Hope or Despair? Formal Models of Climate Cooperation

Jon Hovi, Hugh Ward, and Frank Grundig

1. Introduction

In the aftermath of the Copenhagen (2009), Cancun (2010), Durban (2011), Doha (2012), and Warsaw (2013) meetings, it is difficult to be optimistic about our chances of avoiding increases in global average temperature and associated changes to the climate that will not only damage many communities and countries but also destroy species and eco-systems.[1] In particular, current attempts to construct a regime to deal with climate change hardly seem capable of averting potentially tragic collective action failure. It is widely accepted that the main factor causing such collective action failure is states' calculations about the degree to which it is rational for them to reduce their emissions of greenhouse gases or take other measures that might slow down climate change. Such calculations are understood to be based on economic factors such as perceived loss of competitiveness and opportunities foregone for economic growth, political factors such as the degree of electoral support and the power of the carbon lobby, and institutional factors governing the degree to which pressure from the electorate and organized groups translates into policy. Such factors seem to explain some of the variance in states' willingness to make commitments on climate change (Harrison and McIntosh-Sundstrom 2010). From the perspective of collective action theory, the basic problems are that states (1) do not factor in the spillover effects their actions have on other states and (2) heavily discount future benefits. As a result, they tend to underinvest in

[1] This article is a revised and significantly expanded version of Grundig et al. (2014).

reducing emissions and, because controlling climate change is a global public good, to free-ride on the actions taken by others (Ward 1996; Sandler 1997; Barrett 1999). The efficient Coasian solution, based on informal bargaining from a pre-existing distribution of enforceable property rights is inapplicable, because no such rights over climate stability exist in international law.

The idea that we are witnessing a collective action failure over climate change, together with the idea that states choose efficient means to attain economic and political goals that are largely independent of the domestic and international political processes and institutions built around climate change, make the application of rational choice models relevant. The rational choice approach has been widely criticized, though. In particular, some liberal institutionalists argue that this approach ignores the way political processes, institutions, and ideas frame the possibilities of efficient contracting and, perhaps more significantly, alter states' underlying preferences and the way they perceive problems (Young 1999). To be sure, the fact that climate change is on the international agenda is a clear indication that ideas matter. For good, or quite possibly for ill (Victor 2011), the institutional architecture of the Kyoto Protocol influences developments. The influence of international processes on voters and corporate interests has contributed to the willingness of some states (and notably the EU) to adopt a frontrunner role in terms of climate policy. Nevertheless, climate change illustrates the power of the rational choice approach.

Though the idea of collective action failure seems a powerful one, an important issue is whether we need to develop more sophisticated models. After all, the basic understanding of collective action failure in public good games goes right back to the work of Olson (1965). However, we do need to press further. Even if states' underlying goals are fixed, the best policies for attaining those goals will often depend on what other states do, particularly for the

largest emitters. The importance of such strategic interdependence and the use of game-theoretic tools to model it were introduced into the international relations literature on collective action in the 1980s (Axelrod 1984; Axelrod and Keohane 1985; Keohane 1984). In cases where interaction can be characterized as a game, states' underlying preferences cannot be read off directly from their actions, as is often assumed in accounts of climate change politics found in the policy studies and comparative politics literatures. In a one-shot, two-player Prisoner's Dilemma, it is obvious what a rational player will do, because each player has a dominant strategy of non-cooperation. However, it is likely that such dominant strategies do not exist in climate change politics (Dasgupta and Heal 1980, 21). Nor do such strategies exist in most of the models we review in this article. Game theory helps us understand how states behave in strategically complex environments.

In section 2, we describe some rather discouraging results offered by the formal modeling literature about the prospects for effective climate cooperation. The models we consider highlight the importance of incentives, while downplaying the role of other factors, such as social norms. It is therefore pertinent to say that we agree that norms can sometimes "move mountains" (Barrett 2007, 21). However, incentives are surely important, too. In section 3, we go on to consider some formal models suggesting that there might be light at the end of the tunnel after all. Finally, in section 4 we conclude by outlining three general lessons about international climate cooperation and six more specific lessons about treaty design that can be derived from models reviewed in this article.

Given the large number of models we consider, space does not permit us to describe and analyze each model in detail. Instead we offer a fairly rough sketch of each model and outline the main results. We hope readers will get the main message from each model and encourage

them to consult the referenced literature for more details on models they find particularly interesting.

2. Some Discouraging Results on Climate Change Cooperation

A climate agreement will be effective to the extent that it successfully mitigates anthropogenic climate change. Thus, effectiveness requires broad (and stable) participation, deep commitments by the participating countries, and high compliance rates. Importantly, *all* of these three requirements must be met; meeting only one or two of them is of little or no help (Barrett and Stavins 2003).

Because the avoidance of climate change is a global public good, countries have an incentive to free ride on other countries' mitigation efforts. This will generally cause suboptimal global abatement levels (Finus 2001; Finus and Rundshagen 2003). In the case of the Kyoto protocol (Kyoto 1), at least five different forms of free riding may be distinguished (Hovi et al. 2013). First, a few countries − most importantly the United States − never ratified. Second, Canada ratified the agreement (December 2002) and thus participated initially, but later gave notice of its withdrawal (December 2011). Third, the non-Annex I countries (developing countries) ratified without a legally binding emissions limitation target. Fourth, several East European countries ratified with a legally binding but very lax commitment (the "hot air" problem). Finally, there is the possibility of noncompliance. At the time of writing it is not yet clear whether all remaining Annex I countries actually met their targets for the first commitment period.

The problem of suboptimal global abatement has been addressed by a significant amount of formal work (prominent examples include Barrett 1994, 1999, 2002, 2003; Carraro and

Siniscalco 1992, 1993; Hoel 1992; Tulkens 1979). This formal work has aimed at identifying the conditions for formation of multilateral agreements (or coalitions) that are stable or self-enforcing.[2] The need for self-enforcement (or stability) originates from the anarchic character of the international system, meaning that no supranational authority can be relied upon for enforcement, a point agreed to by both liberal institutionalists and realists.

2.1 Coalition models

Models of international climate cooperation differ both with respect to which stability concept they use and in the way they specify countries' payoff functions. However, they can largely be categorized into two major types − coalition models and repeated-game models (Finus and Rundshagen 2003).[3] Coalition models aim to analyze the conditions for the formation of stable coalitions. A coalition is said to be internally stable if no member can benefit by exiting. Similarly, it is externally stable if no nonmember can benefit by joining.

Carraro and Marchori (2003) distinguish three main coalition formation rules.[4] The open membership rule specifies that each country is free to decide whether it will join or leave the coalition; hence, the coalition accepts as a member any country that wishes to join (e.g. Carraro and Siniscalco 1993). The exclusive membership rule means that a consensus among the existing members is required for a country to join; however, each country is free to exit the coalition (e.g. Yi and Shin 2000). Finally, under the coalition unanimity rule the formation

[2] For a recent discussion of the notion of a self-enforcing agreement, see Grundig et al. (2012).

[3] Most existing models (of both types) picture countries as unitary actors. Notable exceptions include Dietz et al. (2012) and Ward et al. (2001).

[4] Finus and Rundshagen (2009) distinguish and consider the effects of six different rules for coalition formation.

of a coalition requires a consensus among its members, meaning (1) that players are not free to join the coalition and (2) that if a country leaves, the coalition will cease to exist (e.g. Chander and Tulkens 1993).

These three rules entail different incentives for countries. In the setting of a global public good, the open membership rule entails strong incentives for free riding that resemble those found in the Kyoto process. In contrast, the exclusive membership rule resembles the requirements for accession to the EU or the WTO, and may relate to Victor's (2011) vision of a carbon club. Finally, the coalition unanimity rule can have a disciplining effect on countries, because it makes certain types of free riding (e.g. free riding by withdrawal) difficult.

Coalition models depict international cooperation as a two-stage game. At stage 1, countries choose whether to participate (i.e., whether to be a signatory or a nonsignatory). At stage 2, they choose their abatement level (Carraro and Siniscalco 1993; Chander and Tulkens 1992; Hoel 1992).[5] Different models make different assumptions about behavior at each stage (Finus 2008). We here focus on a basic version of the coalition model, while we consider a number of extensions in section 3.1.

At stage 1, countries choose simultaneously (i.e., without knowing other countries' choices) whether they will become a coalition member or not. The basic model assumes that only a single coalition (agreement) can be formed, so countries that do not join have no option but to act individually. Finally, the model assumes open membership, meaning that no country can be barred from joining the coalition if it wants to.

At stage 2, the countries *in* the coalition choose their abatement levels jointly, aiming to maximize the combined payoff of all coalition members. Thus, the coalition internalizes the

[5] In some models nonparticipating countries independently choose their abatement levels in stage three (e.g. Barrett 2005; Nyborg 2014).

external effects across members but not the external effects caused by members on nonmembers. In contrast, each country *not* in the coalition chooses the abatement level that maximizes its individual payoff. The model assumes that no side payments or issue linkages occur, and that decisions are made exclusively on the basis of material costs and benefits. Finally, these costs and benefits are assumed to be known with certainty (Finus 2008, 36).

Coalition models are analyzed by backward induction, meaning that the analyst considers stage 2 before turning to stage 1. Thus, to decide what to do in stage 1, countries must map the implications that each stage 1 option will have at stage 2.

The basic model provides very pessimistic predictions for climate-change cooperation. The reason is that stable coalitions are typically (very) small. Barrett (2005) provides a simple illustration, based on a linear benefit function and a quadratic cost function. Stability then requires that the coalition has exactly three members. This equilibrium is unique with regard to the number of member countries, but not with regard to their identities; as countries are assumed to be identical, any coalition consisting of exactly three countries will be stable. In a world with nearly 200 countries, a coalition of only three countries obviously cannot achieve very much.[6]

2.2 Repeated-game models

Whereas coalition models typically focus on participation, repeated-game models typically focus on compliance; they aim to analyze the conditions under which countries that participate in a climate agreement will meet their commitments.

[6] Several extensions and modifications of the basic model generate more optimistic predictions; see section 3.1.

A game is repeated if it can be reduced to a series of iterations of some smaller game. Applications of repeated-game models to climate change cooperation typically center on the infinitely repeated N-player Prisoner's Dilemma (e.g. Asheim et al. 2006; Barrett 1999; 2003; Finus and Rundshagen 1998; Hovi and Froyn 2008) or some other set-up that resembles the Prisoner's Dilemma (e.g. Asheim and Holtsmark 2009; Heitzig et al. 2011; Kratzsch et al. 2012).

Before the game begins, countries are assumed to enter into an agreement that the parties must enforce throughout the game using credible threats (e.g., Barrett 1999, 2003). The structure of the game enables a country to obtain (at least short-term) net gains by free riding; hence, the agreement must specify a strategy that can enforce compliance.

In most repeated-game models the only leverage available to a country is to threaten retaliation in kind, i.e. to respond to noncompliance by either terminating (permanently) or suspending (temporarily) its own commitment. Most older and some more recent repeated-game literature assumes that cooperation is based on the so-called Grim Trigger strategy, by which even a single case of noncompliance will cause termination of the agreement. However, terminating an agreement will entail that the future gains from cooperation will be lost, thereby harming noncompliant countries and compliant countries alike. Thus, compliant countries will be better off if they abstain from implementing this punishment and simply resume cooperation as if no noncompliance had occurred. A country contemplating noncompliance will rationally foresee this possibility of renegotiation, meaning that the agreement's stability will be undermined.

To deter noncompliance an agreement must therefore be based on a strategy that can sustain the agreement as a renegotiation-proof equilibrium. While several notions of renegotiation proofness exist, most applications to climate change cooperation use Farrell and

Maskin's (1989) notion of weakly renegotiation-proof equilibrium, which requires that not all players be strictly worse off by carrying out the punishment than by renegotiating. In other words, it must be in each party's best interest to conform to the specified strategy and it must be in at least some country's best interest to decline an invitation to renegotiate should a deviation from this strategy occur (Barrett 1999, 2003; Finus and Rundshagen 2003).

Most of the recent repeated-game literature ensures (weak) renegotiation proofness by replacing the Grim Trigger strategy with some other strategy, typically one that prescribes less severe punishment for noncompliance. For example, Barrett (1999) uses a strategy he calls Getting Even and Asheim et al. (2006) use a strategy they call Penance. These two strategies resemble each other in that they both prescribe that a noncompliant country must endure punishment (pay penance) in one period of the repeated game before cooperation can be restarted. Assuming that countries do not discount future payoffs too heavily, these strategies ensure (1) that it is in the noncompliant country's best interest to accept the punishment (because accepting the punishment will cause cooperation to be restarted after one period of punishment and will then last indefinitely unless another case of noncompliance occurs) and (2) that it is in some other country's (or countries') best interest to insist that the punishment be carried out before cooperation can be restarted.

Early repeated-game models teach us that a climate treaty with broad participation and deep commitments is unlikely to be self-enforcing (in the sense of being weakly renegotiation proof). The reason is that renegotiation becomes more attractive the larger the number of parties. Suppose that each country faces only two options in each period, abate and not abate. Suppose furthermore that if member country j fails to abate in a given period, then the agreement requires country j to pay penance in the next period by playing abate while *all* other member countries are allowed to play not abate. By implementing this punishment the

punishing countries will obtain the not abate payoff resulting from just one (other) country playing abate. In contrast, by renegotiating they will obtain the payoff associated with the outcome where *all* member countries play abate. The latter payoff is an increasing function of the number of member countries; thus, the larger the number of member countries, the less likely that the agreement will be renegotiation proof. In other words, the agreement can be renegotiation proof only for a moderate number of participating countries.

Using a repeated-game model, Barrett (2002) demonstrates that it is possible to construct what he calls a consensus treaty, i.e. a global agreement where all countries participate. However, he finds that a consensus treaty is possible only if commitments are unambitious ("shallow"). This finding suggests that a trade-off exists between the depth and the breadth of an agreement.

More generally, infinitely repeated games have infinitely many equilibria if the rate at which players discount future benefits is low enough and the punishment for players breaking away from conditionally cooperative behavior is credible and great enough. Players' interests generally diverge among the set of efficient equilibria. For example, one equilibrium may involve country A delaying the start of cooperation while country B starts immediately, while another may reverse this pattern. Games with divergent preferences over equilibria generate incentives for players to pre-commit to doing less, as in the paradigmatic examples of Chicken and the Battle of the Sexes. Generally, the literature has not dealt with the dangers that arise from multiple equilibria where credible commitments can be made, but these are real dangers in climate change politics where countries can use (and perhaps overstate) the strength of domestic vetoes in order to commit to doing little. If credible, such commitments might cause others to do more, with the danger of a "collision" if too many countries pre-commit (Ward 1996).

Despite the attention on renegotiation-proofness in the literature, relatively little has been written about how to theorize the actual process of climate-change negotiations. However, important exceptions do exist. For example, scholars have analyzed the implications for coalition formation of various aspects of the bargaining process, including North-South bargaining, uncertainty concerning the risk of catastrophic climate change, and the choice of policy instruments. Some of these scholars have used analytical models (Finus and Rundshagen, 1998; Altamirano-Cabrera et al. 2008; Caparrós et al. 2004; Caparrós and Pereau 2013; Urpelainen 2012a), whereas others have used game-theoretically based experiments (e.g. Barrett and Dannenberg 2012).

Noting that multiple efficient ways of sharing the burden of dealing with climate change could be stable under conditional strategies in an infinitely repeated game, Grundig et al. (2001) suggest the application of the Nash-Rubinstein bargaining solution to predict which of these efficient and stable patterns will eventuate. This approach highlights breakdown payoffs and discount factors as determinants of the size of the burden a country will have to take up; however, it says little or nothing about negotiation dynamics. Putnam (1988) argues that international negotiation is deeply affected by the fact that negotiators have to get the consent of domestic veto players before they can settle. In particular, a negotiator who is weak domestically might obtain a good deal by claiming that domestic players would not agree to anything less – a role the US Congress has often played in climate negotiations. Concerning the possibility of bilateral climate deals between the EU and developing countries, Mansfield et al. (2007) show that the likelihood of bilateral deals generally go down with the number of domestic vetoes.

2.3 Other approaches

The pessimistic conclusions reached by the coalition literature and the repeated-game literature sit well with inferences drawn by literature studying the conditions for international cooperation more generally. For example, the so-called law of the least ambitious program states that what can be achieved through international cooperation is limited to the platform advocated by the least enthusiastic party (Underdal 1980, 1998; see also Finus and Rundshagen 1998; Altamirano-Cabrera et al. 2008). The reason is that treaty formation usually requires unanimity (or consensus) among the participating countries, which enables the least enthusiastic party to veto any proposal that is more ambitious than its own. In principle, it might be possible to move beyond the least ambitious platform simply by accepting nonparticipation by the least ambitious country. Notice, however, that the resulting agreement will then still be limited to the platform preferred by the least ambitious *remaining* party. Moreover, in the case of global climate-change cooperation it will likely be difficult to move significantly beyond the least ambitious program by excluding the least ambitious party or parties. One reason is that the world's largest emitters, China and the United States, are among the least ambitious countries. Clearly, if these two countries are omitted, the resulting agreement cannot be effective. Another reason is that the climate negotiations take place within the institutional structure of the UNFCCC, where decisions are made by consensus. China has repeatedly made it clear that it is unwilling to negotiate over climate change cooperation in any other forum.

Even more pessimistic concerning the potential for effective cooperation is the relative-gains literature, a branch of neorealist theory. This literature argues that states' concern with relative gains may further constrain or even completely eliminate the potential for international cooperation (Snidal 1991). Grundig (2006) argues that relative-gains concerns are particularly important in cases that combine significant economic costs and a non-excludable good. Thus, concerns with relative gains might help explain why cooperation to

mitigate climate change (high costs and a non-excludable good) has been far less successful than cooperation to avoid ozone depletion (non-excludable good but only moderate costs) and cooperation on international trade (high costs but excludable good).

3. Light at the end of the tunnel?

Our presentation thus far suggests that the prospects for solving the climate-change problem through international cooperation are very bleak indeed. Unfortunately, these depressing predictions correspond well with the lack of actual progress in the climate negotiations so far. Nevertheless, in this section we ask if there may be some light in the tunnel after all. Existing formal models offer several glimpses of hope. First, the coalition literature provides several ideas for enhancing cooperation by making large(r) coalitions stable. Second, a branch of the repeated-game literature suggests that it might in fact be possible to design a renegotiation-proof climate agreement with broad (or even full) participation. Third, a string of papers have studied the potential of deposit-refund systems to enhance cooperation. Fourth, while the law of the least ambitious program clearly pinpoints some severe constraints on the prospects for cooperation, some scholars have suggested that this law nevertheless has its limits. Fifth, countries may use cooperative probes to build trust. Sixth, some scholars argue that cooperation might emerge in a completely decentralized fashion. Finally, the results from game-theoretically oriented experiments indicate that the prospects for cooperation might be better than the formal results mentioned in section 2 lead us to believe.

3.1. Making larger coalitions stable

We noted in section 2.2 that stable coalitions are typically (very) small, at least within the framework provided by what we have termed the basic version of the coalition model. An

obvious question is therefore whether modifying the assumptions of this basic version might produce more optimistic predictions concerning participation. A number of such modifications have been explored, with varying degrees of success. We here focus on seven – issue linkages, trade restrictions, multiple coalitions, minimum participation clauses, modest targets for emissions reductions, asymmetric countries, and alternative motivation such as equity concerns or strong reciprocity.

Carraro and Siniscalco (1997) use a coalition model in which cooperation on climate change is linked to cooperation on technology R&D. The idea is that such linkage may increase participation, assuming that the fruits of technology R&D is a club good, so that countries that do not participate in climate-change cooperation may be excluded from sharing the fruits of technology R&D.[7] Carraro and Siniscalco show that, given this crucial assumption, full cooperation is possible even with a very large number of countries. However, noting that such linkage is rare in international environmental agreements, others have questioned whether restricting the fruits of technology R&D to countries that cooperate on climate change is possible or if possible, in signatories' best interest (e.g. Barrett 2005).

Kim and Urpelainen (2013) consider whether technology competition in relation to emissions reductions affects states' willingness to cooperate. In stage one, two states decide whether to cooperate over climate change. In stage two, they decide a subsidy level for their carbon reducing sector. States care both about climate change, industry profits (driven by relative competitiveness) and costs to consumers. In symmetric equilibria, the states cooperate in stage one if (1) the costs of subsidization are low enough *and* the states do not care too much about relative competitiveness *or* (2) they care a lot about competitiveness but costs are low enough and profits are high enough with large enough subsidies.

[7] This idea resembles Victor's (2011) idea of a "carbon club".

A second popular suggestion for increasing participation is that signatories impose trade restrictions on non-signatories. Consider a climate treaty that requires signatories to trade only with other signatories. Barrett (1997, 2005) finds that this requirement changes the game into Assurance. If few other countries participate in the treaty, the free rider incentive will dominate the cost of being excluded from trade with signatories, thereby making participation unattractive. However, if sufficiently many other countries participate, the cost from the trade restrictions will dominate the free-rider incentive; in particular, participation will be attractive for every country when all other countries are signatories. If this account is accurate, all that is needed to ensure a stable coalition with full participation is a simple clause stating that the treaty will enter into force only after the critical number of countries have ratified. For such a clause to solve the problem, however, the threat to exclude non-signatories from trade with signatories must be credible. As trade restrictions are often costly for both sides, it is far from obvious that this requirement is actually met (Barrett 1999, Aakre 2013).

A special type of trade restrictions is so-called border-tax adjustments (BTAs), whereby the importing country (say, country A) imposes a CO_2 tax on imported products. The CO_2 tax due for a particular imported product equals the tax due for equivalent products produced in country A. Similarly, exporters from country A receive a refund of the CO_2 tax it paid in country A during the production process (Ismer and Neuhoff 2007). Some scholars claim that BTAs are more credible than more extensive trade restrictions are; however, they are arguably also quite impractical: Calculating the CO_2 emitted during the production of every exported or imported item will likely be cumbersome (Barrett 2003: 388).

Third, several coalition models have modified the basic model's assumption that only a single coalition is possible, thereby opening up for the possibility that more than one climate agreement may be negotiated. Typically, this modification leads to more than one coalition in

equilibrium and thus also to more overall cooperation than in the basic model. As a result, global welfare is usually higher with multiple agreements than with a single global accord (Bloch 1997; Carraro 1999, 2000). Subsequent studies of single versus multiple coalitions that use different assumptions concerning membership rules include Carraro and Marchiori (2003), Finus (2003), and Finus and Rundshagen (2003). These studies support the conclusion that multiple coalitions may be superior to a single coalition.[8]

Fourth, some scholars have studied the impact of adding a stage that precedes the coalition and policy stages. At this first stage, countries choose the minimum number of countries that must participate if the treaty is to enter into force. This decision about the entry-into-force clause is taken on the basis of countries' anticipation of the decision's implications at the second and third stages of the game. Assuming that adoption of an entry-into-force clause requires unanimity, Carraro et al. (2009) show that the presence of such a clause increases the equilibrium number of signatories, even when the clause is endogenous.

Fifth, Finus and Maus (2008) argue that an important reason why only small coalitions are stable is the assumption that coalition members maximize their joint welfare when choosing their emissions in the second stage. The result is ambitious emission reduction targets and high incentives for free riding. They therefore consider the possibility of an agreement based on modest emissions reductions. Interestingly, they find that modesty might pay: It attracts higher participation levels and these higher participation levels may well compensate for less ambitious emissions reduction targets. In short, introducing more modest targets may cause global emissions to decrease in equilibrium.

Sixth, while the basic model assumes that all countries are identical, some scholars have considered the impact of introducing asymmetric countries. With asymmetric countries,

[8] A similar conclusion is reached by Asheim et al. (2006), using a repeated-game framework (see section 3.2).

additional coalitions may become stable through side payments. The reason is that countries that gain a lot from an agreement might be able to compensate countries that would not otherwise benefit from this agreement, thereby enticing the latter countries to join the coalition (and to remain members). Thus, a coalition no longer needs to be internally stable; it suffices that it is *potentially* internally stable (Carraro et al. 2006). The latter condition is met if the sum of benefits generated by the coalition is sufficiently large to permit a redistribution (through side payments) that leaves every member country better off inside the coalition than outside it. Since the set of internally stable coalitions is a subset of the set of potentially internally stable coalitions, it follows that larger (internally) stable coalitions may be possible when countries are asymmetric, provided that side payments are feasible (Holtsmark 2013; McGinty 2007). A major advantage of side payments is that they make it possible to break the link between actual abatement measures and economic burden sharing (McGinty 2014).

Finally, some scholars have begun studying how and to what extent alternative motivations, such as concerns about fairness, might influence participation in international climate agreements. For example, Lange (2006) shows that inequality aversion concerning differences in developed countries' abatement targets can make larger coalitions stable and cause stricter abatement. Similarly, Lange and Vogt (2003) find that in an N-country Prisoner's Dilemma, cooperation by most or even all countries can constitute a Nash equilibrium if countries have a preference for equity. A similar result is obtained by Nyborg (2014), assuming that some or all countries are reciprocators rather than rational actors motivated by self-interest. However, Lange and Vogt (2003) also find that if countries are able to choose their abatement level from a continuum (instead of simply facing a binary choice between cooperate (abate) and defect (not abate), a preference for equity provides no improvement from the usual suboptimal Nash equilibrium.

While Lange and Vogt (2003) consider equity to be a matter of welfare comparisons, Grüning and Peters (2010) consider it to be a matter of observable measures such as abatement activities. A main finding in their model is that the policies of countries participating in a climate agreement tend to converge. This result is driven by the assumption that countries have a reluctance to pursue policies that are very different from those pursued by other participating countries.

3.2 Making renegotiation-proof climate agreements consistent with broad or even full participation

A series of recent articles have questioned the claim that a weakly renegotiation-proof agreement must necessarily entail either moderate participation or shallow commitments. Asheim et al. (2006) show that multiple (e.g., regional) agreements can enhance participation even when the depth of cooperation is taken as a given. Their model builds on Barrett (1999), but admits not only the possibility of negotiating a single global agreement but also that of negotiating two regional agreements. Identifying upper and lower bounds on the number of participating countries in each case, they show that two agreements can sustain a higher number of countries than a single global agreement can. Moreover, they demonstrate that a climate regime based on two agreements Pareto dominates a regime based on a single global agreement. Thus, their results mirror those of Carraro (1999, 2000) and others, using a coalition model (see section 3.1).

Asheim et al.'s (2006) model follows Barrett's (1999) in that noncompliance must be punished by all other participating countries in the perpetrator's *own region*. Although this specification restricts the number of participating countries in each region, the existence of two agreements ensures that the total number of participating countries in the two agreements

combined becomes larger than the number of participating countries in a single global agreement.

Froyn and Hovi (2008) extend Asheim et al.'s (2006) analysis by showing that full participation can be sustained as a weakly renegotiation-proof equilibrium even in a single global agreement. They demonstrate that even when no possibility exists for reducing abatement levels (players face only a binary choice in their model), participation can be increased by limiting the punishment for noncompliance. While Barrett's Getting Even strategy allows *all* other participating countries to punish a noncompliant country, the strategy formulation used by Froyn and Hovi permits only a *subset* of the participating countries to punish. This strategy formulation makes it possible to study how the number of participating countries in equilibrium varies with the number of countries allowed to punish noncompliance. The authors provide lower and upper bounds on the number of punishing countries that is consistent with full participation (in weakly renegotiation-proof equilibrium).

The models studied by Barrett (1999), Asheim et al. (2006) and Froyn and Hovi (2008) assume that countries face a simple binary choice between cooperate (abate) and defect (not abate). Subsequent research has relaxed this assumption. Asheim and Holtsmark (2009) demonstrate that full participation is also possible when countries face a continuum of alternative emission levels. In their model, a Pareto-efficient climate agreement can always be implemented as a weakly renegotiation-proof equilibrium, provided that countries do not discount future payoffs too heavily. This result suggests that one need not choose between a narrow but deep agreement on one hand and a broad but shallow agreement on the other. However, Asheim and Holtsmark's results also demonstrate that designing an enforcement system that makes a broad and deep agreement possible is far from a trivial matter.

The results obtained by Froyn and Hovi (2008) and by Asheim and Holtsmark (2009) are supported by Heitzig et al. (2011), who propose an enforcement system based on a simple dynamic strategy of linear compensation. This strategy redistributes abatement obligations according to past compliance levels, while keeping the overall abatement level constant across periods. Heitzig et al. (2011) show that their strategy can be used to implement any given allocation of emissions reductions, thereby casting further doubt about the existence of a "narrow but deep" versus "broad but shallow" tradeoff.[9]

All models considered thus far in this section treat emissions as a flow variable. Thus, these models assume (usually implicitly) that emissions in a particular period have no lasting effect over time and that the set of possible per period payoffs remains constant over the entire repeated game. These assumptions are arguably implausible for applications to climate change cooperation. Kratzsch et al. (2012) invoke the more realistic assumption that greenhouse gas emissions build up a stock in the atmosphere over time and that it is the current stock that influences the climate. They show that broad or even full participation is possible also when the model takes into account that emitted gases is a stock variable that is depreciated only slowly over time, an important generalization of results from previous studies. Their model also enables them to identify certain effects that are difficult to spot when emissions are modeled as a flow variable. For example, they show that treaties with broad participation are more easily achieved for long-lasting gases than for short-lived ones. The reason is that long-lasting gases induce costs in more periods than short-lived gases do.

[9] This conclusion is also supported by Gilligan (2004). Using a multilateral bargaining model, he shows that such a trade-off does not exist for a wide class of cooperation problems. Gilligan traces the hypothesized broader-deeper trade-off to the assumption that the participating countries must fix their policies at an identical level. In his model, when the multilateral agreement permits the participating countries to fix their policies at different levels, the broader-deeper trade-off ceases to exist.

3.3. Deposit-refund Systems

Inspired by informal suggestions (Finus 2002; Finus 2008a; Gersbach 2006; Gersbach 2008) a string of papers have considered the possibility of using a deposit-refund system to enforce a new climate agreement.[10] The emerging literature includes formal models (Gerber and Wichardt 2009; Gersbach and Winkler 2007; McEvoy 2013) as well as informal analysis (Hovi et al. 2012) and experimental studies (Cherry and McEvoy 2013, McEvoy 2013). The exact design of a deposit-refund system must take into account the type of climate agreement in question. Here we outline and discuss a design for a cap-and-trade type of agreement. [11]

The basic idea is that each member country must deposit a significant amount of hard currency at ratification and make additional yearly deposits until the commitment period begins. Should a member country decline to make further required deposits or fail to meet its emissions limitation target, it will forfeit all or part of its existing deposits (depending on the degree of noncompliance). In contrast, a country that makes all required deposits and meets its target will receive a full refund when the commitment period ends.

As a tool for compliance enforcement, a deposit-refund system has several advantages (Hovi et al. 2012). First, it is simple. Whereas Kyoto's compliance enforcement system is fairly complex, it is straightforward to comprehend a system whereby noncompliance will entail a loss of deposits. Second, punishment does not require cooperation by the noncompliant country, because the climate regime will control the deposits. This provides another contrast to Kyoto's enforcement system, which is based on self-punishment (Barrett

[10] This subsection draws extensively on Hovi and Underdal (2014).

[11] For an application to a climate agreement based on carbon taxes, see Gersbach (2006) and Gersbach and Winkler (2007).

2003). Third, provided that deposits exceed compliance costs, fulfilling one's commitments will be better than being noncompliant *and* forfeiting one's deposits. Fourth, the threatened punishment will be credible, because punishing a non-compliant Party will benefit the other Parties individually as well as collectively. Finally, whereas under the Kyoto Protocol a non-compliant country could escape punishment by withdrawing from the treaty, a deposit system can easily be designed to make such escape infeasible. In particular, withdrawal before the commitment period ends might entail forfeiture of deposits.

In theory, deposit-refund systems may be designed to ensure participation as well as compliance. For example, in a symmetric setting (i.e., all countries are identical), the treaty could state that entry into force will take place only when all countries have ratified and made the required deposits. Such a clause would make free riding through nonparticipation infeasible. As a result, the relative-gains problem is also reduced. When free-riding through nonparticipation is infeasible, relative-gains concerns no longer provide a motive for non-participation − at least not in a symmetric setting.

In practice, however, a deposit-refund system is implausible as a solution to the participation problem. First, the climate-change problem is entangled in many and serious asymmetries (e.g. Victor 2011), which makes the requirement that all countries must participate extremely impractical: If even a single country declines to make required deposits, the treaty will never enter into force. Second, it may not be credible that if one country declines to make required deposits, other countries will abstain from cooperating among themselves. The incentive to participate and make deposits critically hinges on such credibility. Finally, countries facing serious liquidity problems may be particularly reluctant to participate in a climate treaty based on a deposit-refund system. Thus, while a deposit-

refund system could in principle work for compliance enforcement, it must be complemented by other measures (e.g., trade restrictions) for participation enforcement.

3.4. The Law of the Least Ambitious Program Does Not Always Apply

Around three quarters of the regimes coded by Breitmeier, Young, and Zürn (2006) operate on the basis of unanimity or consensus. Moreover, the Framework Convention on Climate Change operates by consensus, too. The Law of the Least Ambitious Program thus seems quite widely applicable. While many regimes formally operate using some version of qualified majority rule, at least on occasion (Hovi and Sprinz 2006), Underdal (1998) notes that the argument might still apply to the country most loath to see action among those vital to progress, for instance to the most loath member of a k-subgroup just large enough to make cooperation worthwhile if all its members cooperate. This would seem a potentially important argument given that perhaps no more than twelve states are crucial to making progress on climate change, based on their percentage of global emissions (Victor 2006). However, there are reasons to doubt that it *always* applies. First, if unanimity makes it difficult to ratchet up effectiveness beyond the level set by independent decision making when a regime is being built up, equally it makes it difficult to revert from an established policy to one with lower effectiveness (Hovi and Sprinz 2006). A "race to the bottom" is unlikely because it requires non-cooperative adjustments by industrialized countries. Second, a partial "race to the top" is likely because many emerging countries stand to gain from reduced negative externalities and the competitiveness problem is limited when the most lucrative export markets are already regulated. Finally, powerful industrialized countries with a high regulatory capacity benefit from a global expansion of regulation and opposition from veto players could be bought out through side payments to ensure unanimity (Barrett 2003; Ward et al. 2001).

In collective action games, side payments are transfers of private goods between players interacting over provision of a public good or concessions on policy dimensions relevant to such provision. Regarding climate change, an example of the former are transfers between the North and South under the Global Environment Facility; while an example of the latter are the flexibility mechanisms under the Kyoto Protocol demanded by the United States and some other countries (Ward et al. 2001). Side payments are usually seen as a trade between agents with different degrees of concern about an issue. Barrett (2003: 335-354) shows that asymmetries in the provision function can lead to stable arrangements where some countries are induced to cooperate through side payments. He considers an n-player Prisoner's Dilemma, where each member of a sub-group of $1 < k < n$ countries can benefit from cooperating so long as all the other members of the k-subgroup do so, but a group larger than k is unstable because for additional members costs exceed marginal benefits from environmental improvement. In the symmetric version of the game, side payments cannot induce extra cooperation. Suppose that one extra country is induced to join by a side payment; then any member of the original k-subgroup will have an incentive to defect, because there are now more than k cooperators and defecting enables it to avoid the cost of making side payments.

However, it is possible to induce more to join in an asymmetric game with two groups of countries, where an extra member in group 2 brings less marginal public good benefits than an extra member in group 1 does. It may then pay a member of the original k-subgroup to stick, given that it is a member of group 1. Defection would then reduce provision of the public good more than it has been increased through adding a member to group 2. Depending on the costs of making side payments and the degree of asymmetry between groups, equilibria in which k members of the first group cooperate and all members of the second group are induced to cooperate through side payments may be possible.

A number of models have provided further insights on how side payments may influence cooperation (e.g. Biancardi and Villani 2010; Carraro et al. 2006; Eyckmans and Finus 2004; Fuentes-Albero and Rubio 2010; Holtsmark 2013; McGinty 2007; McGinty 2011; Ward et al. 2001; Weikard, 2009). Because we considered some of this work in section 3.1 and also because of space constraints, we here focus on the model analyzed by Ward et al. (2001).

Unlike in Barrett's model, in Ward et al.'s model (2001) side payments can also be used to block progress towards a more effective regime. Ward et al. assume the existence of a status quo point on the effectiveness dimension and that any country can veto change. Countries can locate either on the progressive side of the status quo or on the opposite side. One subset of the countries is a progressive coalition, another subset is a laggard coalition, while the remaining countries are unattached. The progressive coalition can make side payments to buy out opposition to progress from unattached countries, but the laggard coalition can attempt to counter such attempted buy-out. The progressive coalition must be highly predominant in its ability to overcome the laggard coalition's efforts if progress is to come about, because the laggard coalition need only focus its attention on one veto, whereas the progressive coalition must bribe all unattached countries that are initially opposed to progress. Even if the progressive coalition is predominant enough to obtain some progress, the *degree* of progress will generally be limited. The progressive coalition member least eager on progress can generally ensure that its desired level of progress (or something close to it) is achieved by limiting its contributions to the funding of side payments.

Ward et al. (2001) re-instate a version of the Law of the Least Ambitious Programme: progress will likely be limited to what the least ambitious member of the progressive coalition wants – if it occurs at all. More work needs to be done in this area, though. While most other work ignores the possible role of side payments from a laggard coalition, Ward et al. (2001)

assume side payments are costless so as to highlight the most progressive equilibrium possible. Moreover, the benefits generated by the use of side payments are assumed to be non-excludable; for example, a country wanting progress could do nothing and still benefit from others making side payments to bring it about. The model does not deal with the issue of collective action failure over who will pay for side payments to be made.

3.5. Trust and networks can make a difference

In reality, country A may be uncertain about whether country B is the type that genuinely wishes to conditionally cooperate or the type that only pretends to have such wishes (Kydd 2007). If B's actions significantly affect A's payoffs, A may take a considerable risk in shifting energy paths, because it may take a long time to become sure that B is not reciprocating and then a long time to switch its own strategy. So initiating unilateral emission cuts, such as those under the EU's 20% emissions cuts (the 20 20 20 by 2020 policy), is a gamble.[12] Why take such a gamble?

Countries' background common knowledge of each other sets the a-priori probability that a country is of the type that actually wishes to conditionally cooperate in a game of incomplete information concerning others' types. As the game progresses, a country may choose to send a costly signal that indicates its type, because only countries of this type would make such a move in equilibrium. Others update their prior beliefs about the country sending the signal, using what they can infer from the signal and the equilibrium; and these updated beliefs support the (perfect Bayesian) equilibrium (Fudenberg and Tirole 1992, 207-241; Kydd 2007, 183-205). Perhaps the EU's "20 20 20 by 2020" policy is a signal of this sort.

[12] The Commission claims significant short-term economic benefits for the EU.

Urpelainen (2012) shows how states may signal whether they are the type that wishes to conditionally cooperate by imposing a tax on a domestic industry as long as the costs to the industry are large enough to deter the type of state that does not wish to cooperate from mimicking the signal. In international crises, states may attempt to reverse a potential escalation towards war by starting with relatively small cooperative gestures which, if reciprocated, lead to further de-escalatory steps (Osgood 1962). By starting small, states may both signal something about their type and learn a lot about others. A similar logic seems to underlie the design of the climate regime's institutional architecture, whereby countries started with a framework convention, aiming to gradually impose tighter standards through adding protocols.

In a game of this type, countries' prior beliefs can be thought of as the degree of background trust they have in each other, and such trust is vital to whether they will risk conditional cooperation (Kydd 2007). Trust can arise during specific negotiations through "cheap talk" (Fudenberg and Tirole 1992, 361-362; Ostrom et al. 1994), but it also arises through numerous interactions, including those in other issue domains than climate. When countries meet each other in the course of routine diplomacy, by direct contact they learn about each other's interests, capabilities and trustworthiness. They also create networks that enable them to learn about each other *indirectly* as information travels through the network. Countries that meet frequently, with many others and in many forums are central to the network, or highly embedded. They are in a position to learn most, but also to affect flows of information, giving them brokerage power (Hafner-Burton et al. 2009; Maoz 2010). Inter-state networks are supplemented by networks between non-state actors such as NGOs, corporations, and scientific bodies, such as those that have come to exist in climate governance (Andonova et al. 2009).

Although networks should not be conceptually confused with trust or social capital (Ostrom and Ahn 2008; cf. Dasgupta 2008), there is strong evidence at the individual level that dense networks are often associated with higher levels of trust. Beliefs about the trustworthiness of countries central to dense networks will have lower variance. Moreover, because a highly embedded country is more involved in a range of international interactions, it has more to lose if its reputation for trustworthiness is harmed by failure to reciprocate on a single issue such as climate change. They may also be more influenced by evolving patterns of international norms (Florini 1996) and information cascades causing states' perceptions of costs of action to decrease. There is also emerging empirical evidence from the environmental realm that that highly embedded countries act more cooperatively (Bernauer et al. 2010; Ward 2006).

3.6. Decentralized cooperation

The general pessimism in the coalition and repeated-game literatures is shared by several studies considering whether unilateral emissions reductions by one or a few countries may cause other countries to follow suit. For example, Hoel (1991) and Bucholz et al. (1998) find that unilateral emissions reductions will unlikely cause other countries to follow suit and could even cause them to *increase* their own emissions. According to their view, unilateral action is at best pointless and at worst counterproductive.

However, a few scholars have recently begun to question this pessimistic view of unilateral policies, arguing that unilateral action may be rational even for a government at the national, regional or local level (or even for an individual firm).[13] For example, Urpelainen (2009) suggests that ancillary local benefits at the national, regional or local level can

[13] For an excellent informal account of the emergence of climate policy in the United States, see Rabe (2004).

motivate unilateral emissions reductions.[14] Moreover, using a two-country, two-period model, Urpelainen (2011) shows how a 'green' government in country A may adopt a climate policy unilaterally in period 1, so as to bind a "brown" government in country A that could be elected in period 2. He also shows that this possibility is contingent both on the climate policy lowering mitigation costs and on what country B will do in period 2, which is contingent on election outcomes in country B.

Luterbacher and Davis (2010) argue that, as an effective global climate agreement becomes more likely, the risks involved in holding on to carbon-intensive technologies will increase. Drawing on work by Milnor and Shapley (1978) on so-called "oceanic games" (i.e. games with an infinite number – an "ocean" – of players) and by Straffin Jr. (1977) on bandwagon effects in U.S. presidential nominations, they find that abandoning investments in carbon-intensive technologies might entail significant first-mover advantages. If some countries, regions, or municipalities begin to introduce regulation to limit the use of carbon-intensive technologies, the risks for other countries, regions or municipalities of continued use of such technologies will increase. Thus, their incentive to switch to low-carbon technologies will also increase. If the size of the coalition of low-carbon countries reaches a certain level, a bandwagon effect may set in and cause a very rapid increase in the number of countries switching to low-carbon technologies. According to Luterbacher and Davis (2010), this bandwagon effect will be stimulated further if the coalition of low-carbon countries is able to use sanctions to motive other countries to join.

[14] On the other hand, using a coalition model Finus and Rübbelke (2012) find that ancillary benefits do *not* enhance the prospects of an efficient global climate agreement. Countries taking private ancillary benefits into account will reduce their emissions *irrespective* of whether an international agreement exists. Thus, when some countries take ancillary benefits into account, other countries will have weaker incentives to join the agreement than they would have if no countries were to take such effects into account.

3.7. Some lessons from the experimental literature on cooperation

As very few climate agreements exist, the possibilities for using field data to test hypotheses about climate cooperation are limited. It is therefore interesting to explore other options. One such option is to use laboratory experiments. A significant number of such experiments have considered the conditions for public goods provision. What can these experiments learn us about the prospects for effective climate cooperation?

The experimental literature on cooperation largely considers variations of the following game:[15] N subjects endowed with z units of a numéraire good (usually money) decide simultaneously how much of their endowment they will keep for themselves and how much they will contribute to a public good for the subject group. Contributions are divided equally among all subjects after being multiplied by a factor between 1 and N. Assuming subjects are rational actors that maximize their own monetary payoff, the unique Nash equilibrium in this game is that every subject keeps its entire endowment and thus contributes nothing to the public good. This equilibrium is inefficient; all subjects would be better off if every subject were to contribute its entire endowment to the public good.

Several experiments add an enforcement stage, allowing subjects to allocate punishment points to other subjects. One allocated punishment point normally detracts three units of the numéraire good from the punished player's payoff and one unit from the punishing player's payoff. Thus, punishment is costly both for the punished subject and for the punishing subject. The subgame-perfect equilibrium in the one-shot version of this public goods game with enforcement is that all subjects keep their entire endowment and that no subject is punished.

[15] This and the next few paragraphs draw extensively on Aakre et al (2014).

Experimental studies of such public goods games with enforcement (e.g. Fehr and Gächter 2000, 2002; Kosfeld et al. 2009) typically permit subjects to play the game a fixed number of times (usually 10). Given the above-mentioned assumptions, the subgame-perfect equilibrium in such a finitely repeated game is that every player keeps its entire endowment in every period, and that no punishment is imposed.

Typically, the behavior observed in experiments deviates significantly from these equilibrium predictions. In experiments without an enforcement stage, average contributions typically begin at 40–60% of the endowment in the first period and then decreases to 10–15% by the last period. In experiments with an enforcement stage, average contributions typically start higher (at 60–70% of the endowment), and increase even further (often to 90–100%) by the last period. Thus, adding enforcement seems to influence behavior significantly, even when enforcement is costly to the enforcers.

The mechanisms producing these results are not very well understood; however, a popular hypothesis is that motivational heterogeneity plays a major role. Quite a few subjects seem to be "reciprocators" (e.g. Fehr et al. 2002), who increase their current contribution if the average contribution in the preceding period was below their own, and reduce their current contribution if the average in the preceding period was above their own., The tendency of average contributions to decline over time is believed to stem from reciprocators' underestimating the portion of purely self-regarding players (existing data suggest that this portion constitutes around one-third of the subject pool in modern societies). Reciprocators' miscalculation concerning the subject pool causes them to make considerable contributions in the first period, and to reduce their contributions as they observe lower average contributions than they expected.

Experiments with enforcement permit subjects to discipline free riders. Subjects often allocate punishment points, even though such allocation is costly for the punishing subject. A common explanation is that "strong reciprocity" plays a role. A strong reciprocator is willing to forego monetary benefits to penalize subjects that do not cooperate. If a purely self-regarding player believes strong reciprocation is sufficiently widespread, and if allocation of punishment points is possible, contributing at a level that avoids punishment may well be a best response (see for example, Fehr and Fischbacher 2005; Gürerk et al. 2006).

Experiments also indicate that the prospects for cooperation are better if climate change mitigation includes some degree of "lumpiness". In particular, cooperation may be easier to sustain if a minimum amount of effort is required to avoid passing a catastrophic threshold. Barrett (2013) finds theoretically that this conclusion holds only provided that the location of the threshold is known and experiments conducted by Barrett and Dannenberg (2012) support this conclusion. Further experiments reported by Barrett and Dannenberg (2014) show that even an uncertain threshold induces subjects to increase their contributions but – importantly – not enough to avoid passing the threshold.[16]

Other experimental studies have considered the conditions under which a deposit-refund system is likely to enhance climate cooperation. One major finding is that such systems may well be very effective in agreements having an entry-into-force clause that requires full participation. In contrast, they will likely be less effective (and might even be counterproductive) in agreements having an entry-into-force clause that does *not* require full participation (Cherry and McEvoy 2012).

[16] Other ideas involving lumpiness have been studied experimentally by McEvoy (2009).

Moreover, some experiments have addressed whether permitting subjects to vote over the nature of an entry-into-force clause influences cooperation. For example, Cherry et al. (2014) let subjects in a public goods game vote on the number of members required to form before they decide whether to join the agreement and contribute to the public good. They find that subjects tend to successfully introduce an efficient entry-into-force rule when full participation is optimal but not when less than full participation is optimal (i.e., when only a subset of countries are required to participate to solve the problem). They also find that introducing heterogeneous payoff functions aggravates this effect.

Finally, experiments have also been used to consider the importance of equity concerns in the climate negotiations. For example, Dannenberg et al. (2010) conducted an online experiment with subjects who had experience from international climate policy. Using two non-strategic games to measure the subjects' inequality aversion, they find that equity concerns do play a role, although regional differences in climate policy are driven more by differences in national interests than by differences in equity concerns.

Assuming that lab experiments are relevant for international climate cooperation, several of these experimental results offer some reason for optimism. For example, they suggest that a potential for moderate cooperation levels may exist even *without* enforcement and that even very high cooperation levels may be sustainable *with* enforcement. Interestingly, enforcement seems to encourage cooperation even when it is based on threats that are not credible (in the narrow sense that they are costly to implement). The latter result could mean that the credibility requirements imposed in most formal models are excessively strict.

4. Conclusions

The models considered in this paper offer two somewhat conflicting main messages. The first is that the prospects for effective climate cooperation are rather bleak. Climate change mitigation entails huge incentives for free riding and curbing those incentives successfully constitutes a formidable challenge. Many formal models suggest that a stable (self-enforcing) agreement is possible only if the number of signatories is (very) small or if the signatories' commits are shallow.

However, a second message is that a new and carefully designed climate agreement could nevertheless make a significant difference. We end by summarizing three general lessons about international climate cooperation and six more specific lessons about treaty design that can be derived from models reviewed in this article.

The first general lesson is that each country's best course of climate action is likely to depend on what other countries do. This interdependence provides a strong rationale for international coordination; in particular, it suggests that the efforts to design a new and more effective climate agreement should continue despite the rather disappointing achievements so far. In particular, unilateral climate policies – although certainly worthwhile as a supplement to international cooperation – cannot be expected to cut global emissions to the extent required to avoid dangerous anthropogenic climate change.

A second general lesson is that countries' climate policies (or lack of such policies) are motivated to a considerable extent by incentives, that is, the costs and benefits generated by different policy options. Formal models of public goods provision suggest that curbing emissions of greenhouse gases are associated with strong incentives to free ride. The existence of such incentives provides a powerful explanation of what caused the climate change problem. Although countries' behavior is certainly influenced also by other factors

(such as norms), the formal models literature suggests that changing the incentives to free ride must be part of any viable solution.

Finally, the possibility of a climate threshold (beyond which catastrophic climate change will occur) might influence the prospects of climate cooperation. In particular, cooperation may be easier to sustain if some minimum amount of effort is required to avoid passing the threshold. Importantly, however, it seems that this conclusion holds only provided that the threshold's location is known. The latter conclusion provides a caution for those who believe that the risk of a climate disaster will make countries more cooperative concerning climate change mitigation.

However, equally interesting contributions of the models reviewed in this article concern more specific aspects of the design of a new and more effective climate treaty. Many results concerning such aspects contribute to mapping the conditions under which a particular design feature is likely to prove helpful. Identifying such conditions is one of the major strengths of formal models.

First, a carefully chosen entry-into-force clause might help attract more signatories. A demanding clause is particularly likely to have this effect. Indeed, in a symmetric setting, an entry-into-force clause requiring *all* countries to participate may even be able to sustain full participation as an equilibrium. However, the real-world setting of climate negotiations is clearly *not* symmetric, so more research is needed on what type of entry-into-force clause may be expected to work best in *asymmetric* settings.

Second, limiting the targets for emissions reductions can induce more countries to participate with binding emissions reduction commitments. Interestingly, a broad but shallow agreement can – given certain conditions – be more effective than a narrow but deep

agreement. This result suggests that negotiators might be well advised to choose a design with moderate targets − at least in the initial phases of a new climate treaty.

Third, if countries are asymmetric – either in terms of their capacity for contributing to climate change mitigation or in terms of how much they benefit from such mitigation – progressive countries might be able to offer side payments to unattached countries, thereby making larger coalitions stable. However, if laggard countries can also offer side payments to unattached countries, they might be able to reduce or even nullify this positive effect on participation.

Fourth, a new and more effective climate agreement will likely require potent enforcement. However, designing a potent enforcement system that is also politically feasible is a great challenge. Importantly, deterring only one type of free riding (e.g., noncompliance) may simply shift free riding to other types (such as non-ratification, ratification with no commitment, ratification with only a very shallow commitment, or withdrawal). To ensure a new climate agreement's effectiveness, the enforcement system must therefore be able to deter all types of free riding.

Fifth, and related to the previous point, formal models provide insights into the strengths and weaknesses of specific proposals for enforcement systems. For example, they have been used to study the possible merits of trade sanctions, of deposit-refund systems, and of enforcement based on specific reciprocity. The results establish the conditions under which different enforcement systems may or will be effective and highlight potential problems with each type of system. One major conclusion is that effective enforcement through issue-specific reciprocity is possible and consistent with broad participation but requires rather intricate designs and probably presupposes too much flexibility to be politically attractive. Another is that a deposit-refund system might be able to ensure high compliance but would

likely have to be supplemented with other measures (such as some kind of trade restrictions) to ensure also high participation with deep commitments.

Finally, an emerging branch of formal modeling considers how countries' *motivation* influences the prospects for climate cooperation. In particular, these models study the effect of replacing the standard assumption that all countries are fully rational and purely self-interested with an assumption that at least some countries are reciprocators or have a preference for equity. Such models entail far more optimistic results concerning the potential for cooperation than standard models do. In particular, they offer some encouragement to environmental NGOs and other green pressure groups: If such NGOs and pressure groups could convince the governments in sufficiently many countries that climate change is better seen in terms of equity or reciprocity than in terms of national interests, they would also significantly enhance the likelihood of effective climate cooperation.

References

Aakre S (2013) Enforcing Compliance in Climate Agreements: Any Role for Costly Trade Restrictions? CICERO, Oslo: Unpublished working paper

Aakre S, Helland L, Hovi J (2014) When Does Informal Enforcement Work? BI Norwegian Business School, Oslo: Unpublished working paper

Altamirano-Cabrera JC, Finus M, Dellink R (2008) Do Abatement Quotas Lead to More Successful Climate Coalitions? The Manchester School 76(1):104-129

Andonova LB, Betsill MM, Bulkeley H (2009) Transnational Climate Governance. Global Environmental Politics 9(2):52–73

Asheim GB, Froyn CB, Hovi, Menz FC (2006) Regional versus Global Cooperation on Climate Control. Journal of Environmental Economics and Management 51(1):93–109

Asheim GB, Holtsmark B (2009) Renegotiation-Proof Climate Agreements with Full Participation: Conditions for Pareto-Efficiency. Environmental and Resource Economics 43(4):519–533

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

Axelrod R, Keohane RO (1985) Achieving Cooperation under Anarchy: Strategies and Institutions. International Organization 25(4):866–874

Barrett S (1994) Self-enforcing International Environmental Agreements. Oxford Economic Papers 46(4):878–894

Barrett S (1999) A Theory of Full International Cooperation. Journal of Theoretical Politics 11(4):519–541

Barrett S (2002) Consensus treaties. Journal of Institutional and Theoretical Economics 158(4):529–547

Barrett S (2003) Environment and Statecraft: The Strategy of Environmental Treaty Making. Oxford University Press, Oxford

Barrett S (2005) The Theory of International Environmental Agreements. In: Mäler K-G, Vincent JR (eds) Handbook of Environmental Economics, Vol 3. Elsevier, p 1458–1514

Barrett S (2007) Why Cooperate? The Incentive to Supply Global Public Goods. Oxford University Press, Oxford

Barrett S, Dannenberg A (2012) Climate Negotiations under Scientific Uncertainty. PNAS 109(43):17372–17376

Barrett S, Dannenberg A (2014) Negotiating to Avoid "Gradual" versus "Dangerous" Climate Change. An Experimental Test of Two Prisoners' Dilemmas. In: Cherry T, Hovi J, McEvoy D (eds) Toward a New Climate Agreement. Conflict, Resolution and Governance, Routledge, London, p 61–90

Barrett S, Stavins RN (2003) Increasing Participation and Compliance in International Climate Agreements. International Environmental Agreements: Politics, Law and Economics 3(4):349–376

Bernauer T, Kalbhenn A, Koubi V, Spielker G (2010) A Comparison of International and Domestic Sources of Global Governance Dynamics. British Journal of Political Science 40(4):509–538

Biancardi M, Villani G (2010), International Environmental Agreements with Asymmetric Countries. Computational Economics 36(1):69-92

Bloch F (1997) Non-cooperative Models of Coalition Formation in Games with Spillovers. In: Carraro C, Siniscalco D (eds) New Directions in the Economic Theory of the Environment. Cambridge University Press, Cambridge UK

Breitmeier H, Young OR, Zürn M (2006) Analyzing International Environmental Regimes: From Case Study to Database. MIT Press, Cambridge MA

Buchholz W, Haslbeck C, Sandler T (1998) When does Partial Cooperation Pay? Finanzarchiv 55(1):1–20

Caparrós A, Pereau JC (2013) Forming Coalitions to Negotiate North-South Climate Agreements. Environment and Development Economics 18(1):69–92

Caparrós A, Pereau JC, Tazdaït T (2004) North-South Climate Change Negotiations: A Sequential Game with Asymmetric Information. Public Choice121 (3–4):455–480

Carraro C (1999) The Structure of International Agreements on Climate Change. In: Carlo Carraro (ed) International Environmental Agreements on Climate Change. Kluwer Academic Publishers, Dordrecht

Carraro C (2000) The Economics of Coalition Formation. In: Gupta J, Grubb M (eds) Climate Change and European Leadership. Kluwer Academic Publishers, Dordrecht

Carraro C, Eyckmans J, Finus M (2006) Optimal Transfers and Participation Decisions in International Environmental Agreements. Review of International Organizations 1(4):379–396

Carraro C, Marchiori C (2003) Stable Coalitions. In: C. Carraro (ed.) The Endogenous Formation of Economic Coalitions. Edward Elgar, Cheltenham

Carraro C, Siniscalco D (1992) The International Dimension of Environmental Policy. European Economic Review 36(2–3):379–387

Carraro C, Siniscalco D (1993) Strategies for the International Protection of the Environment. Journal of Public Economics 52(3):309–328

Carraro C, Siniscalco D (1997) R&D Cooperation and the Stability of International Environmental Agreements. In: Carraro C (ed) International Environmental Agreements: Strategic Policy Issues. Edward Elgar, Cheltenham

Carraro C, Siniscalco D (1998) International Environmental Agreements: Incentives and Political Economy. European Economic Review 42(3–5):561–572

Carraro C, Marchiori C, Oreffice, S. (2009) Endogenous Minimum Participation in International Environmental Treaties. Environmental and Resource Economics 42(3):411–425

Chander P, Tulkens H (1992) Theoretical Foundations of Negotiations and Cost Sharing in Transfrontier Pollution Problems. European Economic Review 36(2–3):388–398

Chander P, Tulkens H (1993) Strategically Stable Cost-Sharing in an Economic-Ecological Negotiation Process. In: Mäler K-G (ed) International Environmental Problems: An Economic Perspective. Kluwer Academic Publishers, Dordrecht

Cherry TL, McEvoy DM (2012) Enforcing Compliance with Environmental Agreements in the Absence of Strong Institutions: An Experimental Analysis. Environmental and Resource Economics 54(1):63–77

Cherry TL, McEvoy DM, Stranlund J (2014) International Environmental Agreements with Endogenous Minimum Participation and the Role of Inequality. In: Cherry T, Hovi J, McEvoy D (eds) Toward a New Climate Agreement. Conflict, Resolution and Governance. Routledge, London, p 93–105

Dannenberg A, Sturm B, Vogt C (2010) Do Equity Preferences Matter for Climate Negotiators? An Experimental Investigation. Environmental and Resource Economics 47(1):91–109

Dasgupta P, Heal G (1980) Economic Theory and Exhaustible Resources. Cambridge University Press, Cambridge

Dasgupta P (2008) Economic Progress and the Idea of Social Capital. In Dasgupta P, Serageldin I (eds) Social Capital a Multifaceted Perspective. World Bank, Washington

Dietz M, Marchiori C, August T (2012) Domestic Politics and the Formation of International Environmental Agreements. Working Paper. Grantham Research Institute, London

Eyckmans J, Finus M (2004) An Almost Ideal Sharing Scheme for Coalition Games with Externalities. CLIMNEG Working Paper 62, University of Leuven (KUL)

Farrell J, Maskin E (1989) Renegotiation in Repeated Games. Games and Economic Behavior 1(4):327–60

Fehr E, Fischbacher U (2005) The Economics of Strong Reciprocity. In: Gintis H, Bowles H, Boyd RT, Fehr E (eds) Moral Sentiments and Material Interests. The Foundations of Cooperation in Economic Life. Cambridge Mass.: MIT Press

Fehr E, Fischbacher U, Gächter S (2002) Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms. Human Nature 13(1):1-25

Fehr E, Gächter S (2000) Cooperation and Punishment in Public Goods Experiments. The American Economic Review 90(4):980-94.

Fehr E, Gächter S (2002) Altruistic Punishment in Humans. Nature 415:137-40.

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

Finus M (2002) Game Theory and International Environmental Cooperation: Any Practical Application? In Böhringer C, Finus M, Vogt C (eds) Controlling Global Warming: Perspectives from Economics, Game Theory and Public Choice. Edward Elgar, Cheltenham

Finus M (2003) New Developments in Coalition Theory: An Application to the Case of Global Pollution. In: Marsiliani L, Rauscher M, Withagen C (eds) Environmental Policy in an International Perspective. Kluwer Academic Publishers, Dordrecht

Finus M (2008) Game Theoretic Research on the Design of International Environmental Agreements: Insights, Critical remarks, and Future Challenges. International Review of Environmental and Resource Economics 2(1):29–67

Finus M (2008a) The Enforcement Mechanisms of the Kyoto Protocol: Flawed or Promising Concepts? Letters in Spatial and Resource Sciences 1(1):13–25

Finus M, Maus S (2008) Modesty May pay! Journal of Public Economic Theory 10(5): 801–826

Finus M, Rübbelke DTG (2013) Public Good Provision and Ancillary Benefits: The Case of Climate Agreements. Environmental and Resource Economics 56(2):211–226

Finus M, Rundshagen B (1998) Renegotiation-proof Equilibria in a Global Emission Game when Players Are Impatient. Environmental and Resource Economics 12(3):275–306

Finus M, Rundshagen B (2003) Endogenous Coalition Formation in Global Pollution Control: A partition Function Approach. In: Carraro C (ed) Endogenous Formation of Economic Coalitions. Edward Elgar, Cheltenham

Finus M, Rundshagen B (2009) Membership Rules and Stability of Coalition Structures in Positive Externality Games. Social Choice and Welfare 32(3):389–406

Florini A (1996) The Evolution of International Norms. International Studies Quarterly 40(3):363–389

Froyn CB, Hovi J (2008) A Climate Regime with Full Participation. Economics Letters 99(2):317–319

Fudenberg D, Tirole J (1992) Game Theory. MIT Press, Cambridge MA

Fuentes-Albero C, Rubio SJ (2010) Can International Environmental Cooperation Be Bought? European Journal of Operational Research 202(1):255-64

Gersbach, H. (2006) The Global Refunding System and Climate Change. Working paper, CER-ETH Zürich

Gersbach H (2008) A New Way to Address Climate Change: A Global Refunding System. Economists' Voice July 2008. Available from: www.bepress.com/ev

Gersbach H, Winkler R (2007) On the Design of Global Refunding and Climate Change. Discussion Paper 6379, CEPR

Gerber A, Wichardt PC (2009) Providing Public Goods in the Absence of Strong Institutions. Journal of Public Economics 93(3–4):429–439

Gilligan MJ (2004) Is There a Broader-Deeper Trade-Off in International Multilateral Agreements? International Organization 58(3):459–48

Grundig F (2006) Patterns of International Cooperation and the Explanatory Power of Relative Gains: An Analysis of Cooperation on Global Climate Change, Ozone Depletion, and International Trade. International Studies Quarterly 50(4):781–801

Grundig F, Hovi J, Ward H (2014) Modeling Climate Cooperation. In: Luterbacher U, Sprinz, DF (eds) International Relations and Global Climate Change (2nd ed). MIT Press, Cambridge MA

Grundig F, Ward H, Zorick E (2001) Modeling Global Climate-Change Negotiations. In: Luterbcher U, Sprinz D (eds) International Relations and Global Climate Change. MIT Press, Cambridge MA, p 153–182

Grundig F, Hovi J, Underdal A, Aakre S (2012) Self-enforcing Peace and Environmental Agreements. Toward Scholarly Cross-fertilization? International Studies Review 14(4): 522–540

Grüning C, Peters W (2010) Can Justice and Fairness Enlarge International Environmental Agreements? Games 1(2):137–158

Gürerk Ö, Irlenbusch B, Rochenbach B (2006) The Competitive Advantage of Sanctioning Institutions. Science 312 (5770):108–11

Hafner-Burton EM, Kahler M, Montgomery AH (2009) Network Analysis for International Relations. International Organization 63(2):559–592

Harrison K, McIntosh-Sundstrom L (2010) Conclusion: The Comparative Politics of Climate Change. In: Harrison K, McIntosh-Sundstrom L (eds) Global Commons and Domestic Decisions. MIT Press, Cambridge MA

Hoel M (1991) Global Environmental Problems: The Effects of Unilateral Actions Taken by One Country. Journal of Environmental Economics and Management 20(1):55–70

Hoel M (1992) International Environmental Conventions: The Case of Uniform Reductions of Emissions. Environmental and Resource Economics 2(2):141–159

Holtsmark B (2013) International Cooperation on Climate Change: Why is there so Little Progress? In Fouquet, R (ed) Handbook on Energy and Climate Change. Edward Elgar, Cheltenham

Hovi J, Greaker M, Hagem C, Holtsmark B (2012) A Credible Compliance Enforcement System for the Climate Regime. Climate Policy 12(6):741–754

Hovi J, Skodvin T, Aakre S (2013) Can Climate Change Negotiations Succeed? Politics and Governance 1(2):138–150

Hovi J, Sprinz DF (2006) The Limits of the Law of the Least Ambitious Program. Global Environmental Politics 6(3):28–42

Hovi J, Underdal A (2014) Implementation, Compliance, and Effectiveness of Policies and Institutions. In Luterbacher U, Sprinz, DF (eds) International Relations and Climate Change (2nd ed). MIT Press, Cambridge MA

Ismer R, Neuhoff K (2007) Border tax adjustment: a feasible way to support stringent emission trading. European Journal of Law and Economics 24:137–164

Keohane RO (1984) After Hegemony. Cooperation and Discord in the World Political Economy. Princeton University Press, Princeton NJ

Kosfeld M, Okada A, Riedl A (2009) Institution Formation in Public Goods Games. The American Economic Review 99(4):1335–55

Kratzsch U, Sieg G, Stegemann U (2012) An International Agreement with Full Participation to Tackle the Stock of Greenhouse Gases. Economics Letters 115 (3):473–476.

Kydd AH (2007) Trust and Mistrust in International Relations. Princeton University Press, Princeton NJ

Lange A (2006) The Impact of Equity-preferences on the Stability of International Environmental Agreements. Environmental and Resource Economics 34: 247–267

Lange A, Vogt C (2003) Cooperation in International Environmental Negotiations Due to a Preference for Equity. Journal of Public Economics 87 (9–10): 2049–2067

Luterbacher U, Davis P (2010) Explaining Unilateral Cooperative Actions: The Case of Greenhouse Gas Regulations. Monash University Law Review 36(1):121–138

Mansfield ED, Milner HV, Pevehouse JC (2007) Vetoing Cooperation: The Impact of Veto Players on Preferential Trading Arrangements. British Journal of Political Science 37(3):403–32

Maoz Z (2010) Networks of Nations: The Evolution, Structure, and Impact of International Networks, 1816–2001. Cambridge University Press, New York

McEvoy DM (2009) Not It: Opting out of Voluntary Coalitions that Provide a Public Good. Public Choice 142(1):9–23

McEvoy DM (2013) Enforcing Compliance with International Environmental Agreements Using a Deposit-refund System. International Environmental Agreements 13(4):481–496

McGinty M (2007) International Environmental Agreements among Asymmetric Nations. Oxford Economic Papers 59(1):45–62

McGinty M. (2011) A Risk-Dominant Allocation: Maximizing Coalition Stability. Journal of Public Economic Theory 13(2):311–325

McGinty M (2014) Improving the Design of International Environmental Agreements. In: Cherry T, Hovi J, McEvoy D (eds) Toward a New Climate Agreement. Conflict, Resolution and Governance. Routledge, London, p 128–142

Milnor JW, Shapley LS (1978) Values of Large Games II: Oceanic Games. Mathematics of Operations Research 3(4):290–307

Nyborg, K (2014) Reciprocal Climate Negotiators: Balancing Anger against Even More Anger. Working paper Department of Economics, University of Oslo. Available from: http://folk.uio.no/karineny/papers_files/ClimateTreatiesWithReciprocity.pdf

Olson M (1965) The Logic of Collective Action: Public Goods and the Theory of Groups. Harvard University Press, Cambridge

Osgood C (1962) An Alternative to War or Surrender. University of Illinois Press, Urbana IL

Ostrom E, Ahn TK (2008) The Meaning of Social Capital and its Link to Collective Action. In: Svendsen GT, Svendsen GL (eds) Handbook on Social Capital: The Troika of Sociology, Political Science and Economics. Edward Elgar, Northampton MA, p 17–35

Ostrom E, Gardner R, Walker J (1994) Rules, Games and Common-Pool Resources. Michigan University Press, Ann Arbor MI

Putnam R (1988) Diplomacy and Domestic Politics: The Logic of Two-Level Games. International Organization 42(4):427–60

Rabe B (2004) Statehouse and Greenhouse: The Evolving Politics of American Climate Change Policy. Brookings Institution Press, New York

Sandler T (1997) Global Challenges: An Approach to Environmental Political and Economic Problems. Cambridge University Press, Cambridge

Straffin PD Jr. (1977) The Bandwagon Curve. American Journal of Political Science 21(4):695–709

Tulkens H (1979) An Economic Model of International Negotiations Relating to Transfrontier Pollution. In: Krippendorff K (ed) Communication and Control in Society. Gordon and Breach, New York

Underdal A (1980) The Politics of International Fisheries Managements: The Case of the Northeast Atlantic. Columbia University Press, New York

Underdal A (1998) Introduction. In: Underdal A (ed) The Politics of International Environmental Management. Kluwer, Dordrecht

Urpelainen J (2009) Explaining the Schwarzenegger Phenomenon: Local Frontrunners in Climate Policy. Global Environmental Politics 9(3):82-105

Urpelainen J (2011) Can Unilateral Leadership Promote International Environmental Cooperation? International Interactions 37(3):320-339

Urpelainen J (2012) Costly Adjustments, Markets and International Reassurance. British Journal of Political Science 42(4):679–704

Urpelainen J (2012a) Technology Investment, Bargaining, and International Environmental Agreements. International Environmental Agreements 12(2):145-163

Victor DG (2006) Toward Effective International Cooperation on Climate Change: Numbers, Interests and Institutions. Global Environmental Politics 6:90–103

Victor DG (2011) Global Warming Gridlock: Creating More Effective Strategies for Protecting the Planet. Cambridge University Press, Cambridge

Ward H (1996) Game Theory and the Politics of the Global Warming: The State and Beyond. Political Studies 44(4):850–71

Ward H (2006) International Linkages and Environmental Sustainability: The Effectiveness of the Regime and IGO Networks. Journal of Peace Research 43(2):149–166

Ward H, Grundig F, Zorick E (2001) Marching at the Pace of the Slowest: A Model of International Negotiations over Global Climate Change. Political Studies 49(3):438–61

Weikard HP (2009) Cartel Stability under Optimal Sharing Rule. Manchester School 77(5):575-93

Yi S-S, Shin H (2000) Endogenous Formation of Research Coalitions with Spillovers. International Journal of Industrial Organization 18(2):229–256

Young O (1999) Governance in World Affairs. Cornell University Press, Ithaca NY