We discuss both the normal form representation of a game and the extensive form representation of a game. We investigate the role of solution concepts including the Nash equilibrium and the subgame perfect equilibrium. We study examples such as the prisoner's dilemma, repeated prisoner's dilemmas, the ultimatum game, games of treaty participation, games of treaty ratification, and game theoretic approaches to bargaining. We investigate a basic framework for studying what happens when countries have a continuous choice about how much they reduce their emissions.

In Section 3, we examine situations where players can cooperate with each other and form coalitions, which may then behave non-cooperatively when interacting with other coalitions. We introduce transferrable utility games, in which players within a coalition can make payments to each other. We examine the concept of the core, where a grand coalition that includes all players does form, and no smaller coalitions have an incentive to break away. We discuss an interesting result, due to Chander & Tulkens (1997), that if a grand coalition for reducing emissions was to dissolve into singletons when any coalition breaks away, then full cooperation is possible. We also discuss non-cooperative mechanisms for a coalition formation, and apply this to the question of whether cooperation is more likely among a grand coalition, or among several smaller coalitions.

In Section 4, we look at applications of implementation theory to climate change. We examine mechanisms for getting players to agree to a socially optimal outcome, including one which also induces them to reveal their mitigation costs. We also look at some mechanisms for providing public goods when there may not be strong institutions, and how that relates to emissions reductions in an international context. Section 5 concludes.

2 Non-cooperative Games and Climate Change

In non-cooperative games, players make decisions independently. We define some of the relevant ways of representing non-cooperative games and solution concepts. We illustrate these definitions with a number of examples that are relevant to climate change.

2.1 Normal Form Games and the Nash Equilibrium

Definition 2.1 The *normal form representation* of a game specifies

1. the set of *players* in the game (in the context of climate change these will often be *countries*), N;

2. a set S of *strategy combinations*, each strategy combination assigns a strategy to each player;

3. and the set of payoffs $\Pi = \{\pi_i : i \in N\}$ received by each player for each possible strategy combination. Each payoff π_i assigns a real number (the utility[1]) to a strategy combination.

[1] It is possible to define strategic games more generally in terms of a preference relation for each player on the set of strategy combinations (Osborne, 2003, Chapter 2). It follows from ordinal utility theory that if a preference relation satisfies certain axioms, then it is representable by a utility function (Berger,

The normal form representation of a game is sometimes also known as the strategic form of a game.

When we consider a player i and strategy combination s, we will often write s_{-i} to denote the strategies of players other than i, and write $s = (s_i, s_{-i})$.

Definition 2.2 A *Nash equilibrium* for a normal form representation of a game is a strategy combination $s^* = (s_i^*, s_{-i}^*)$ where for all players $i \in N$, we have that

$$\pi_i(s_i^*, s_{-i}^*) \geq \pi_i(s_i, s_{-i}^*). \tag{1}$$

In other words, in the Nash equilibrium every strategy is the best response to the best strategies of the other players.

An important variation of the concept of a normal form game allows players to play *mixed strategies.* Instead of choosing a particular strategy, each player assigns a probability to each strategy (Osborne & Rubinstein, 1994, Chapter 3).

Example 2.1 (The Prisoner's Dilemma). The problem of achieving cooperation to reduce greenhouse gas emissions is related to the *prisoner's dilemma.* All countries are collectively better off if they reduce their emissions, but each country is individually better off if they continue to pollute. We shall now describe a two player prisoner's dilemma. Each player has two possible strategies $\{Pollute,\ Abate\}$. We represent the payoffs by using the *payoff matrix* notation. The two rows correspond to the two possible actions of the first player; the two columns correspond to the possible actions of the second player; the numbers in each box correspond to the payoffs for each player, with the payoff for the first player listed first.

$$\begin{array}{cccc} & & \multicolumn{2}{c}{\text{Player 2}} \\ & & \textit{Abate} & \textit{Pollute} \\ \text{Player 1} & \textit{Abate} & (10,10) & (0,11) \\ & \textit{Pollute} & (11,0) & (1,1) \end{array} . \tag{2}$$

The strategy pair $(Pollute, Pollute)$ is a Nash equilibrium because given that the second player chooses $Pollute$, the first player is better off choosing $Pollute$ than choosing $Abate$, and vice-versa. None of the other strategy combinations are Nash equilibira because in each case at least one player can improve their payoff by changing their strategy. The

1980, Chapter 2), (Ok, 2007, Section B.4).

Assessing the impact on utility of climate change is complicated by several factors (Garnaut, 2008a), (Stern, 2006): the damages are uncertain, so players are interested in impact on *expected utility*; damages can include impacts on non-market goods such as ecosystems; many impacts occur in the future and affect future generations, so their valuation depends on a discount rate that is likely to take into account the pure rate of time preference, the marginal elasticity of consumption, and the expected rate of economic growth; and depends on the risk aversion of the player. Because unmitigated climate change presents potentially catastrophic risks, the possible impact of highly damaging outcomes can dominate the expected damage function (Weitzman, 2009).

strategy pair $(Abate, Abate)$ is known as the *social optimum*, because the collective payoff (the sum of each player's payoff) is maximised. For this example the Nash equilibrium has a much lower collective payoff than the social optimum.

Climate change is similar to a prisoner's dilemma, but countries don't just make a decision about whether to pollute or not, they make a decision about how much to reduce their emissions. The following example models this situation.

Example 2.2 (The Global Emissions Game with Continuous Strategy Space). This example is based on (Finus, 2001, Chapter 9), a more general version of this game that applies to transboundary pollutants that are not necessarily global pollutants is described in Finus (2003). This game has a continuous strategy space in that each player chooses how much pollution to emit, rather than whether to pollute or not. This game describes a global pollutant, in that the the damages from the pollutant on each player depend on the total amount of pollution emitted by all of the players. This game could apply to greenhouse pollution, and also to pollutants that affect the ozone layer. This game does not examine the dynamic aspects of pollution.

Players can be thought of as countries. We assume that the set of players, N, has size n. Let e_i be the emissions from country i. The utility π_i of country i is given by

$$\pi_i = \beta_i(e_i) - \phi_i\Big(\sum_{j\in N} e_j\Big) \tag{3}$$

where β_i are the emissions benefit functions and have the property that the derivative is strictly positive ($\beta_i' > 0$) and the second derivative is not positive ($\beta_i'' \leq 0$); ϕ_i are the emissions damage functions and we assume that their derivatives is strictly positive ($\phi_i' > 0$) and the second derivative is non-negative ($\phi_i'' \geq 0$). In other words, the marginal benefits from emissions decrease with emissions, but the marginal damages from emissions increase.

To calculate the Nash equilibrium, we first work out what the best response for country i is if the emissions for all of the other countries are given. This is done by differentiating (3) with respect to e_i. The first order conditions

$$\frac{\partial \pi_i}{\partial e_i} = 0 \tag{4}$$

imply that

$$\beta_i'(e_i) = \phi_i'\Big(\sum_{j\in N} e_j\Big). \tag{5}$$

By taking the total derivative of (5) and applying the implicit function theorem, it is possible to show (see (Finus, 2001, p. 126) or (Finus, 2003, Appendix 2)) that e_i can be expressed as a function of the emissions of the other countries. We call this function the *best reply function*, and we write $e_i = r_i(e_{-i})$ where e_{-i} is the emissions from countries other than i. It also follows that for $j \neq i$,

$$\frac{dr_i}{de_j} = \frac{\phi_i''}{\beta_i'' - \phi_i''}. \tag{6}$$

It is interesting to note that because ϕ_i'' is non-negative and β_i'' is not positive, it follows that (6) implies that if some country j reduces its emissions compared to the Nash equilibrium, then country i's best reply is to increase its emissions. This is because if j reduces their emissions, the total damages are lower, so the marginal damage function ϕ' is not as steep.

The Nash equilibrium can be obtained by substituting the best reply functions into each other and solving for the remaining variable. Suppose that the emission benefit functions are given by

$$\beta_i(e_i) = b(de_i - \frac{1}{2}e_i^2), \tag{7}$$

and the emission damage functions are given by

$$\phi_i(e_i) = \frac{c}{2}\Big(\sum_{j \in N} e_j\Big)^2. \tag{8}$$

Then the Nash equilibrium emissions are given by

$$e_i^* = \frac{bd}{b + 2c}. \tag{9}$$

If there were no damages from emissions (so that $c = 0$) then the Nash equilibrium would be $e_i^* = d$. The social optimum is given by $e_i = bd/(b+4c)$. So the Nash equilibrium does involve some emission reductions, but less than optimal emission reductions.

Situations where the non-cooperative outcome is sub-optimal are known as *social dilemmas*. The above situation assumes that all participants have complete information about the payoffs for each other; assumes that decisions are made independently; does not take into account communication between the participants; and does not consider how a central authority could enforce agreements among participants about their choices. If these assumptions are not true, then it is much less certain that a suboptimal non-cooperative outcome will occur (Ostrom, 2009). When there is communication, decisions are not made independently, and participants can make enforceable agreement, cooperation may be more likely. But if participants do not have complete information, cooperation may become more difficult, because players could have an incentive to misrepresent their preferences.

2.2 Extensive Form Games and the Subgame Perfect Equilibrium

The normal form representation of a game hides the sequential nature of strategy and decision making. By contrast, extensive form games study the sequential nature of games explicitly. An extensive form game represents the game as a tree. At each node of the tree, except for terminal nodes, one of the players makes a decision that determines which node is reached next. Terminal nodes determine the payoffs of the game.

Definition 2.3 An *extensive form game with perfect information* (Osborne, 2003, Chapter 5) specifies

1. the players N in the game;
2. a set of sequences of nodes in the game (*terminal histories*) with the property that no terminal history is a proper subsequence of any other terminal history;
3. a function (known as the *player function*) that assigns a player to any sequence h that is a proper subsequence of a terminal history – the player function can be thought of as specifying the player whose turn it is after h;
4. the payoffs for each player at each possible end node.

Given a history h, the set of all *actions* available to the player who moves after h is

$$A(h) = \{a \ : \ (h, a) \text{ is a history}\}.$$

A *strategy* of a player i in an extensive game with perfect information is a function that assigns an action in $A(h)$ to each history h after which it is a player i's turn to move. A strategy combination s determines a terminal history $O(s)$, known as the *outcome* of s. Associated with an extensive form game is a normal form representation that we will call the *strategic form* of the extensive form game. The strategic form has the same players and strategy combinations as the extensive form game, and the payoffs are given by the payoffs at the end nodes of each outcome of the extensive form game. If the longest terminal history of a game is finite, then we say that it has *finite horizon.*

The strategy combination s^* in an extensive game with perfect information is a *Nash equilibrium* if for every player $i \in N$ and strategy s_i,

$$\pi_i(O(s^*)) \geq \pi_i(O(s_i, s^*_{-i})). \tag{10}$$

The Nash equilibrium of an extensive form game is the Nash equilibrium of its strategic form.

Let Γ be an extensive form game with perfect information and player function P. For any non-terminal history h of Γ, the *subgame* $\Gamma(h)$ following the history h is the following extensive game:

1. the players are the same as those for Γ;
2. the terminal histories are sequences h' such that (h, h') is a terminal history of Γ;

3. the player $P(h, h')$ is assigned to the proper subhistory h' of the terminal history (h, h');

4. the payoff in $\Gamma(h)$ associated with h' is equal to the payoff in Γ associated with (h, h').

Definition 2.4 A *subgame perfect equilibrium* is a strategy combination constituting a Nash equilibrium in every subgame of the entire game. Equivalently, for every player $i \in N$, every history h after which it is player i's turn to move, and strategy s_i,

$$\pi_i(O_h(s^*)) \geq \pi_i(O_h(s_i, s^*_{-i})) \tag{11}$$

where $O_h(s)$ is the terminal history consisting of h followed by the actions generated by playing strategy s after h.

We will make extensive use of the subgame perfect equilibrium.

Example 2.3 (The Ultimatum Game). In the ultimatum game, there are two players and a sum of money. The first player proposes how to divide up the sum of money, and the second player chooses whether to accept or reject the proposal. If the second player rejects the proposal, neither player receives anything.

Assume that there is a smallest division of the sum of money available (1 cent say), that we denote by ε. Assume that the total amount of money available is equal to 1 (\$1 say) and that 1 is an integer multiple of ε. The ultimatum game can be represented by an extensive form game with two stages. In the first stage the first player chooses an amount of money $x \in [0, 1]$ which is also an integer multiple of ε. In the second stage the second player chooses whether to accept the offer or not. If the second player accepts, the payoffs are $(1 - x, x)$; if not, the payoffs are $(0, 0)$.

Because the ultimatum game has finite horizon, it is possible to find the subgame perfect equilibrium using a technique known as *backwards induction.* In this technique the subgame perfect equilibria for the 'last' subgames are calculated first. Then taking these actions as given, we calculate the equilibria for preceding subgames and so on. For the ultimatum game, we first consider the subgames where the second player either accepts or rejects an offer from the first player. For any offer $x > 0$, the second player's optimal response is to accept the offer. In the subgame when the offer is $x = 0$, the second player is indifferent about whether to accept or not. There are therefore two equilibrium strategies for the second player. Either to all accept all payoffs (including $x = 0$), or to accept all payoffs except for $x = 0$.

Let us now consider the subgame perfect equilibrium strategy for the first player. There are two possibilities:

- If the second player accepts all offers, the first player's optimal strategy is to make the offer $x = 0$, and then recieve the payoff 1.

- If the second player accepts all offers except $x = 0$, the first player's optimal strategy is to make the offer $x = \varepsilon$, and recieve the payoff $1 - \varepsilon$.[2]

Both of the above possibilities are subgame perfect equilibria, but unless the first player is certain that the second player will accept all offers including $x = 0$, they are better off making the offer $x = \varepsilon$.

Let us now characterise the Nash equilibria of the ultimatum game. The first player chooses an amount x in the unit interval $[0, 1]$ that is a multiple of ε. The second player chooses a function

$$f : [0, 1] \mapsto \{\mathit{Accept}, \mathit{Reject}\}.$$

A strategy combination (x, f) is a Nash equilibrium if $f(x) = \mathit{Accept}$ and there is no $y < x$ such that $f(y) = \mathit{Accept}$. The first player would not want to decrease their offer because the second would reject it; the second would not want to reject the offer because then they would get nothing. Another Nash equilibrium is the combination $x = 0$, and $f(x) = \mathit{Reject}$ for all x. So any possible off could be a Nash equilibrium. In this sense the Nash equilibrium is a weaker concept than the subgame perfect equilibrium.

Experiments where people have played the ultimatum game have consistently found that the first player will usually offer significantly more money to the other player than the subgame perfect equilibrium, and the second player will be unlikely to accept the offer if they are offered less than 30 per cent of the total amount (Güth *et al.* , 1982).

It has been argued by Fehr & Gächter (2000) that the ultimatum game provides evidence that economic agents don't just base their decisions on pure self interest, and reciprocal considerations play an important role in people's actions. It has been argued (Barrett, 2003, pp. 299–301) that the ultimatum game also provides evidence that an international environmental agreement is more likely to be stable if it is perceived by its parties to be fair.

Equity considerations play an important role in many proposals for how greenhouse gas emissions should be allocated in a post-Kyoto protocol. They are one of the reasons why many proposals (such as Baer *et al.* (2000); Garnaut (2008b); Meyer (2000)) suggest that all countries should eventually be allocated the same amount of per-capita emissions. A shorter transition to equal per-capita emissions would be fairer than a longer transition because a longer transition rewards high per-capita emitters for having high per-capita emissions. But even a very short transition to equal per-capita emissions could be considered to be unfair because different countries have different historical emissions. Stern (2009) states (p. 153) that

> To suggest that we should all be entitled to emit roughly equal amounts by 2050 is to say that, at the end of the drinking spree, we should be using glasses

[2] If there was no smallest division of the sum of money, then no offer $x > 0$ would be optimal, because $x/2$ would be better. In this case the only subgame perfect equilibrium corresponds to the offer $x = 0$.

> of the same size. It is difficult to see this as a particularly equitable division of the entitlements to the reservoir, since this type of equality takes no account of all the 'drinking' that has gone on over the previous two hundred years.

One alternative approach is for a global total emissions budget that takes into account historical emissions (Pan *et al.* , 2000; Project Team of the Development Research Centre of the State Council, 2009). This approach could be considered to be equitable because no country imposes an external cost on any other. It would also mean that many countries would have already used up significantly more than their emissions budget. Both approaches would be unlikely to be acceptable to many developed countries. They would be unlikely to ratify a treaty if it is based on one of these approaches. We discuss the issue of treaty ratification in Example 2.4.

Example 2.4 (Treaty Ratification). After an international treaty is negotiated, it then has to be ratified by its participants. This can be modelled as a two stage game. In Stage 1, the players negotiate the treaty; in Stage 2, each country decides whether to ratify the treaty. For some countries, for example the United States, ratification can be difficult. The United States requires 67 out of 100 Senate votes in order to ratify a treaty. By backwards induction, for negotiators in Stage 1 to play the subgame perfect equilibrium, they will take into account that a treaty will have to be sufficiently aligned with the domestic interests of the United States, in order for it to be ratified by the United States (Barrett, 2003, p. 148).

It is possible to modify Definition 2.3 so that players can make simultaneous moves. Instead of having the player function assign a player to a subhistory, it assigns a set of players. The game also needs to be consistent – the actions corresponding to a subhistory is the same as the actions of the players assigned by the player function to that subhistory. The formal definition is as follows:

Definition 2.5 An *extensive form game with perfect information and simultaneous moves* specifies

1. the players N in the game;

2. a set of sequences of nodes in the game (*terminal histories*) with the property that no terminal history is a proper subhistory of any other terminal history;

3. a function (known as the *player function*) that assigns a set of players to any sequence h that is a proper subsequence of a terminal history;

4. for each proper subhistory h of each terminal history and each player i contained in the set of players assigned to h by the player function, a set of actions $A_i(h)$;

5. the payoffs for each player at each possible end node.

It is *consistent* in that h is a terminal history if and only if either

1. h has the form $(a^1, \ldots, a^k)$ for some integer k, the player function is not defined at h and for every $l = 0, \ldots, k-1$, a^{l+1} is a list of actions of the players assigned by the player function to $(a^1, \ldots, a^l)$, or
2. h has the form $(a^1, a^2, \ldots)$ and for every non-negative integer l, the element a^{l+1} is a list of actions of the players assigned by the player function to $(a^1, \ldots, a^l)$.

Example 2.5 (The Treaty Participation Game). This example is based on Chapter 7 of Barrett (2003). This and related games are sometimes known as conjectural variation models (Finus, 2001, Section 13.2) or cartel formation games (Finus & Rundshagen, 2003). We consider the situation that there are two players and the final payoffs are the same as for the prisoner's dilemma (Example 2.1). This game can be divided into three stages.

Stage 1 All players simultaneously choose whether to be a signatory or a non-signatory.

Stage 2 Signatories choose whether to play *Abate* or *Pollute*, with the objective of maximising their collective payoff.

Stage 3 Non-signatories choose simultaneously whether to play *Abate* or *Pollute*.

The subgame perfect equilibrium can be determined by backwards induction, so consider Stage 3 first. The Nash equilibrium of the prisoner's dilemma is for players to play *Pollute*, so non-signatories will play *Pollute*.

We now consider the Stage 2 subgame. If there is one signatory, they will anticipate that the non-signatory will play *Pollute* in Stage 3, and so will also play *Pollute*. If both countries are signatories, they will collectively choose to play *Abate*, because that will maximise their collective payoff.

In the Stage 1 game, if country Y decides not to become a signatory, then country X is indifferent about becoming a signatory. If country Y decides to become a signatory, country X is strictly better off if it becomes a signatory. Country X is therefore not worse off by becoming a signatory regardless of the other players strategy. The subgame perfect equilibrium therefore has all countries becoming signatories. When countries can make a continuous choice about their abatement, they will still choose the optimal abatement level (Barrett, 2003, p. 207).

The extension of the treaty participation game to more than two players has been investigated in (Barrett, 1994) and (Barrett, 2003, Chapter 7). Using a framework similar to that of Example 2.2, Barrett considers an agreement where signatories maximise their collective benefits, while non-signatories maximise their individual benefits. Each player is assumed to have the same emissions cost and benefit functions. Suppose that there are n players, and α is the proportion of players that sign an international environmental agreement, so that it has $n\alpha$ signatories. Let $\pi_n(\alpha)$ be the payoff for a non-signatory, and let $\pi_s(\alpha)$ be the payoff for a signatory. An international environmental agreement is said to be *self-enforcing* if

$$\pi_n(\alpha - 1/n) \leq \pi_s(\alpha) \ \text{ and } \ \pi_n(\alpha) \geq \pi_s(\alpha + 1/n). \tag{12}$$

In other words, an agreement is self-enforcing if no signatory can benefit from dropping out of the agreement and no non-signatory can benefit from joining the agreement. Barrett found that self-enforcing agreements would be likely to have significantly less that full participation. A similar result has also been obtained by Carraro & Siniscalco (1993).

This illustrates a serious barrier to full international cooperation – even when there is an international agreement, countries can have an incentive not comply with the agreement, to not participate in the agreement, possibly by dropping out of the agreement. Measures that may encourage include reciprocal measures, side payments, issue linkage, and trade restrictions[3](Barrett & Stavins, 2003). One possible reciprocal measure is for countries to reduce their emissions by a lesser amount if there is less participation (Barrett, 2003, Chapter 11). Another possible method is to threaten to dissolve the treaty altogether (see Chander & Tulkens (1997) or Chapter 10 of Barrett (2003)). The problem with these punishments in the context of greenhouse gas emissions is that they hurt signatories as much as non-signatories. Threats to substantially increase greenhouse gas emissions are unlikely to be credible and involves impacts that are experienced decades into the future. An alternative way to punish non-cooperation is to link cooperation with another issue, such as trade. Another issue that can be linked to cooperation on reducing emissions is cooperation on research and development. It may however be difficult to prevent the benefits from research and development cooperation from spilling over to other countries.

Cooperation on global warming is automatically linked to trade through a phenomenon known as *carbon leakage.* If a country unilaterally reduces emissions, it could lead to reduced production of some internationally traded emissions intensive goods. This can in turn increase the price of the good. The increased price could then drive increased production of the good in an overseas country that has not reduced its emissions, leading to economic benefits and an increase in emissions for the non-cooperating country.

There are several ways that trade can be linked with cooperation. One way is through trade restrictions. There is a precedent for this – trade restrictions were incorporated into the Montreal Protocol on Ozone Depleting Substances. It has been suggested that the trade restrictions "were indispensable to the protocol's effectiveness" and also helped to drive the ratification process (Benedick, 1991).

The issue of carbon leakage can also be addressed through border tax adjustments. When a country has a price on carbon, a border tax adjustment consists of either: (i) the imposition of a carbon price on imported products that corresponds to a similar

[3] Trade restrictions can also be thought of as a form of issue linkage

tax borne by domestic products; and/or (ii) an exemption from paying a carbon price for the production of exported products. It is likely that border tax adjustments would be allowed under World Trade Organisation rules (Tamiotti *et al.* , 2009). Under the Montreal Protocol, countries accounted for their production of ozone depleting substances, subtracted their exports, and added their imports. Countries were effectively accounting for their consumption of ozone depleting substances. If a country applies border tax adjustments on both exports and imports when it imposes a carbon price, it is effectively putting a price on the *consumption* of emission intensive goods rather than the *production* of emissions.

Barrett (1997) examined the role of trade sanctions by analysing a game structure involving both countries and polluting firms. This is an extensive form game with five stages: in Stage 1, countries collectively decide whether to employ trade sanctions, and if so, under what conditions; in Stage 2, countries simultaneously choose whether to be a signatory to an agreement or not; in Stage 3, signatories choose their abatement levels; in Stage 4, non-signatories choose their abatement levels; finally in Stage 5, firms choose their outputs.

Barrett found than for some choices of parameters, when there were trade sanctions there would be two equilibria. One with no signatories and one with all countries being signatories. The equilibrium with everyone being signatories is preferable and this one can be realised by introducing a minimum participation level into the treaty. The treaty only becomes effective if at least a minimum amount of countries have become signatories. A similar result was obtained by Lessmann *et al.* (2009), who used an integrated assessment model and found that the imposition of tariffs would increase the level of participation of a treaty.

It is also possible to link trade with cooperation by applying a tax to fossil fuels that are exported to a non-cooperating country. Hoel (1994) has suggested that policies that affect both the supply and demand of fossil fuels are superior to policies that affect only the supply or only the demand of fossil fuels. A cartel that exports fossil fuels will capture less rents if other countries reduce their consumption due to an international climate agreement. It would then be in the interests of the cartel to apply a tax on the exported fossil fuel (Bråten & Golombek, 1998).

A final way that trade is linked to cooperation is in international negotiations through implicit or explicit threats to directly link trade to cooperation.

The *American Clean Energy and Security Act of 2009*, also known as the Waxman-Markey Bill, proposes to create a cap and trade emissions reduction scheme in the United States, and includes provisions for border tax adjustments. At the time of writing, it has passed the United States House of Representatives and the status of Senate legislation is uncertain.

The Waxman-Markey Bill proposes to introduce a program for border tax adjustments known as an ‘international reserve allowance program’ if by 2018, the United States is not party to an internationally binding agreement in which major greenhouse gas emitting countries contribute equitably to the reduction of emissions and satisfies the following criteria:

1. it has provisions that address carbon leakage between parties and non-parties;
2. it does not prevent countries from addressing carbon leakage between different parties;
3. it has agreed remedies for any party that fails to meet their emission reduction obligations.

The program would apply to eligible industries that are emissions intensive and trade exposed. Importers of covered goods would need to purchase international reserve allowances, whose price would be equal to the auction price of emissions allowances that are required for compliance with the cap and trade scheme that would also be introduced by the Waxman-Markey Bill. It would not apply to imports from countries that meet any of the following criteria:

1. they are party to an agreement that the United States is also party to, and the country has greenhouse gas reduction commitments 'at least as stringent' as those of the United States;
2. they are party to a multilateral or bilateral agreement for the eligible sector to which the United States is also a party;
3. they have an emissions or energy intensity for that industry that is not greater than that of the United States for that industry;
4. they have been identified by the United Nations as a least developed country;
5. they are responsible for less than 0.5 percent of global greenhouse gas emissions and less than 5 percent of United States imports for the eligible industry.

It is interesting that there is an emphasis on the strength of emissions reduction commitments rather than just participation in the agreement. The analysis above suggests that trade measures can increase the stability of an agreement, but whether trade measures can be used to directly increase other countries mitigation commitments is unclear. There is a risk that countries would use trade measures to shift the burden of emissions reductions from themselves to other countries, and that this could undermine cooperation. Some parts of the legislation are somewhat vague, the legislation does not specify what is meant by "at least as stringent" when if refers to greenhouse gas reduction commitments at least as stringent as those of the United States. For example, it is feasible that a country such as India could be subject to border tax adjustments, even though in 2005 its per-capita emissions were approximately 14 times less than the per-capita emissions of the United States.

India and other developing countries have been highly critical of the border tax adjustment provisions in the Waxman-Markey Bill. The United States President Barak Obama has also been critical, stating "At a time when the economy worldwide is still deep in recession and weve seen a significant drop in global trade, I think we have to be very careful about sending any protectionist signals out there." There is a risk that if any border tax adjustments or other trade measures are not considered to be fair, then the international climate regime would also not be considered to be fair, undermining cooperation.

If an in an international climate agreement is self-enforcing, for reasons to do with issue linkage or otherwise, will the agreed targets be more likely to be close to socially optimal, or less likely? A related question is whether binding or non-binding targets are more likely to be strong targets. Game theory suggests that when an agreement is self-enforcing, players will act under the assumption that other players will comply with the agreement; when an agreement is not, players are likely to assume that other player will not comply. If an agreement had strong penalties for non-participation, countries may be willing to accept targets than they would otherwise accept in order to participate. This may suggest that binding targets are more likely to be close to socially optimal targets than non-binding ones.

However, when countries agree to binding targets, the risks associated with these targets being costly is greater. There is less risk associated with a country agreeing to a non-binding target, because if a non-binding is difficult to comply with, little is lost by not complying. Victor (2007) asserts that with international cooperation on the North Sea, the Baltic Sea, and acid rain in Europe, nonbinding commitments backed by senior politicians were more effective than binding commitments. For the European acid rain regime, ambitious non-binding commitments to reduce nitric oxide and nitrogen dioxide pollutants were adopted by a smaller number of countries alongside a less ambitious binding convention to address the same pollutant. A domestic mechanism for implementing such an approach is described in Section 4 of Wood & Jotzo (2009).

Example 2.6 (Repeated Prisoner's Dilemmas). Repeated prisoner's dilemmas are discussed in (Osborne & Rubinstein, 1994, Chapter 8) and (Finus, 2001, Chapter 5). For repeated prisoner's dilemmas with finite horizon, the only Nash equilibrium consists of players not cooperating in each turn of the game. When games have infinite horizon the 'folk theorems' of game theory tell us that these games have a huge amount of different subgame perfect equilibria. These results suggest that cooperative behaviour is more likely if players have a long term perspective, and have a strategy for punishing players who do not cooperate.

In (Axelrod, 1984), repeated versions of a prisoner's dilemma with the following payoff matrix were studied.

$$\begin{array}{llcc} & & \multicolumn{2}{c}{\text{Player 2}} \\ & & \textit{Cooperate} & \textit{Defect} \\ \text{Player 1} & \textit{Cooperate} & (3,3) & (0,5) \\ & \textit{Defect} & (5,0) & (1,1) \end{array} . \tag{13}$$

Axelrod organised two computer tournaments where players would submit algorithms that determine whether to play a cooperative or noncooperative choice on each move, taking into account the history of the game so far. The first tournament received 14 entries and each game would consist of 200 moves. The second tournament received 62 entries, this time each game would have a 0.00346 chance of ending after each move (so the game would not have finite horizon).

In both tournaments an algorithm called *Tit for Tat* won. *Tit for Tat* starts by cooperating, then in subsequent moves it plays the preceding move played by its opponent. Axelrod analysed the highest scoring strategies and found that they would have several properties in common: they were *nice*, in that they would not defect before their opponent does; they were *forgiving*, they would fall back to cooperating if their opponent does not continue to defect; but they would also be *retaliatory* in that they would immediately defect after an "uncalled for" defection from the other player.

Because repeated games often have a large amount of subgame perfect equilibira, a stronger concept, known as the 'renegotiation proof equilibrium' has been developed (Farrell & Maskin, 1989).

Because countries make decisions about their emissions over time, and change their emissions over time, a repeated prisoner's dilemma is in many ways a more appropriate game to study than a single shot prisoner's dilemma. But there are some important limitations to using repeated games to study greenhouse gas abatement. The most important limitation is that greenhouse gases are a stock pollutant – the damages from greenhouse gases are related to cumulative emissions rather than the emissions during any particular year. Another limitation is that there is quite a bit of delay before the amount of radiative forcing (and therefore damage) from one tonne of greenhouse gases is maximised.

2.3 Bargaining

The question of bargaining is very relevant to the issue of achieving international cooperation to reduce greenhouse gas emissions. In many ways, international climate negotiations are a bargaining process.

Game theory has looked at bargaining from a number of different perspectives. One of the first game-theoretic approaches to bargaining was introduced by Nash (1950), who proposed a solution that maximises the product of each player's improvement in utility. Schelling (1970) examined bargaining from both game theoretic and behavioural viewpoints, as well as looking at problems of strategy and conflict. Schelling discusses the use of pre-commitment to restrict one's options and strengthen one's bargaining position. For this reason Schelling describes bargaining power as "the power to bind oneself". An example of the use of pre-committment to strengthen one's bargaining position is the difficulty of treaty ratification in the United States Senate, as described in Example 2.4.

The following example shows how bargaining can be treated as an extensive form game, and why the subgame perfect equilibrium is important.

Example 2.7 (Split the Pie). This extensive form game was proposed by Rubinstein (1982). Two players are bargaining over a pie, that we assume to be of size 1. A partition of the pie is identified with a number s in the unit interval, which we interpret as the portion of the pie that is received by the first player. Each player in turn offers a portion of the pie to the other player, who will either accept it (ending the game) or reject it.

Each player prefers larger partitions of the pie to smaller partitions and prefers the bargaining to take a shorter amount of time to a longer amount of time. Rubinstein shows that for any partition s in the unit interval, s is induced by a Nash equilibrium. Rubinstein shows that this is not necessarily the case for the subgame perfect equilibrium. The subgame perfect equilibrium is calculated for the situation when there are fixed bargaining costs, and for when there are fixed discounting factors. A more general version of this game is described in (Osborne & Rubinstein, 1994, Chapter 7).

3 Coalitions

There have been debates in the game theory literature on whether a cooperative outcome is more likely to arise from a 'grand coalition' of all countries, or from smaller coalitions. There is also a parallel political debate on the role of the UN in international negotiations. Game theory analyses coalitional behaviour from a variety of perspectives. One such perspective is a cooperative game theory approach, which we examine in Section 3.1. Another perspective is described in Section 3.2 where we examine non-cooperative approaches to coalition formation, and the role of externalities.

3.1 Cooperative Game Theory and the Core

Cooperative game theory investigates situations where groups of players may form coalitions that enforce cooperative behaviour. For cooperative games, the outcomes of interest consist of a partition of the players into coalitions, and actions for each coalition. Players in a coalition behave cooperatively with each other, and non-cooperatively with respect to other players and coalitions. The core is a concept that can be used to analyse the stability of a grand coalition of all players.

Definition 3.1 Let N be a set of n players. A *coalition* is a subset S of N. A *payoff vector* (also known as an imputation) for N is an n-dimensional real vector $\pi = (\pi_1, \ldots, \pi_n)$, and we write $\pi(S) = \sum_{i \in S} \pi_i$ for any coalition $S \subseteq N$. A *characteristic function* v (also known as a coalitional function) is a function which assigns a real number to each coalition.

We say that a characteristic function v is *zero-normalised* if $v(\{i\}) = 0$ for $i = 1, \ldots, n$; and that v is *super-additive* if $v(S \cup T) \geq v(S) + v(T)$ for any disjoint subsets S, T of N.

An n-player *game in coalitional form with transferrable utility* (also called a TU-game) is defined by a set of players N, and characteristic function v, and denoted (N, v). The *core* of (N, v) is defined by

$$C(N, v) = \{\ \pi : \pi(N) = v(N) \text{ and } \pi(S) \geq v(S) \text{ for all } S \subseteq N\ \}. \tag{14}$$

The core is the set of possible outcomes in which no coalition can break away from a grand coalition in such a way that all of its members are better off. The core, being a set, always exists, but can be empty.

Example 3.1 (The γ-Core of Chander & Tulkens (1997)). This example is based on (Chander & Tulkens, 1997), which is also discussed in Chapter 13 of (Finus, 2003) and Chander & Tulkens (2008). We use the same basic framework as in Example 2.2. Let $\pi_i(e^S, e^{N \backslash S})$ be the payoff for a country i in a coalition S which has e^S emissions, and with the other countries emitting $e^{N \backslash S}$ emissions. Assume that each of the countries in $N \backslash S$ maximise their individual benefits, while countries in S maximise their collective benefits. The γ-characteristic function of a coalition S is the sum of the utilities of each member os S, assuming that members of $N \backslash S$ behave non-cooperatively. It is given by

$$v_\gamma(S) = \sum_{i \in S} \pi_i(e^S, e^{N \backslash S}). \tag{15}$$

The core of the associated TU-game can be thought of as the set of possible payoff vectors for the countries in a grand coalition where no coalition will benefit if the grand coalition dissolves into singletons when any coalition breaks away from it. The payoffs depend both on a country's emissions and a transfer t_i of payments received by country i that satisfies $\sum_{i \in N} t_i = 0$. The total payoff for country i is given by

$$\pi_i = \beta_i(e_i) - \phi_i\Big(\sum_{j \in N} e_j\Big) + t_i. \tag{16}$$

Chander and Tulkens show that the γ-core is non-empty by constructing a payoff vector that is contained in it. Let $\bar{e}_i$ be country i's Nash equilibrium emissions and let e_i^* be country i's social optimum emissions. The values for t_i chosen are

$$t_i = (\beta_i(\bar{e}_i) - \beta_i(e_i^*)) - \frac{\phi_i'(\sum_{j \in N} e_j^*)}{\sum_{k \in N} \phi_k'(\sum_{j \in N} e_j^*)} \Big(\sum_{k \in N} \beta_k(\bar{e}_k) - \beta_k(e_k^*) \Big). \tag{17}$$

This choice of t_i corresponds to an element of the γ-core if any of the following conditions hold:

1. damage functions are linear;
2. for all $S \subset N$ with $|N \backslash S| \geq 2$, and for all $i \in S$, $\sum_{k \in N \backslash S} \phi'_k(e^*) \geq \phi'_i(\bar{e})$; or
3. countries are symmetric.

The result of Example 3.1 suggests that socially optimal emission reductions could be possible, but it has been questioned whether this outcome is feasible. The threat that each countries will break into singletons if one or more countries leave the grand coalition may not be credible. Finus has suggested (Finus, 2001, Section 13.3.3) that variants of the treaty participation game (as discussed in Example 2.5) may be a more feasible approach. Finus also points out (Finus, 2001, Section 13.3.3) that the cost sharing rule provides countries with an obvious incentive to misrepresent their environmental preferences and abatement costs. In Example 4.6, we shall describe a mechanism where it is optimal for players to state their their true abatement costs.

Although the cost sharing rule (17) may not be practical or feasible, it is still important because it demonstrates that the core can be non-empty. This is significant because it has been shown (Serrano, 1995), (Okada & Winter, 2002) that it is possible to design extensive form games (which can be thought of as a bargaining game) whose subgame perfect equilibria are elements of the core. This relates to the ‘Nash program’ (Nash, 1953; Serrano, 1997) to link cooperative and non-cooperative game theory by finding non-cooperative procedures that yield cooperative outcomes as their solution concepts.

We note that the core for a global warming game that does not assume that countries in $N \backslash S$ dissolve into singletons has been studied by Uzawa (2003). In this case the core may be empty. Uzawa also investigated the situation where utility is non-transferrable.

3.2 Coalition Formation and Externalities

The fully cooperative result from Chander and Tulkens described in Example 3.1 contrasts with the less cooperative results from Barrett (1994) and Carraro & Siniscalco (1993) that we discussed in Section 2.2. This has lead to a debate in the game theory literature about whether cooperation on climate change is best achieved among all countries working together, or among smaller coalitions. The debate has been surveyed by Tulkens (1998) and ten years later by Chander & Tulkens (2008). Tulkens (1998) described the results of Barrett (1994) and Carraro & Siniscalco (1993) as the small stable coalitions (SSC) thesis, and the results of Chander & Tulkens (1997) as the grand stable coalition (GSC) thesis. The role of coalitions in the different approaches is different – under the SSC approach, the ‘bad guys’ who do not cooperate are singletons, outside of any coalition; under the GSC approach, the ‘bad guys’ who do not cooperate form a coalition.

When there are coalitional externalities, assumptions about the coalitions that do not contain a particular player change the value of the characteristic function for that player.

This in important when analysing issues such as the core, and the stability of a grand coalition. Chander & Tulkens (2008) point out that an alternative to using characteristic functions is a 'partition function' that also takes as its input a partition of the other players into coalitions.

The approach of Barrett (1994) and Carraro & Siniscalco (1993) has the property that the number of non-trivial coalitions is restricted to one; the use of a partition function facilitates going beyond this assumption. The following definition of a partition function is from Maskin (2003).

Definition 3.2 Let N be a set of n players and let $\mathscr{C}$ be a partition of N into disjoint coalitions. For each partition $\mathscr{C}$ and coalition $C \in \mathscr{C}$, the *partition function* $v(\cdot, \cdot)$ assigns a number $v(C, \mathscr{C})$, which is interpreted as the payoff for coalition C given the partition $\mathscr{C}$.

The partition function is *zero-normalised* if $v(\{i\}, \mathscr{C}) = 0$ for all $i \in 1, \ldots, n$ and all partitions $\mathscr{C}$ of N. It is *super-additive* if $v(C_1 \cup C_2, \mathscr{C}_{12}) \geq v(C_1, \mathscr{C}) + v(C_2, \mathscr{C})$ for any partition $\mathscr{C}$ of N and coalitions $C_1, C_2 \in \mathscr{C}$, where $\mathscr{C}_{12}$ is the same as $\mathscr{C}$ but with C_1 and C_2 replaced by $C_1 \cup C_2$.

Finus & Rundshagen (2003) have applied partition functions to climate change coalitions. They consider a two-stage game, each stage can also be analysed as a game: in the first stage countries choose their coalitions; and in the second stage, coalitions choose their optimal strategy. They consider a large variety of different approaches to how countries choose their coalitions, including the approach of Barrett (1994). These approaches model the process of coalition formation as an extensive form non-cooperative game. The size and nature of the coalitions that form depend very much on this process. Some of these processes (such as the Barrett (1994) approach) have very small coalitions, but in some cases a grand coalition was possible. Buchner & Carraro (2006) have also used this two-stage process, and incorporated it with a six-region economic model *FEEM-RICE*. How coalition formation can be treated as a non-cooperative game has been discussed in more general context by Bloch (1996), Ray & Vohra (1997), Yi (1997), and Maskin (2003). Yi (1997) also found that different rules of coalition formation lead to different predictions about stable coalition structures.

For some games coalition formation imposes a positive or negative externality on other players (Maskin, 2003), (Yi, 1997), (de Clippel & Serrano, 2008). With the basic framework that we use to analyse climate change (Example 2.2), coalition formation imposes a positive externality – when a group of countries form a coalition, their emissions will be lower than when they act individually in a non-cooperative way.

Definition 3.3 Let $\mathscr{C}$ be a partition of a set N of n players into disjoint coalitions. Let $C_1, C_2 \in \mathscr{C}$ be two of these coalitions, and let $\mathscr{C}_{12}$ be the partition that forms when C_1 and C_2 merge. For some other coalition $C \in \mathscr{C}$, we say that C_1 and C_2 impose *no coalition* *externality* on C if merging has no effect, i.e.

$$v(C, \mathscr{C}) = v(C, \mathscr{C}_{12});$$

the coalitions C_1 and C_2 impose a *positive coalition externality* on C if merging increases C's payoff

$$v(C, \mathscr{C}) < v(C, \mathscr{C}_{12});$$

the coalitions C_1 and C_2 impose a *negative coalition externality* on C if merging decreases C's payoff

$$v(C, \mathscr{C}) < v(C, \mathscr{C}_{12}).$$

There is a large variety of non-cooperative coalition formation games that have been studied. Some of them involve players making simultaneous moves, some involve sequential moves. We list some of these games below:

- The *treaty participation game* that was described in Section 2.2 has been studied by Carraro & Siniscalco (1993), Barrett (1994), Finus & Rundshagen (2003) and others. A variant of this game is where players also consider the impact of other defections that could arise if a player defects from a coalition (Carraro & Moriconi, 1997; Finus & Rundshagen, 2003), and leads to a more cooperative outcome.

- The *equilibrium binding agreement game* was introduced by Ray & Vohra (1997) and has also been studied by Finus & Rundshagen (2003). The starting point is a grand coalition C. Then a smaller coalition, c, may split away from the grand coalition. In the following step, members of either c or $C \backslash c$ may propose further deviations. This process continues until no player wants to split up into finer partitions.

- *Open membership games* have been studied by Yi (1997) and Finus & Rundshagen (2003). In these games, players can freely join coalitions and no outsider is excluded from a coalition.

- *Exclusive membership games* have been studied by Finus & Rundshagen (2003), Hart & Kurz (1983), Yi (1997), and Yi & Shin (2000). Players first simultaneously list the players who they are willing to join a coalition with. In one type of exclusive membership game, known as the Δ-Game, two players are in the same coalition if and only if they are on each others list. In another exclusive membership game, the Γ-Game, a group of players are in the same coalition if and only if their lists are all identical. In their model with symmetric countries, Finus and Rundshagen found that larger coalitions were sustained by the Γ-Game than by the open membership game or the cartel formation game.

- Bloch (1996) and Finus & Rundshagen (2003) have examined the *sequential move unanimity game*. We start with an exogenous ordering of players. The first player proposes a coalition to which they would like to belong; each prospective member then is asked (according to the same ordering) whether they accept the proposal; if all proposed members agree, then a coalition is formed and the remaining players may form a coalition according to the same process; if a proposed member disagrees, they can then propose their own coalition.

- Maskin (2003) introduced a sequential process that is also based on an exogenous ordering, and proved that when externalities were negative, a grand coalition forms (for up to three players). A counter-example was provided by de Clippel & Serrano (2008) to this statement when there was more than three players. Maskin also provided examples of positive externality games where a grand coalition would not form.

- Aghion *et al.* (2007) compared two specific bargaining processes, in order to understand whether multilateral approaches are more likely to lead to cooperation on trade or bilateral processes were. They only modelled three players, and found that for the processes that they investigated, a grand coalition would form, even when coalitional externalities were positive.

The processes described above is a non-cooperative approach to coalition formation. A significant question in game theory is which non-cooperative processes can implement concepts in cooperative game theory. We will discuss how to design non-cooperative games with cooperative solutions in the next section.

What do coalition formation processes tell us about the role of the UNFCCC in climate negotiations, if it does indeed tell us anything? Without a credible threat, or forms of issue linkage such as trade restrictions, a fully cooperative grand coalition seems unlikely. But this does not in itself make the UNFCCC process unimportant, even though it is in many ways based on consensus, that can be easily blocked. It is possible to have negotiations among smaller groups in parallel with the UNFCCC negotiations. This has been occurring in fora such as the *Major Economies Forum on Energy and Climate*, and this is likely to continue. Negotiations among smaller groupings could be complements to the UN process, rather than substitutes.

The fact that more cooperation is likely to occur with exclusive membership games than with open membership games could have implications for how to get the most cooperation from a coalition formation process. In some ways, the exclusive membership games are similar to what arises when countries with emissions trading schemes are considering the possibility of linking their carbon markets. Countries that establish an emissions trading scheme may want to it it with those of other countries for efficiency reasons. But countries would be reluctant to link with a country whose emissions trading rules are significantly different (Jotzo & Betz, 2009), or whose mitigation commitment is much less ambitious. This suggests that carbon market linkage has important strategic implications.

4 Implementation Theory

Implementation theory addresses the key game-theoretic question that needs to be answered in order to address a social dilemma. How can non-cooperative games be designed so that their solution (often a Nash equilibrium or subgame perfect equilibirum) corresponds to a socially optimal outcome? We now will proceed with a formal treatment of the concepts from implementation theory. We will then examine some mechanisms that relate to public good provision or pollution reduction, and discuss their relevance to climate change mitigation. The reader is referred to Jackson (2001) for a more detailed summary of the main concepts of implementation theory.

Let N be a set of n *players*, and let A be a set of possible *outcomes*. Let a player i have a preference relation R_i on A; if player i prefers an outcome a to another outcome b, or is indifferent, we say that aR_ib. An example of a preference relation is when each player i assigns a utility u_i to each outcome, in which case, aR_ib if and only if $u_i(a) \geq u_i(b)$.

A *social choice correspondence* F maps profiles of preferences $R = \{R_1, \dots, R_n\}$ into the set of outcomes, i.e. $F(R) \subset A$. When $F(R)$ is a singleton, F is called a *social choice function*. A social correspondence tells us what outcomes are desirable, given a preference profile. We have made extensive use of the *social optimum*, a social choice function that maximises the sum of the utilities of each player.

A *mechanism* or *game form* is a pair (M, g) consisting of a product of 'message spaces' or 'strategies' $M = M_1 \times \ldots \times M_n$, and an *outcome function* $g : M \to A$. The main difference between a mechanism an non-cooperative game is that the result of the mechanism is given by an outcome, rather than a payoff. A *solution concept* S specifies the behaviour of players who have preferences R, given a mechanism (M, g). Given (M, g, R), S specifies a subset of M. The outcome function will then lead to an *outcome correspondence* that is given by

$$O_S(M, g, R) = \{a \in A \;:\; \text{there exists } m \in S(M, g, R) \text{ such that } g(m) = a\}. \tag{18}$$

Important examples of solution concepts include the Nash equilibrium and the subgame perfect equilibrium.

A social choice correspondence is *implemented by the mechanism* (M, g) via a solution concept S if the outcome correspondence coincides with the image of the social choice correspondence. In other words,

$$O_S(M, g, R) = F(R). \tag{19}$$

A field that is closely related to implementation theory is *mechanism design*. The mechanism design problem involves finding mechanisms where the outcome correspondence contains the the social choice correspondence, but where there could be other solutions as well, i.e. $O_S(M, g, R) \supset F(R)$.

The use of subgame perfect equilibrium as a solution concept is particularly important, because there exist situations where a choice function cannot be implemented in a single stage via Nash equilibrium, but can be implemented in several stages via subgame perfect equilibrium (Moore & Repullo, 1988).

Example 4.1 This illustrative example is based on Moore & Repullo (1988). Suppose that there is a club with a set N of members that are designing a constitution – a mechanism (M, g) for making decisions. This mechanism could, for example, be a voting procedure, or a consensus based decision procedure. A social choice function F, together with the member's preferences R, determine the decision $F(R)$ that would be preferred. The members preferences R, may be known to each other, but unknown to outsiders, such as a court. For this reason, instead of directly using $F(R)$ to make a decision, the constitution specifies an outcome function g, based on messages M, both of which can be verified.

An interesting property of this mechanism is that there is no social planner, such as a government, that implements the mechanism. An example of such a club could be the UNFCCC, where the decision making body (the 'conference of parties') mostly makes decisions using consensus.

Because mitigation of climate change is a global public good, it is useful for us to consider non-cooperative games whose solutions implement a public good. We shall now consider some more examples of games that do this.

Example 4.2 (Subscription Games). Bagnoli & Lipman (1989) describe a relatively simple game for providing public goods using voluntary contributions. Each player voluntarily commits any amount of their choice towards the cost of the public good. The public good is considered to be discrete – the example of a single streetlight or multiple streetlights is described. Players 'pledge' to make contributions towards completion of the project. If the total amount of contributions is enough to provide the public good, then players must pay and the good is provided. If the total amount of contributions is not enough, each player's contribution is refunded and none the public good is not provided. Bagnoli and Lipman model this process with a normal form game. They show that this game has a solution that satisfies a solution concept known as 'undominated perfect equilibrium'. This solution provides the public good and implements the core of the economy. This mechanism is also sometimes known as a *provision point mechanism.*

An extensive form version of this game is described by Admati & Perry (1991). They call this game the *subscription game.* The structure of this game is similar to the game in Example 2.7 in that players take turns to make an offer. For simplicity, assume that there are two players. Players alternate in pledging contributions to complete the project. The game ends if and when the total amount of contributions exceeds the cost of the good. Let c_i be the total contribution from player i, let k be the cost of the public good, let v be the benefit of the public good for each player and let T be the first time such that

$c_1 + c_2 > k$. We assume that the payoffs for each player are given by

$$\pi_i(T, c_1, c_2) = \begin{cases} \delta^T(v - c_i) & \text{if } c_1 + c_2 \geq k \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Admati and Perry prove that the subgame perfect equilibrium of this game is as follows.

1. If $k > 2v$, so that the total cost of the good is greater than its benefits, each player subscribes nothing and the good is not provided.

2. If $v(1 - \delta) < k < 2v$, then the good is provided. The first player subscribes in the first move
$$c_1^* = \frac{k - v(1 - \delta)}{1 + \delta}, \tag{21}$$
and the second player subscribes in their first turn
$$c_2^* = k - c_1^*. \tag{22}$$

3. If $k < v(1 - \delta)$, then the first player subscribes k in the first move and the good is provided.

4. If $k = v(1 - \delta)$, then there are two equilibria. In one equilibrium the first player subscribes k in the first move and the good is provided. In the other equilibrium the second player subscribes k in their first turn and the good is provided.

Admati and Perry also consider the game where players are not refunded their commitments if the good is not provided. In this case, in equilibrium the good will not be provided unless the value of the good to each player is greater than the cost of the good.

Marks & Croson (1998) have performed experiments which also suggest that the subscription game can be successful. Advantages of the subscription game are that it is reasonably simple, and that it does not require a strong sanctioning institution such as a government that can enforce the desired contributions to public goods.[4] Bagnoli and Lipman state

> "Even the analysis of mechanisms which are put forth as 'plausibly useful', such as Groves-Clarke taxes, is focused on mechanisms that a government might actually wish to impose and rarely on mechanisms which private individuals might jointly use. Perhaps for this reason, the literature on private provision of public goods has basically ignored the implementation literature, hypothesized particular games, and demonstrated, among other things, that these games do not have efficient outcomes." (Bagnoli & Lipman, 1989, p. 596)

[4] Some sort of institution may still be required to ensure players do not renege on their commitments, and to provide the public good once it has been paid for.

Example 4.3 (Using Bargaining to Resolve a Non-cooperative Game). Attanasi *et al.* (2010) develop a bargaining process that they call a *confirmed proposal* mechanism that can lead to cooperative outcomes. They describe two mechanisms: a game form with 'confirmed conditional proposals', and a game form with 'confirmed unconditional proposals'. We describe the game form with confirmed conditional proposals below. These mechanisms are a little similar to Rubinstein's alternating offers 'split the pie' game (Example 2.7), but do not have a time factor. There is an underlying game (such as a prisoner's dilemma) that determines the payoffs for each player. There are two players, with 'strategy spaces' S_1, S_2 for the underlying game. Pairs of strategies can be thought of as outcomes, and the underlying game treats these outcomes as strategies that determine utility functions (which in turn determine the players' preference profiles). The mechanism proceeds as follows:

Stage 1.1 Player 1 communicates to Player 2 their intention to follow a strategy $s_1^1 \in S_1$, if the bargaining process arrives at an agreement.

Stage 1.2 Player 2 responds to Player 1's proposal by communicating their intention to follow strategy $s_2^1 \in S_2$ if Player 1 is willing to confirm their strategy.

Stage 1.3 Player 1 has a choice about whether to confirm their strategy or not. If so, then the two players choose strategies (s_1^1, s_2^1); if not, the players proceed to Stage 2.

Stage 2.1 The reply of Player 2 in Stage 1.2 becomes Player 2's new proposal, i.e. $s_2^2 = s_2^1$.

Stage 2.2 Player 1 announces their intention to follow strategy $s_1^2 \in S_1$, which must be different from their proposal from Stage 1.1 (i.e. $s_1^2 \neq s_1^1$).

Stage 2.3 Player 2 must choose whether or not to confirm the strategy profile $(s_1^2, s^2, 2)$. If so, the bargaining process ends; if not, they return to Stage 1, but with the proposal of Player 1 in Stage 1.1 being the same as their proposal from Stage 2.2, and the proposal of Player 2 in Stage 1.2 being different from their proposal in Stage 2.1.

Attanasi *et al.* (2010) show that when the underlying game is a prisoner's dilemma (so that the players' preference profiles lead to a prisoner's dilemma), the game has a subgame perfect equilibrium that induces the cooperative outcome in the first bargaining stage. In other words, the cooperative outcome is implemented by the confirmed conditional proposal mechanism, when the player's preferences are a prisoner's dilemma. They also found that experiments using this mechanism with human subjects sustained high amounts of cooperation.

The above bargaining mechanism has some similarities to scenarios that can occur in international climate negotiations. Sometimes a country will state what they are prepared to so as part of a 'comprehensive international agreement' or something similar. They may later confirm whether their proposal is an actual commitment or not. For example, some countries, including Japan and New Zealand, made commitments under the Copenhagen Accord that are conditional on actions from other countries. During the negotiations at Copenhagen in 2009, Russia stated that a previous commitment from their President for a 2020 emission level that is between 20 and 25 percent less than their 1990 commitment level was based on "a substantive, comprehensive agreement as a solution to the long-term cooperative action track of the negotiations".

Another form of conditionality that takes place during climate negotiations is when countries state that they will make an unconditional commitment, but are willing to increase their emission reductions based on the commitments of others. For example, at Copenhagen the EU had an unconditional commitment to reduce its emissions by 20 percent compared to 1990 levels by 2020, but would be willing to reduce its emissions by 30 percent compared to 1990 levels if there was a sufficient commitment from other countries. Australia made an unconditional commitment to reduce its emission by 5 percent compared to 1990 levels by 2020, and a commitment to increase that by up to 15 percent if certain commitments were met, and to 25 percent if certain other conditions were met. These kinds of approaches are in many ways similar to the provision point mechanism of Example 4.2, but are also similar to the 'matching abatement commitment' approach described below.

Example 4.4 (Matching Abatement Commitments). A game where players commit to reducing their emissions by a multiple of other player's targets on top of their unconditional commitments is considered by Boadway *et al.* (2009). Each country chooses a *matching rate* and its level of *direct abatement.* The game proceeds as follows:

Stage 1 Each country i simultaneously chooses matching rates m_{ij} that correspond to country j's direct abatement.

Stage 2 Each country i simultaneously chooses their direct abatement levels a_i. After Stage 2, the total abatement commitment of country i is

$$A_i = a_i + \sum_{j \neq i} m_{ij} a_j. \tag{23}$$

Stage 3 Countries engage in trading of their emission quotas to equalise the marginal benefits of emissions across all countries.

Boadway *et al.* (2009) show that the subgame perfect equilibrium of this process achieves the efficient level of pollution abatement. They extend their model to a situation with two time periods, and treat the pollutant as a stock pollutant (so that it can build up in the atmosphere). They show that the above process also efficiently allocates emissions across periods.

The previous three mechanisms all have cooperative outcomes, and all are based on some sort of conditionality. The fact that their solutions are cooperative suggests that a cooperative approach to climate mitigation is possible. This contrasts with less optimistic views, such as from Brennan (2009), who states that the grounds for hope are "decidedly thin".

These mechanisms require that countries can make a commitment that they cannot backtrack from at each stage of the mechanism. It suggests that if an international legal architecture is devised for cooperation on climate change, a mechanism that makes 'legally binding' conditional commitments possible would be desirable. A novel approach to get countries to make commitments that they will not backtrack from is described below.

Example 4.5 (A Deposit Based Compliance Mechanism). A two-stage mechanism to provide public goods when there are not strong institutions has been described by Gerber & Wichardt (2009). We assume that there is an underlying public goods game such as the game in Example 2.2. In Stage 1, each player is required to pay a deposit. In Stage 2, there are two possible outcomes. If not all players paid the deposit in Stage 1, then the deposits are refunded and the underlying public goods game is played. If all players paid the required deposit, then in Stage 2 players are required to make a pre-specified contribution to the public good. If a player makes the contribution, they get their deposit back. If their contribution is less than what was specified, they do not.

Gerber and Wichardt show that provided the deposits satisfy a certain inequality, and the payoffs for each player are greater when all players contribute the specified contribution than when nobody does, then it is a subgame perfect equilibrium for each player to contribute the specified amount to the public good. The mechanism discourages players from reneging from their commitments because by making a deposit prior to the contribution stage, they make it a dominant strategy to comply with the agreement. The action of paying the deposit can be thought of as a way for players to execute their own punishment, rather than have to punish anyone else.

An institution is required to collect deposits, monitor players' contributions, and refund deposits. The institution does not have to implement the provision of the public good itself, or enforce punishments of free-riders.

In many of the situations described here, such as Example 2.2, players' preferences are known. An important issue in implementation theory is how to find mechanisms that induce players to reveal their preferences. A significant problem with achieving international cooperation is that players often have strong incentives to misrepresent their abatement costs and environmental preferences. In the climate negotiations, countries have an incentive to exaggerate their abatement costs in order to negotiate a weaker target for themselves or reduce the likelihood of being committed to a target. We shall now examine a mechanism that induces players to reveal their true abatement costs.

Example 4.6 (An Auction that Induces Truthful Bidding). We describe an auction mechanism for pollution licences in which the amount of licences are endogenous to the auction. Players have an incentive to bid truthfully and their payments are equal to the externality they impose.[5] This mechanism was introduced in (Montero, 2008), and applications to global warming have are described in (Montero, 2007b). We consider an administrator and some players. In a national setting, the administrator could be a government and the players could be firms. In an international setting, the administrator could be some other sort of institution and the players could be countries.

The mechanism works in the following way. Firstly the administrator specifies a damage function for the pollutant. Then each player submits a schedule of how much they are willing to pay for each licence. Finally the administrator announces the price and amount of permits auctioned, and also a payment to each player. The role of the payment is to induce each player to bid truthfully.

Let ϕ be the emissions damage function which satisfies $\phi' > 0$ and $\phi'' \geq 0$. Let P_i be the emission demand function for player i, which satisfies $P_i' \leq 0$. The demand function can be thought of as the marginal benefits of emissions for player i, the derivative of the emissions benefit function. This setting is slightly different to the setting of Example 2.2 in that we are dealing with an aggregate damage function rather than a damage function for each player. When a player i submits a bid, it submits what it claims to be its benefit function $\hat{P}_i$ that is supposed to approximate P_i. Because $\phi' > 0$, it is possible to define an inverse function S, which we call the administrator's supply function, that satisfies $S(\phi'(x)) = x$ and represents the amount of licences that the administrator would be willing to sell at a particular price. Because $\hat{P}_i' < 0$, it has an inverse function $\hat{X}_i$, the demand schedule, which satisfies $\hat{X}_i(\hat{P}_i(x)) = x$.

For each player i, the administrator computes the residual supply function S_i, using the other players reported demand schedules,

$$S_i(p) = S(p) - \sum_{j \neq i} \hat{X}_j(p). \tag{24}$$

Each S_i satisfies $S_i' > 0$, and so has an inverse that we call the residual marginal damage function ϕ_i', satisfying $\phi_i'(S_i(p)) = p$. The administrator clears the auction by determining a price p_i and a number of licences l_i such that

$$p_i = \hat{P}_i(l_i) = \phi_i'(l_i), \tag{25}$$

and so

$$l_i = S_i(p_i) = \hat{X}_i(p_i). \tag{26}$$

[5] Mechanisms that induce players to bid truthfully and pay the cost of the externality that they impose are known as Vickrey-Clarke-Groves mechanisms. Another mechanism with these properties is described by Dasgupta *et al.* (1980).

If we substitute (24) in (26), we have that

$$S(p_i) = \sum_j \hat{X}_j(p_i). \tag{27}$$

So p_i does not depend on i, and each player pays the same licence price p.

Each player is also paid by the administrator an amount $\alpha_i(l_i)$, which is set to be

$$\alpha_i(l_i) = 1 - \frac{\phi_i(l_i)}{\phi_i'(l_i)l_i}, \quad \text{where} \quad \phi_i(l_i) = \int_0^{l_i} \phi_i'(z)dz. \tag{28}$$

It is optimal for each player to bid their true demand curve by setting $\hat{P}_i = P_i$, regardless of what other players bid (Montero, 2008, Proposition 3).

We shall now look more closely at how this mechanism will work in the context of climate change. It will be simpler to first consider how it would work in a national setting, where a government is using this mechanism to auction some annual permits for emitting greenhouse gases. Greenhouse gases are stock pollutants, so different levels of greenhouse gas pollution from a single country in a single year will have a very small effect on the marginal damage function (also known as the social cost of carbon) ϕ'. We can therefore assume that there exists a very small $\varepsilon > 0$ such that $\phi'' < \varepsilon$, and that there is some price $\tilde{p} > 0$ such that $\phi'(x)$ is approximately equal to $\tilde{p}$ for all x. It then follows that

$$\begin{aligned} \alpha_i(l_i) &\approx 1 - \frac{\tilde{p}l_i}{\tilde{p}l_i} \\ &= 0. \end{aligned}$$

So when ε is so small that it is negligible, the auction mechanism will be equivalent to a carbon tax. This is consistent with players having less of an incentive to misrepresent their abatement costs under a carbon tax. However, there are reasons why a government would choose a different damage function. Stern (2009) notes (p. 101) that

> The social cost of carbon is calculated by estimating the damages created by the emissions of an extra unit of carbon, keeping in mind that this extra unit results in higher concentrations of greenhouse gases in the atmosphere over the very long term. That calculation depends on the assumptions made on the future path of the economy and of emissions, the strength of the carbon cycle shaping absorption, climate sensitivity, distributional and intertemporal values, and so on. If you also factor in assumptions about probability distributions of all those things and attitudes to risk, you can get a huge range of estimates for the social cost of carbon. This means that the social cost of carbon is a very weak and unreliable peg for policy.

A commonly suggested alternative to a carbon tax is an emissions trading scheme, which can be thought of as having the government choose some amount of licences $\tilde{l}$; letting

$\phi'(x) = 0$ for $x < \tilde{l}$; and letting $\phi'(x)$ be infinite for $x > \tilde{l}$. A hybrid approach is described in (Roberts & M. Spence, 1976), where the administrator approximates a marginal damage function with a piecewise constant function. The auction described here can be thought of an a mechanism for implementing general hybrid approaches to control emissions.

Montero (2007b) suggests that one way for countries to find out their mitigation costs would be to have an auction at the national level. This could then inform an auction at the international scale. In an international situation, the amount of emissions involved are greater than at the national level. The emissions from a single player i are likely to have a greater impact on the marginal damage function ϕ', so ϕ'' is likely to be greater, and α_i is likely to be greater. A significant barrier to implementing this mechanism is that it is not revenue neutral. Players are unlikely to want to participate if it involves significant wealth transfers to a third party. A system of paybacks is described in Montero (2007a) which comes close to being budget balancing, and becomes closer to being budget balancing when there are more players.

There are some other barriers that would need to be addressed if this mechanism was to be used internationally in practice. An institution would be required to both administer the auction and transfer the payments. There would also be the issues of potential non-compliance and non-participation. Finally, a mechanism for players to agree on a damage function would be required. This mechanism may be somewhat more complicated than some of the other mechanisms described in this section. Regardless of whether it is feasible at the international scale, it is of theoretical interest, and informs decisions about the choice of instrument for pricing greenhouse gas emissions.

5 Conclusion

In its simplest form, climate change mitigation is a prisoner's dilemma. The prisoner's dilemma has a Nash equilibrium that involves players acting non-cooperatively in a manner that is socially sub-optimal. When countries have a continuous choice about how much to pollute, the Nash equilibrium involves much more pollution than is optimal. This is why climate change is sometimes known as a social dilemma.

Normal form games such as this help us to understand the free-rider problem, but do not tell is about the sequential nature of strategic behaviour. Being able to do this is important for addressing the social dilemma.

Extensive form games that have more than one stage, such as the treaty participation game, can have solutions that are more cooperative, as their subgame perfect equilibrium. For two players, the treaty participation game implements a cooperative outcome. But for more than two players, there is only partial cooperation. Ways to address this may include the use of punishments; and issue linkage, possibly involving trade. The subgame perfect equilibrium also helps us to understand the process of treaty ratification and its strategic implications.

However, solution concepts such as the Nash equilibrium and the subgame perfect equilibrium do have limitations. Experiments using the ultimatum game illustrate this, by illustrating that human behaviour not only influenced by strategic considerations, it is also influenced by reciprocity and conceptions of fairness. Disciplines which help us understand human behaviour in these situations is a very useful complement to the use of game theoretic solution concepts.

When game theory is used to help us understand coalitions, outcomes that are more cooperative than the treaty participation game are possible. A socially optimal outcome has been predicted by Chander & Tulkens (1997), using cooperative game theory and the concept of the γ-core. However, this outcome is based on a threat that might not be credible, so may not be realistic. But many non-cooperative models of coalition formation have subgame perfect equilibria that are more cooperative than predicted by the treaty participation game, including several that were studied by Finus & Rundshagen (2003). Because one of the coalition formation processes, the exclusive membership game, has as a significantly more cooperative solution that some of the others, it may be the case that carbon market linkage can help facilitate a cooperative outcome.

There are several strong results about mechanisms that implement a cooperative outcome via subgame perfect equilibrium when there is a social dilemma. These include subscription games (Example 4.2), bargaining based on confirmed proposals (Example 4.3), and approaches where countries 'match' each others pollution abatement commitments (Example 4.4). All of these approaches make use of conditionality. This suggests that when countries are willing to increase their emission reduction commitment if others do the same, cooperation is more likely. It also suggests that cooperation would be more likely if an international mechanism were to exist that would allow countries to make a binding conditional commitment.

Game theoretic approaches inform our understanding of participation and compliance in international agreements, the role of coalitions, and the role of conditionality when bargaining over emission reductions. This can help us understand the social dilemma associated with climate change and provide insights that may help us address it.

Acknowledgements

The author would like to thank Stephen Howes, Frank Jotzo, and David Pannell for comments on earlier versions of the text. This work was supported by the Australian Government Department of the Environment, Water, Heritage and the Arts through the Environmental Economics Research Hub.

References

Admati, A.R, & Perry, M. 1991. Joint Projects without Commitment. *Review of Economic Studies*, **58**(2), 259–76.

Aghion, P., Antras, P., & Helpman, E. 2007. Negotiating free trade. *Journal of International Economics*, **73**(1), 1 – 30.

Attanasi, G., Gallego, A. G., Georgantzis, N., & Montesano, A. 2010. *Non-cooperative games with confirmed proposals*. Working Paper. LERNA Travaux No. 10.02.308.

Axelrod, R. M. 1984. *The evolution of cooperation*. Basic Books, New York.

Baer, P., Harte, J., Haya, B., Herzog, A. V., Holdren, J., Hultman, N. E., Kammen, D. M., Norgaard, R. B., & Raymond, L. 2000. Climate Change: Equity and Greenhouse Gas Resposibility. *Science*, **289**(5488), 2287.

Bagnoli, M., & Lipman, B. L. 1989. Provision of Public Goods: Fully Implementing the Core through Private Contributions. *Review of Economic Studies*, **56**(4), 583–601.

Barrett, S. 1994. Self-Enforcing International Environmental Agreements. *Oxford Economic Papers*, **46**, 878–894.

Barrett, S. 1997. The strategy of trade sanctions in international environmental agreements. *Resource and Energy Economics*, **19**, 345–361.

Barrett, S. 2003. *Environment and Statecraft*. Oxford Oxfordshire: Oxford University Press.

Barrett, S., & Stavins, R. 2003. Increasing Participation and Compliance in International Climate Change Agreements. *International Environmental Agreements: Politics, Law and Economics*, **3**, 349–376.

Benedick, R. E. 1991, 1998. *Ozone Diplomacy: New Directions in Safeguarding the Planet*. Enlarged edn. Cambridge, Massachusetts: Harvard University Press.

Berger, J. O. 1980. *Statistical decision theory, foundations, concepts, and methods*. Springer-Verlag, New York.

Black, R. 2010. Copenhagen climate summit undone by 'arrogance'. *BBC News Online*, March.

Bloch, F. 1996. Sequential Formation of Coalitions in Games with Externalities and Fixed Payoff Division. *Games and Economic Behavior*, **14**, 90–123.

Boadway, R., Song, Z., & Tremblay, J.-F. 2009 (May). *The Efficiency of Voluntary Pollution Abatement when Countries can Commit*. Working Papers 1205. Queen's University, Department of Economics.

Bråten, J., & Golombek, R. 1998. OPEC's Response to International Climate Agreements. *Environmental and Resource Economics*, **12**, 425–442.

Brennan, G. 2009. Climate change: a rational choice politics view. *The Australian Journal of Agricultural and Resource Economics*, **53**(3), 309–326.

Buchner, B., & Carraro, C. 2006. Parallel Climate Blocs: Incentives to Cooperation in International Climate Negotiations. *In:* Guesnerie, R., & Tulkens, H. (eds), *The Design of Climate Policy Conference Volume of the 6th CESIfo Venice Summer Institute*. MIT Press, Cambridge, Massachusetts.

Carraro, C., & Moriconi, F. 1997. International Games on Climate Change Control. *Fondazione Eni Enrico Mattei.*

Carraro, C., & Siniscalco, D. 1993. Strategies for the international protection of the environment. *Journal of Public Economics*, 309–328.

Chander, P., & Tulkens, H. 1997. The Core of an Economy with Multilateral Environmental Externalities. *International Journal of Game Theory*, **26**(3), 379–401.

Chander, P., & Tulkens, H. 2008. Cooperation, Stability, and Self-enforcement in International Environmental Agreements. *Pages 165–186 of:* Guesnerie, R., & Tulkens, H. (eds), *The Design of Climate Policy.* MIT Press, Cambridge, Mass.

Dasgupta, P., Hammond, P., & Maskin, E. 1980. On Imperfect Information and Optimal Pollution Control. *Review of Economic Studies*, **47**, 857–860.

de Clippel, G, & Serrano, R. 2008. *Bargaining, Coalitions and Externalities: A Comment on Maskin.* Working Paper. Brown University.

Farrell, J., & Maskin, E. 1989. Renegotiation in Repeated Games. *Games and Economic Behaviour*, **1**, 327–160.

Fehr, E., & Gächter, S. 2000. Fairness and Retaliation: The Economics of Reciprocity. *The Journal of Economic Perspectives*, **14**(3), 159–181.

Finus, M. 2001. *Game theory and international environmental cooperation.* Edward Elgar, Cheltenman, UK ; Northampton, MA.

Finus, M. 2003. Stability and design of international environmental agreements: the case of transboundary pollution. *Pages 82–158 of:* Folmer, H., & Tietenberg, T. (eds), *International yearbook of environmental and resource economics 2003/4.* Edward Elgar, Cheltenman, UK.

Finus, M., & Rundshagen, B. 2003. Endogenous Coalition Formation in Global Pollution Control: A Partition Function Approach. *Pages 199–244 of:* Carraro, C. (ed), *The Endogenous Formation of Economic Coalitions.* Edward Elgar, Cheltenman, UK.

Garnaut, R. 2008a. *The Garnaut climate change review : final report.* Cambridge University Press, Port Melbourne, Vic. :.

Garnaut, R. 2008b. *Targets and Trajectories.* Supplementary Draft Report. Commonwealth of Australia.

Gerber, A., & Wichardt, P. C. 2009. Providing public goods in the absence of strong institutions. *Journal of Public Economics*, **93**, 429–439.

Güth, W., Schmittberger, R., & Schwarze, B. 1982. An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior and Organization*, **3**, 367–388.

Hart, S., & Kurz, M. 1983. Endogenous Formation of Coalitions. *Econometrica*, **51**(4), 1047–1064.

Hoel, M. 1994. Efficient Climate Policy in the Presence of Free Riders. *Journal of Environmental Economics and Management*, **27**, 259–274.

Jackson, M. O. 2001. A crash course in implementation theory. *Social Choice and Welfare*, **18**, 655–708.

Jotzo, F., & Betz, R. 2009. Australias emissions trading scheme: opportunities and obstacles for linking. *Climate Policy*, **9**, 402–414.

Lessmann, K., Marschinski, R., & Edenhoffer, O. 2009. The effects of Tariffs on Coalition Formation in a Dynamic Global Warming Game. *Economic Modelling*, **26**(3), 641–649.

Little, A. 2010. Copenhagen Accord is the priority, says U.S. climate envoy. But what about a binding treaty? *Grist Magazine*, January.

Marks, M., & Croson, R. 1998. Alternative rebate rules in the provision of a threshold public good: An experimental investigation. *Journal of Public Economics*, **67**, 195–220.

Maskin, E. 2003. *Bargaining, Coalitions, and Externalities.* Presidential Address to the Econometric Society. Institute for Advanced Study, Princeton.

Meyer, A. 2000. *Contraction and Convergence: The Global Solution to Climate Change.* Green Books for the Schumacher Society, Totnes, U.K.

Montero, J.-P. 2007a. An Auction Mechanism for the Commons: Some Extensions. *Cuadernos de Economa (Latin American Journal of Economics)*, **44**(130), 141–150.

Montero, J.-P. 2007b. An auction mechanism in a climate policy architecture. *Pages 327–339 of:* Aldy, Joseph E., & Stavins, Robert N. (eds), *Architectures for Agreement: Addressing Global Climate Change in the Post-Kyoto World.* Cambridge University Press, New York.

Montero, J.-P. 2008. A Simple Auction Mechanism for the Optimal Allocation of the Commons. *American Economic Review*, **98**(1), 496–518.

Moore, J., & Repullo, R. 1988. Subgame Perfect Implementation. *Econometrica*, **56**(5), 1191–1220.

Nash, J. 1950. The Bargaining Problem. *Econometrica*, **18**(2), 155–162.

Nash, J. 1953. Two-person Cooperative Games. *Econometrica*, **21**(1), 128–140.

Ok, E. 2007. *Real Analysis with Economic Applications.* Princeton: Princeton University Press.

Okada, A., & Winter, E. 2002. A Non-cooperative Axiomatization of the Core. *Theory and Decision*, **53**(1), 1–28.

Osborne, M. J. 2003. *An Introduction to Game Theory.* Oxford University Press, USA.

Osborne, M. J., & Rubinstein, A. 1994. *A Course in Game Theory.* The MIT Press.

Ostrom, E. 2009. *A Polycentric Approach for Coping with Climate Change.* Policy Research Working Paper 5095. World Bank.

Pan, J., Chen, Y., Wang, W., & Li, C. 2000. *Carbon Budget Proposal: Global Emissions under Carbon Budget Constraint on an Individual Basis for an Equitable and Sustainable Post-2012 International Climate Regime.* Working Paper. Research Centre for Sustainable Development, Chinese Academy of Social Sciences.

Project Team of the Development Research Centre of the State Council, People's Republic of China. 2009. Greenhouse gas emission reduction: A theoretical framework and global solution. *Pages 389–408 of:* Garnaut, R., Song, L., & Woo, W. T. (eds), *China's New Place in a World in Crisis.* Canberra, Australia: ANU E Press.

Ray, D., & Vohra, R. 1997. Equilibrium Binding Agreements. *Journal of Economic Theory*, **73**, 30–78.

Roberts, M. J., & M. Spence, M. 1976. Effluent Charges and Licences under Uncertainty. *Journal of Public Economics*, 193–208.

Rubinstein, A. 1982. Perfect Equilibrium in a Bargaining Model. *Econometrica*, **50**(1), 97–109.

Schelling, T. C. 1970. *The Strategy of Conflict.* Harvard University Press.

Serrano, R. 1995. A Market to Implement the Core. *Journal of Economic Theory*, **67**(1), 285–294.

Serrano, R. 1997. A comment on the Nash program and the theory of implementation. *Economics Letters*, **55**(2), 203–208.

Stern, N. 2006. *The economics of climate change : the Stern review.* Cambridge, UK ; New York : Cambridge University Press.

Stern, N. 2009. *The Global Deal : Climate Change and the Creation of a New Era of Progress and Prosperity.* PublicAffairs ; New York.

Tamiotti, L., Olhoff, A., Teh, R., Simmons, B., Kulaçoğlu, V., & Abaza, H. 2009. *Trade and Climate Change.* Tech. rept. World Trade Organisation and United Nations Environment Programme.

Tulkens, H. 1998. Cooperation versus free-riding in international environmental affairs: two approaches. *Pages 30–44 of:* Hanley, N., & Folmer, H. (eds), *Game Theory and the Environment.* Edward Elgar, Cheltenham, England.

Uzawa, H. 2003. Global Warming as a Cooperative Game. *Pages 193–239 of:* Uzawa, Hirofumi (ed), *Economic Theory and Global Warming.* Cambridge University Press, New York.

Victor, D. G. 2007. Fragmented carbon markets and reluctant nations: implications for the design of effective architectures. *Pages 133–160 of:* Aldy, Joseph E., & Stavins, Robert N. (eds), *Architectures for Agreement: Addressing Global Climate Change in the Post-Kyoto World.* Cambridge University Press, New York.

Weitzman, M. 2009. On Modeling and Interpreting the Economics of Catastrophic Climate Change. *The Review of Economics and Statistics*, **91**(1).

Wood, P. J., & Jotzo, F. 2009. *Price Floors for Emissions Trading.* Research Report. Environmental Economics Research Hub, The Australian National University.

Yi, S. S. 1997. Stable Coalition Structures with Externalities. *Games and Economic Behaviour*, **20**, 201–237.

Yi, S.-S., & Shin, H. 2000. Endogenous formation of research coalitions with spillovers. *International Journal of Industrial Organization*, **18**(2), 229 – 256.