DISCUSSION PAPER SERIES

IZA DP No. 8866

Reciprocal Climate Negotiators

Karine Nyborg

February 2015

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Reciprocal Climate Negotiators

Karine Nyborg
University of Oslo
and IZA

Discussion Paper No. 8866
February 2015

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8866
February 2015

ABSTRACT

Reciprocal Climate Negotiators*

International climate negotiations have been troubled by mutual mistrust. At the same time, a hope seems to prevail that once enough countries moved forward, others would follow suit. If the abatement game faced by climate negotiators is a Prisoners' Dilemma, and countries are narrowly self-interested, such a hope seems unfounded. However, if countries display reciprocity – a preference to repay meanness by meanness and kindness by kindness – their willingness to abate will be conditional on others' abatement. I show that a full or majority coalition can be stable. This requires, however, that a majority of countries have relatively strong reciprocity preferences. No coalition participation is always stable. In addition, a stable minority coalition may exist; if so, it is weakly larger than the maximum stable coalition with standard preferences, but is characterized by mutually negative sentiments.

JEL Classification: F53, H87, Q54

Keywords: international environmental agreements, reciprocity, coalitions

Corresponding author:

Karine Nyborg
Department of Economics
University of Oslo
P.O.Box 1095 Blindern
N-0317 Oslo
Norway
E-mail: karine.nyborg@econ.uio.no

* I am grateful to Michael Hoel, Jon Hovi, Simon Dietz, Alessandro Tavoni, Natalia Montinari, Aart de Zeeuw, and seminar/conference participants at the Paris Environmental and Energy Economics Seminar, the Grantham Institute, 5th World Congress of Environmental and Resource Economists, University of Lund and the Graduate Institute in Geneva for helpful discusson and comments. Thanks to the Research Council of Norway for funding through the NORKLIMA programme, and to the Beijer Institute of Ecological Economics for hosting me during an early stage of this work. The author is part of CREE (Oslo Centre for Research on Environmentally friendly Energy) and ESOP (Centre for the Study of Equality, Social Organization, and Performance).

1 Introduction

> "Mr. President, this is the worst meeting I've been to since the eighth-grade student council." (*Secretary of State Hilary Clinton to President Barack Obama at his arrival at the Copenhagen climate change summit in 2009.)*[1]

Observers have argued that mistrust and anger have been hindering international climate negotiations.[2] Correspondingly, a hope seems to prevail that if expectations about other countries' intentions could only be improved, a global agreement on emission reductions may be within reach.[3]

From the perspective of standard economic theory, it is not clear why improved beliefs might help, however. Apparently, international climate negotiators are faced with a situation of the Prisoners' Dilemma type: while limiting global warming would benefit all countries, individual countries' incentives to abate their own greenhouse gas emissions are weak (Barrett 2003). In a (static or finitely repeated) Prisoners' Dilemma situation, trust is not essential for players with narrowly self-interested preferences: they will defect, regardless of their expecations.

In line with this, the theoretical literature on international environmental agreements has established a whole array of mostly pessimistic results, showing that any stable climate treaty is likely to have very few signatories and/or involve unambitious emission reduction goals (Barrett 1992, 1994, 2003, Carraro and Siniscalco 1993, Hoel 1992).[4] Sadly, the outcomes from international climate negotiations seem, so far, to confirm this.

Nevertheless, such pessimism has been questioned based on results from behavioral and experimental economics (Grüning and Peters 2007, Burger and Kolstad 2009). A substantial body of research has, indeed, established that in the field and the laboratory alike, groups sometimes manage to cooperate in Prisoners' Dilemma-like situations – even in the absence of external enforcement (Ostrom 1990, Camerer 2003, Zelmer 2003). Nevertheless, the same literature also confirms that cooperation frequently fails (op.cit., Tavoni et al. 2011, Barrett and Dannenberg 2012, 2014).

In the present paper, I consider whether the set of potentially stable international environmental agreements changes if players have *reciprocal preferences,*

[1] As quoted by Landler and Cooper (2010).

[2] "In such a poisonous atmosphere, no meaningful progress is possible." Zammit-Lucia (2013), commenting on the COP 19 in Warsaw.

[3] "The United States and China hope that by announcing these targets now, they can inject momentum into the global climate negotiations and inspire other countries to join in coming forward with ambitious actions as soon as possible" (White House 2014).

"By unveiling clear pledges in Beijing to cap China's emissions by 2030 and further cut America's by 2025, [Presidents Barack Obama and Xi Jinping] have injected hope into a process where little existed before. [...] Mr Xi's pledge will make it harder for others to hang back and raises the chances of a multilateral deal." (Financial Times editorial, 12.11.14).

[4] The literature based on repeated games does include some more optimistic results as well, however; see Froyn and Hovi (2008), Kratzsch et al. (2012), Heitzig et al. (2011).

a phenomenon often favored by behavioral economists as an explanation to observed cooperation patterns (Fehr and Gächter 2000a, Sobel 2005, Croson et al. 2006, Croson 2007).

Combining a simple participation model from Barrett (2003) with a reciprocity model based largely on Rabin (1993), I show that three stable coalition sizes may be feasible: zero participation; a minority coalition; and a majority or even full participation coalition. For the latter to be stable, however, reciprocity preferences must be strong and widespread – possibly more so than the experimental literature indicates. The minority coalition, which is weakly larger than the maximal stable coalition with standard preferences, does not require this. The model thus provides a possible rationale for the existence of a relatively small "coalition of the willing". Nevertheless, this minority coalition barely improves on welfare, compared to the no compliance situation – just as in the standard preferences case.

Reciprocity can be defined as a *preference* for repaying mean (kind) intentions by mean (kind) actions. This should not be confused with a reciprocal *strategy,* like tit-for-tat; to distinguish the two, Sobel (2005) uses the term 'intrinsic reciprocity' for what I call reciprocal preferences or just reciprocity. Formal modelling approaches include Rabin (1993), Levine (1998), Dufwenberg and Kirschsteiger (2004), Falk and Fischbacher (2006) and Cox et al. (2007).

While the experimental economics literature finds that humans are more cooperative than implied by narrowly selfish preferences (Camerer 2003), these findings can hardly be explained by altruism alone. Models of altruism (e.g. Andreoni 1988, 1990) typically predict voluntary contributions to public goods to be *decreasing* in others' contributions, whereas the empirical evidence suggests that this relationshop is *increasing* (Nyborg and Rege 2003, Croson 2007). Moreover, in public good game experiments, hardly any subjects are unconditional altruists who keep contributing if others do not; a substantial share of subjects, however, are *conditional* cooperators, who contribute more the more others contribute (Ledyard 1995, Fischbacher et al. 2001, Fischbacher and Gächter 2006, 2010, Croson et al. 2006). Martinsson et al. (2013) summarize findings from experiments across the world, concluding that conditional cooperators tend to constitute a majority, or close to a majority, of subjects: their list comprises Colombia with 63%, Vietnam 50% (Martinsson et al. 2013); Switzerland 50% (Fischbacher et al. 2001); Denmark 70 % (Thöni et al. 2012); Russia 56% (Herrmann and Thöni 2009); USA 81%, Switzerland 44%, and Japan 42% (Kocher et al. 2008).

Conditional cooperation can sometimes be explained by inequity aversion (Fehr and Schmidt 1999, Bolton and Ockenfels 2000). Like altruism, inequity aversion models define preferences over material outcomes only. A growing body of experimental evidence indicates, however, that people also care about others' *intentions* (Fehr and Gächter 2000a, Camerer 2003). In ultimatum games, for example, responders tend to reject inequitable offers from proposers; but if offering an equitable share was not an option for the proposer, responders are considerably more likely to accept (Falk et al. 2003). Such behavior can be

explained neither by altruism nor inequity aversion.[5]

Unlike the altruist, a reciprocal person is not generally kind. Rather, reciprocity is about anger and gratitude, retaliation and reward. Reciprocity may help secure cooperation, but can also be very destructive (Rabin 1993). An altruistic or inequity averse person would never destroy valuable resources for the sake of revenge; a reciprocal person might do precisely that (Sobel 2005).

The idea of studying social preferences within the framework of international environmental agreement models is not new. Hoel and Schneider (1997) show that if there is some non-environmental cost of breaking agreements, the size of equilibrium coalitions is enlarged. Van der Pol et al. (2012) find that 'community altruism', where signatories care about other signatories but not about non-signatories, increases treaty participation. Lange and Vogt (2003) show that inequity aversion can increase coalition size; if the abatement choice is discrete, they find that even the fully cooperative outcome may be feasible. Lange (2006) allows heterogeneity between countries, and finds that inequity aversion with respect to abatement targets across industrialized countries makes larger coalitions feasible.[6]

In spite of the reciprocity concept's popularity in behavioral economics, formal models have rarely been employed in the applied literature. To my knowledge, Hadjiyiannis et al. (2012) is the only preexisting formal analysis of reciprocal preferences in the context of international environmental agreements. While I study coalition participation in an N-player game with discrete abatement choices, Hadjiyiannis et al. (2012) are concerned with compliance, assuming continuous abatement and only two players. They find that reciprocity can facilitate cooperation, but only if the abatement level required to be viewed as 'fair' by the other player is low; whenever countries' fairness view is more demanding, they find that reciprocity is detrimental to cooperation.

Formal reciprocity models tend to become analytically very complex.[7] To keep the analysis tractable, I use a very simple model of participation in international environmental agreements in which abatement choices are binary, abatement costs and environmental benefits are linear, and all countries are identical. Concerning the modeling of reciprocity, I follow Rabin (1993) closely, but modify his approach to allow for more than two players. It turns out that within the game I study, reciprocity can be expressed in a simple and tractable way.[8]

[5] See also Frans de Waal's capuchin monkey fairness experiment on https://www.youtube.com/watch?v=-KSryJXDpZo&feature=player_detailpage. The monkey in the video could possibly be acting strategically, but its behavior can be explained neither by altruism nor inequity aversion.

[6] See also Grüning and Peters (2007), Kolstad (2013).

[7] Standard game theory defines preferences over outcomes only. If players care intrinsically about others' intentions, one may need to define preferences over beliefs. For this reason, Rabin (1993) applies psychological game theory (Geanakoplos et al., 1989) in his 2-player analysis. With N players, the set of potentially relevant beliefs easily become excessively complex; furthermore, as pointed out by Dufwenberg (2008), research on psychological games is still in its infancy.

[8] More precisely, reciprocity can be defined as a function of own and others' strategies (not beliefs per se), permitting me to use an approach by Segal and Sobel (2007) rather than

Even if individuals were reciprocal, it would not follow automatically that *countries* behave reciprocally. Experimental findings are somewhat mixed regarding the cooperativeness of groups versus individuals (Kocher and Sutter 2007, Kugler et al. 2007, Hauge and Røgeberg 2014). Below, I explore possible stable coalitions *if* countries act as if they have reciprocal preferences. I do not claim that countries *do* have reciprocal preferences. However, a democratic government wanting to be re-elected may well act according to reciprocal preferences if it believes that the median voter is reciprocal. Similarly, if government leaders or negotiating officials hold reciprocal preferences, this may of course influence their negotiation behavior.[9]

2 The non-cooperative abatement game

Consider the simple one-shot global abatement game with $N \geq 2$ identical countries described by Barrett (2003, Ch.7). Each country i can choose either to abate ($q_i = 1$) or to pollute ($q_i = 0$). The material payoff for country i, π_i, consists of its environmental benefits from abatement (compared to some baseline) less its own abatements costs, given by

$$\pi_i = b(Q_{-i} + q_i) - cq_i \tag{1}$$

where $b > 0$ is the environmental benefits to the individual country due to one unit of abatement (by any country), $c > 0$ is a fixed per unit abatement cost, and $Q_{-i} = \sum_{j=1}^{N} q_j - q_i = \sum_{j \neq i} q_j$ denotes the sum of abatement by countries other than i. Moreover, $b < c$ and $bN > c$ (i.e. $N > c/b$). If countries' preferences coincide with their material payoffs as given by eq. (1), and this is common knowledge, the above constitutes an N-player Prisoners' Dilemma game. Pollute ($q_i = 0$) is then a strictly dominant strategy with non-cooperative play; nevertheless, each country would have been better off if everyone had chosen Abate instead.[10] The dilemma is illustrated in Figure 1: regardless of how many others abate, country i's material payoff is always strictly higher if it pollutes itself.

INSERT FIGURE 1 ABOUT HERE

psychological game theory.

[9] It is also conceivable that reciprocal preferences might represent behavior of self-interested countries acting strategically in a bigger game of international relations in general. This remains to be explored, however. For a formal model of issue linkage, see e.g. Conconi and Perroni (2002).

[10] This must be the case since $\pi(1, Q_{-i}) = bQ_{-i} + b - c$ while $\pi(0, Q_{-i}) = bQ_{-i}$. Thus $\pi(1, Q_{-i}) - \pi(0, Q_{-i}) = b - c < 0$ (by assumption). Hence, for any Q_{-i}, $q_i = 1$ yields strictly lower material payoff for i than $q_i = 0$. If all play Abate, payoff for each country is $bN - c > 0$. If all play Pollute, payoff for each is 0.

2.1 Defining reciprocity

Assume now that country i's utility u_i depends on its material payoff π_i as well as reciprocity concerns R_i, where linear separability is assumed for simplicity:

$$u_i = \pi_i + \alpha R_i \tag{2}$$

Below, "payoff" will refer to material payoff π_i, while "utility" or "preferences" refer to u_i.

Rabin (1993) assumed that reciprocity consists, essentially, of two parts: First, the negative (positive) emotion of being treated badly (nicely); second, the satisfaction of repaying by being mean (kind) in return. Consider the following story: Paul pays Ann's bill at a restaurant. Ann thinks Paul does this to insult her, which makes her feel bad (the first part). However, Ann's pain is reduced if, when leaving the restaurant, she tells Paul that he's a snobbish fool, insulting him back (the second part).[11]

Let f_{ij} denote i's kindness towards j, and let $\tilde{f}_{ji}$ be i's belief about j's kindness towards i (for $i \neq j$). $f_{ij} < 0$ (> 0) means that i is being mean (kind). Extending Rabin's 2-player normal form game to allow for $N > 2$ players, I define the reciprocal part of utility as follows:

$$R_i = \frac{1}{N-1}[\sum_{j \neq i} \tilde{f}_{ji} + \sum_{j \neq i} f_{ij}\tilde{f}_{ji}] \tag{3}$$

where the sums are over all $j = 1, ..., N$ for whom $j \neq i$. That is, each binary reciprocity relationship consists of the two parts discussed above (the pain of being treated badly, represented by $\tilde{f}_{ji}$, plus the pleasure of repayment, represented by $f_{ij}\tilde{f}_{ji}$).[12] R_i is given by the average of each bilateral reciprocity relation. That is, I assume that a country cares about the average relationship between itself and each other country, while being unconcerned about the relations between others.

"Kindness", f_{ij}, could potentially be defined in many ways. Here, I follow the intuition of Rabin (1993): the more material payoff I am trying to secure to you, compared to what I *could* have secured to you, the kinder I am.

Let σ_i be i's strategy, let σ_{-i} denote the strategies of everyone other than i, and let $\tilde{\sigma}_{-i}$ denote i's belief about the strategies of everyone else. Accordingly, let $\pi_j(\sigma_i, \sigma_{-i})$ denote the material payoff j will get as a function of i's and others' strategies. Then, $\pi_j(\sigma_i, \tilde{\sigma}_{-i})$ is the material payoff i is trying to secure to j, given i's beliefs.

Let $\pi_{ij}^{\max}$ denote the maximum of $\pi_j(\sigma_i, \tilde{\sigma}_{-i})$ with respect to σ_i (the most i could secure to j for a given set of beliefs), and let $\pi_{ij}^{\min}$ denote the minimum

[11] As hinted at by this example, misunderstandings can lead to gridlock in the relationship between reciprocal players; different norms, cultures, fairness views, affluence and/or histories hardly make things easier.

[12] Usually, only the second part is behaviorally relevant: even if you are pained by someone else's (believed) bad intentions, you may be left to take those intentions as given. In the present analysis, I still need to include both parts, since each can be behaviorally relevant when coalitions behave cooperatively.

of $\pi_j(\sigma_i, \tilde{\sigma}_{-i})$ (the least i could secure to j for given beliefs). Then, I define kindness from i to j as

$$f_{ij} = \frac{\pi_j(\sigma_i, \tilde{\sigma}_{-i}) - \pi_{ij}^e}{\pi_{ij}^{\max} - \pi_{ij}^{\min}} \tag{4}$$

where

$$\pi_{ij}^e = \frac{1}{2}(\pi_{ij}^{\max} + \pi_{ij}^{\min}). \tag{5}$$

If $\pi_{ij}^{\max} = \pi_{ij}^{\min}$, then $f_{ij} = 0$. Note that although I have suppressed this in the notation, $\pi_{ij}^{\max}$, $\pi_{ij}^{\min}$, and π_{ij}^e are all functions of the beliefs about others' strategies, $\tilde{\sigma}_{-i}$.[13]

With this specification, kindness depends on the payoff i tries to secure to j, compared to a fair or "equitable" payoff π_{ij}^e. The "equitable" payoff is the average of the least and most i could have secured to j (given i's beliefs). Finally, still following Rabin, this is normalized by $(\pi_{ij}^{\max} - \pi_{ij}^{\min})$, the range of payoffs i could have secured to j. The latter can be interpreted to mean that kindness is measured relatively to i's power vis-a-vis j.

If I choose the strategy that gives you the highest possible material payoff, given my beliefs about your and others' strategies, I am being maximally kind. If I choose the strategy that gives you the least possible material payoff, given my beliefs about yours and others' strategies, I am being minimally kind. With this specification, thus, what matters is what I try to secure to you compared with the options I think I have, not how much I sacrifice to do so.

2.2 Reciprocity in the non-cooperative abatement game

In the non-cooperative abatement game presented above, there is only one way for i to influence j's payoff: i's choice of pollute or abate.

Taking others' behavior as given, i can secure no more to j than $\pi_{ij}^{\max} = b(Q_{-i} + 1) - cq_j$. Similarly, i can secure no less to j than $\pi_{ij}^{\min} = b(Q_{-i}) - cq_j$. Defining the equitable payoff π_{ij}^e as the average between these two, according to (5), yields

$$\pi_{ij}^e = b(Q_{-i} + \frac{1}{2}) - cq_j. \tag{6}$$

Inserting this in (4), country i's kindness towards j simplifies to

$$f_{ij} = q_i - \frac{1}{2}. \tag{7}$$

Since environmental quality is a pure public good, i is always equally kind or mean to everyone else. Moreover, i's kindness does not depend on others' strategies.[14]

[13] Rabin (1993) distinguishes between the *minimum Pareto efficient* payoff a player could have secured to another, and the *minimum* payoff a player could have secured to another. I am disregarding this distinction here.

[14] This is due to the assumed linearity of the environmental benefits (and costs being independent of others' efforts).

Thus, i's belief about j's kindness, $\tilde{f}_{ji}$, can quite naturally be assumed to depend, similarly, on j's strategy only:

$$\tilde{f}_{ji} = q_j - \frac{1}{2}. \tag{8}$$

Inserting from eqs. (7) and (8), using $Q_{-i} = \sum_{j \neq i} q_j$ and that $f_{ij} = f_{ik}$ for all $j, k \neq i$, R_i can now be written as a function of own and others' strategies as follows:

$$R_i = (\frac{Q_{-i}}{N-1} - \frac{1}{2})(q_i + \frac{1}{2}) \tag{9}$$

This expression says that reciprocity concerns depend on the average kindness of others – judged by their abatement choices – and one's own abatement choice.

Inserting this into the utility function (2) defines reciprocal utility as a function of own and others' strategies:

$$u_i = u(q_i, Q_{-i}) = b(Q_{-i} + q_i) - cq_i + \alpha(\frac{Q_{-i}}{N-1} - \frac{1}{2})(q_i + \frac{1}{2}). \tag{10}$$

Segal and Sobel (2007) showed that some psychological games can be reformulated assuming that players have preferences over strategies, rather than beliefs, and developed solution concepts for such games. Below, I will be using their definition of Nash equilibria in such games.

2.3 Nash equilibria

Let us now turn to abatement decisions in the case where all countries act non-cooperatively. Given others' strategies, $q_i = 1$ (abate) is (weakly) preferred to $q_i = 0$ (pollute) if $u(1, Q_{-i}) - u(0, Q_{-i}) \geq 0$, or

$$\frac{Q_{-i}}{(N-1)} \geq \frac{c-b}{\alpha} + \frac{1}{2}, \tag{11}$$

implying that the share of others who abate must be at least $(c-b)/\alpha + \frac{1}{2}$. This corresponds to strictly more than a majority (since $c > b$ and $\alpha > 0$).

Define now $\hat{Q}_{-i}$ as the number of others abating that would make i exactly indifferent between abating and polluting:

$$\hat{Q}_{-i} = (\frac{c-b}{\alpha} + \frac{1}{2})(N-1) \tag{12}$$

Whenever the number of other countries that abate exceeds $\hat{Q}_{-i}$, reciprocal concerns are sufficiently strong to outweigh the material incentive to free-ride. Whenever $Q_{-i} < 1/2$, reciprocity *reinforces* the incentive to pollute as compared to the model with standard preferences.

Note that $\hat{Q}_{-i}$ is strictly decreasing in α: the stronger the reciprocity preferences, the lower the number of abating others needed to make abatement individually preferable. Nevertheless, $\hat{Q}_{-i}$ is always strictly more than half of

the others. If $\alpha < 2(c-b)$, there exists no $\hat{Q}_{-i}$ such that $\hat{Q}_{-i} \leq N-1$, and pollution is individually preferred regardless of others' abatement.

Following Segal and Sobel (2007), a Nash equilibrium is a strategy profile for which every agent i's strategy maximizes U_i, given that i's expectations about how his opponents will play the game are considered fixed. Let Q be the total number of countries that abate. The following proposition then demostrates that although the symmetric one-shot climate game is a Prisoners' Dilemma in material payoffs, it becomes a coordination game in utilities.

Proposition 1 *In the non-cooperative abatement game with identical, reciprocal countries, i) $Q = 0$ is a Nash equilibrium. ii) If $\alpha > 2(c-b)$, $Q = N$ is a Nash equilibrium. iii) If $\alpha > 2(c-b)$, the following situation is a Nash equilibrium: every country i uses a mixed strategy such that $q_i = 1$ with probability p and $q_i = 0$ with probability $1-p$, where $p = (c-b)/\alpha + 1/2$. In this situation, every country i is indifferent between abate and pollute. There is no pure strategy Nash equilibrium in which countries use different strategies.*

Proof. See the Appendix. ■

These results are illustrated in Figure 2. $\hat{Q}_{-i}$ represents a tipping point in the model. If at least $\hat{Q}_{-i}$ others abate, the reciprocal benefits from abatement are sufficiently large to make it individually rational for every remaining country to abate too. Thus, if reciprocity is strong enough, abatement by every country is a Nash equilibrium. Conversely, if fewer than $\hat{Q}_{-i}$ others abate, the reciprocal benefits from abatement are too small to make abatement individually rational.

The last sentence of Proposition 1 may be somewhat surprising. If everyone has the same preferences and still use different pure strategies in Nash equilibrium, it must be because each is indifferent between the two pure strategies. In the present game, this is not possible because utility depends on what *others* do (see the Appendix for details).

INSERT FIGURE 2 ABOUT HERE

Using eq.(10), it is easy to establish that the Nash equilibrium $Q = N$ is Pareto superior to $Q = 0$: If $Q = 0$, the utility of each country is

$$u_i = u(0,0) = -\frac{1}{4}\alpha < 0 \tag{13}$$

while if $Q = N$, we have

$$u_i = u(1, N-1) = bN - c + \frac{3}{4}\alpha > 0. \tag{14}$$

2.4 What if some countries do not have reciprocal preferences?

If only some countries are reciprocal, while the others care only about material self-interest π_i, $Q = N$ cannot be a Nash equilibrium. However, if reciprocity is strong enough and widespread enough, a high abatement Nash equilibrium, in which a majority of countries abate, still exists.

Proposition 2 *Assume that preferences are given by*

$$u_i = \pi_i + \alpha_i R_i$$

where $\alpha_i \in \{0, \alpha\}$. *Let* $A \leq N$ *be the number of countries with* $\alpha_i = \alpha$, *while* $N - A$ *is the number of countries with* $\alpha_i = 0$. *Then, if*

$$A > \frac{N+1}{2}$$

and

$$\alpha \geq 2(c-b)\frac{N-1}{2A-N-1}$$

there are two pure strategy Nash equilibria in the non-cooperative abatement game, represented by $Q = 0$ *and* $Q = A$, *respectively.*

Proof. See the Appendix. ■

Note that Proposition 2 requires that a strict majority of countries are reciprocal. As demonstrated above, the tipping point $\hat{Q}_{-i}$ is always strictly larger than a majority; hence, if less than half are reciprocal, the tipping point cannot be reached. Furthermore, each reciprocal country must have an even stronger preference for reciprocity than what was required for the full abatement equilibrium in Proposition **1.**

3 Coalition participation with reciprocity

Let us now turn to the treaty participation game extensively studied in the literature on international environmental agreements. Consider a three-stage game as follows (Barrett 2003, Ch. 7):

Stage 1: Every country i chooses whether or not to be part of the coalition;

Stage 2: Signatories decide their strategies collectively, aiming to maximize the coalition's total payoff;

Stage 3: Non-signatories choose their strategies non-cooperatively.

3.1 The standard preferences case

Consider first the standard case where each country maximizes its own payoff π_i (see e.g. Barrett 2003, Wagner 2001). The game is solved by backward induction. In Stage 3, pollute is a strictly dominant strategy for non-cooperative

players, so every non-signatory will pollute. Given this, the joint payoff of a coalition S of k countries is $\sum_{s \in S} \pi_s = k(bk - c)$ if they all abate, and 0 if they all pollute. Hence, in Stage 2, the coalition will prefer its members to abate if $k \geq c/b$. Given this, countries decide in Stage 1 whether to join S.

A coalition of size k is said to be stable if it satisfies the requirements of internal as well as external stability, see Wagner (2001). Internal stability requires that when $k-1$ others are members, and you are a member, it is better for you to stay than to leave. External stability requires that if k others are members, but you are not, it is better for you to stay outside.

Following Wagner (2001), let $\Pi_s(k)$ denote the material payoff of a signatory country as a function of the number of signatories k. Similarly, let $\Pi_n(k)$ denote the material payoff of a non-signatory country as a function of the number of signatories k. Then internal stability requires $\Pi_s(k) \geq \Pi_n(k-1)$, while external stability requires $\Pi_n(k) \geq \Pi_s(k+1)$.

If no other country takes part in the coalition, country i will not prefer to form an abating coalition on its own (if it is at all meaningful to speak of a coalition of one). Thus, $k = 0$, the coalition of zero members, is stable. However, with standard preferences, there is another possibility as well. Let k^0 be the smallest integer such that $k^0 \geq c/b$. A coalition of size k^0 is stable (Barrett 2003, Ch. 7.6): a country expecting $k^0 - 1$ others to join will join too, because its participation is required to make the other signatories abate (which they will do only if $c/b \leq k < c/b + 1$); for the same reason, a signatory of a coalition of size k^0 will stay. If the coalition is larger than k^0, the individual signatory will prefer to leave, since it is not pivotal for the coalition's abatement.[15]

The implication is, unfortunately, that coalition formation can improve the sum of countries' payoffs only very slightly compared to the non-cooperative outcome of no abatement. This is illustrated in Figure 3.

INSERT FIGURE 3 ABOUT HERE

If $k^0 = c/b$, the coalition will provide no net benefits at all to its members compared to the no abatement case, since their environmental benefits exactly outweigh their abatement costs. There will still be a net benefit to non-members, who free-ride on the coalition's efforts. Signatories cannot gain from leaving because if one of them does, the coalition collapses (does not abate); hence, the relevant alternative for a signatory country is that *no-one* abates. If $k^0 > c/b$, the existence of the coalition secures a strictly positive gain to coalition members as well.

For example, assume that $N = 100$, $b = 2$ and $c = 3$. Then, $c/b = 3/2$, hence $k^0 = 2$. With a coalition of 2 countries, each signatory gets a payoff of 1, while each non-signatory gets a payoff of 4. Had all 100 countries abated, each of them would instead have received a payoff of 197.

[15] If k is larger than $c/b + 1$, signatory s faces the same freeriding incentive as in the non-cooperative game: the coalition abates regardless of whether s stays, and s's abatement cost c is not outweighted by the corresponding benefits to s, b.

3.2 Defining reciprocity in the three-stage game

Assume now that every country i has reciprocal preferences as given by eq. (2) above, where $\alpha > 0$, and where R_i is given by eqs. (3) - (5). Suppose also that a coalition S, if formed, collectively maximizes the sum of its members' utilities

$$\sum_{s \in S} u_s = \sum_{s \in S} (\pi_s + \alpha R_s) \tag{15}$$

with respect to q_s for every signatory $s \in S$. Assume that the coalition always chooses the same abatement strategy q_s for every member s.

A strategy σ_i for country i now consists of a plan, for any given beliefs about others' strategies, of whether to join the coalition in Stage 1 and, if a non-signatory, whether to abate or pollute in Stage 3. If i's strategy implies joining, i's abatement is determined by the coalition's policy in Stage 2.

In the non-cooperative case, a country's impact on others depended only on its own abatement choice. In the three-stage coalition game, i's impact on j may also depend on others' strategies – more precisely, on whether i is pivotal for the coalition's abatement or not. If i is not pivotal, its power to change others' payoff is just as limited as it was in the non-cooperative game, and kindness can be calculated as before. If i is pivotal, its power is considerably larger: it can then secure or cancel out abatement efforts not just from itself, but from others too.

It turns out, however, that even for pivotal players, the kindness function (4) can be simplified into exactly the same expression as before: $f_{ij} = q_i - \frac{1}{2}$. (This is shown formally as part of the proof of Proposition 6 below.) As mentioned above, the kindness measure is essentially a relative one. In the coalition formation game, a pivotal player's choice has a larger impact on others – but since a player's kindness is normalized by this player's power, the kindness function ends up being unchanged.[16]

3.3 Stable coalitions with reciprocity

Let $U_s(q_s, k)$ denote the utility of a signatory country as a function of the coalition's abatement policy for each of its members q_s and the number of signatories k. Similarly, let $U_n(q_n, k)$ denote the utility of a non-signatory country n as a function of its own abatement choice q_n and the number of signatories k.[17]

In the following, I look for coalitions which are externally and internally stable, in the following sense:

[16] With a different specification of kindness, results might of course change; exploring this would, however, require a separate analysis.

[17] Note the distinction as compared to the notation $u(q_i, Q_{-i})$ from the non-cooperative case: $u(q_i, Q_{-i})$ is the same function for all i, and gives i's utility as a function of i's own and others' behavior. $U_m(q_m, k)$ is a different function depending on whether $m = s$ or $m = n$, where m is i's coalition membership status; moreover, the second variable of $U_m(q_m, k)$ is the number of coalition members, which may or may not correspond to the number of others abating, Q_{-i}.

Definition 3 *A coalition of size k is internally stable if $U_s(q_s, k) \geq U_n(q_n, k-1)$, and expectations are correct in the sense that every $s \in S$ expects $k-1$ other countries to be signatories, while every $n \notin S$ expects k other countries to be signatories.*

Definition 4 *A coalition of size k is externally stable if $U_n(q_n, k) \geq U_s(q_s, k+1)$, and expectations are correct in the sense that every $s \in S$ expects $k-1$ other countries to be signatories, while every $n \notin S$ expects k other countries to be signatories.*

Proposition 5 below establishes that the empty coalition is stable. Intuitively, if no-one joins, there are no signatories in Stage 3, which means that everyone plays non-cooperatively; consequently, we can apply Proposition 1, part i, which says that with non-cooperative play, zero abatement is a Nash equilibrium.

Proposition 5 *The no-cooperation situation $k = 0$, in which no country is a signatory and all countries pollute, is stable.*

Proof. See the Appendix. ■

If the preference for reciprocity is sufficiently strong, the grand coalition is also stable. This may not be surprising, given that this is a possible Nash equilibrium even in the non-cooperative game.

Proposition 6 *If $\alpha > 2(c-b)$, the grand coalition (the coalition abates, and $k = N$) is stable.*

Proof. See the Appendix. ■

Consequently, with sufficiently strong reciprocity, extremely cooperative as well as extremely uncooperative outcomes can be stable: Low as well as high expectations about others' intentions can be self-fulfilling. This provides a rationale for the view that mistrust and anger can hinder international climate negotiations, and that improved expectations about others' abatement intentions could make a global agreement on abatement feasible.

Moreover, as established by the proposition below, there may exist a third stable coalition size k^1. This resembles the small, stable coalition size k^0 from the payoff-maximizing countries case, but k^1 is weakly larger than k^0. When k^1 is stable, it is always a minority coalition. Consequently, the model provides one possible explanation for existence of a small, but not minimal "coalition of the willing".

Proposition 7 *Assume that $N > 13$ and that $c/b \leq (N+2)/3$. Then, there exists an externally and internally stable coalition consisting of k^1 countries, such that $\frac{N-1}{2} > k^1 \geq k^0 \geq c/b$, for which the coalition abates while non-signatories pollute, and where k^1 is defined as the smallest integer such that $k^1 \geq \underline{k}$, where $\underline{k} = \frac{2c(N-1)+\alpha(N+2)}{2b(N-1)+3\alpha}$.*

Proof. See Appendix. ■

If the number of countries is not too small, and the cost-benefit ratio is modest, a stable minority coalition of k^1 countries exists regardless of the strength of reciprocity concerns (α). The number k^1 itself is weakly increasing in α (with an upper boundary at $\frac{N+2}{3}+1$), but if α becomes sufficiently small, k^1 coincides with k^0.

INSERT FIGURE 4 ABOUT HERE

While the details are given in the proof in the appendix, the main intuition is illustrated in Figure 4. $\underline{k}$ corresponds to c/b in the standard preferences case in the following sense: $\underline{k}$ is the lowest k for which a country is indifferent between *everyone*, including itself, polluting, and being a signatory in an abating coalition. k^1 is the lowest integer weakly above $\underline{k}$. It is easily seen that when $k = \underline{k}$, non-signatories are better off than signatories. However, like in the standard preferences case, a signatory considering to leave cannot take others' abatement as given, because if it leaves, the coalition will collapse.

Since k^1 is a minority coalition, others' behavior is, on average, mean. When $k = k^1$, no-one really wants to abate: not only is it materially unprofitable, but everyone would also like to punish others for their polluting behavior. So how can k^1 be stable?

To understand this, consider the situation of a signatory when $k = k^1$. There is, after all, a small group of $k^1 - 1$ others who are behaving nicely to you. They are too few to make you want to be nice yourself. Still, they do represent a small island of kindness in a mean world. If you stop being nice to them, they will stop being nice to you. The island of kindness will disappear; there will be only meanness left in the world.

3.4 Coalition participation if some countries are not reciprocal

Finally, consider the case where only some countries are reciprocal. In this case, the grand coalition is not feasible, but there may still exist stable, abating coalitions of strictly positive size.

Indeed, if reciprocity preferences are sufficiently strong and widespread, both $k = 0$ and $k = A$ are stable. Moreover, although k^0 was not stable with only reciprocal countries unless $k^0 = k^1$, it is now possible that k^0 as well as k^1 are stable even if $k^0 \neq k^1$.

Assume, like in Proposition 2, that preferences are given by

$$u_i = \pi_i + \alpha_i R_i$$

where $\alpha_i \in \{0, \alpha\}$. Let $A \leq N$ be the number of countries with $\alpha_i = \alpha$, and let $N - A$ be the number of countries with $\alpha_i = 0$.

If the conditions for Proposition 2 hold, i.e. if A and α are sufficiently large, we know that there is a Nash equilibrium in the non-cooperative game in

which every reciprocal country abates, while every payoff-maximizing country pollutes. Consequently, under those same assumptions, there is a corresponding stable majority coalition $k = A$ in the three-stage game.

If reciprocity preferences are too weak and/or the number of reciprocal countries is too small, no such majority coalition can be stable. Even if a stable coalition $k = A$ does exist, it is not necessarily realized, since other, smaller coalition sizes are stable too. In particular, the no participation coalition is always stable (see the proof for Proposition 5). Again, expectations will tend to be self-fulfilling.

If the assumptions for Proposition 7 hold, and if $A \geq \underline{k} = \frac{2c(N-1)+\alpha(N+2)}{2b(N-1)+3\alpha}$, there is a stable minority coalition size $k^1 \geq k^0$. This holds whether the coalition of $k = A$ is stable or not.[18]

In fact, when only some countries are reciprocal, then even if $k^1 > k^0$, both k^0 and k^1 can be stable. For k^0 to be stable, it must consist only of non-reciprocal countries; for k^1 to be stable, it must consist only of non-reciprocal countries.

4 Empirical relevance

The model presented above is of course highly stylized. While assuming continuous abatement and/or heterogeneous country size would clearly be more realistic, this would require separate analyses; one would, for example, need to reconsider how to model "kindness".

In the simple model with identical countries presented above, we must have $\alpha/2 > c - b$ for the grand coalition to be stable (Proposition 6).[19] If not all countries are reciprocal, strictly more than half of them must be so in order for a majority coalition to be stable.

Even though *countries'* preferences cannot necessarily be inferred from *individual* behavior in small-scale laboratory experiments, a glance at the results from the literature on public good game experiments may be of interest here. These results are not exclusively encouraging.

As noted in the introduction, researchers have found that typically, about half of experimental subjects are conditional cooperators – but the share varies between countries, and is often at or slightly below 50 percent (Martinsson et al., 2013).

In public good game experiments in the lab, cooperation is rarely sustained over time. The typical finding is that players contribute substantially in one-shot games and in the first round of repeated games, but as the game is repeated,

[18] Note, however, that reciprocity concerns must be relatively strong for k^1 to be substantially larger than k^0. If, for example, $b = 2$, $c = 3$, $N = 100$ and $\alpha = 2$ (for all countries), k^1 and k^0 coincide at $k^1 = k^0 = 2$. With $\alpha = 10$ instead, $k^0 = 2$, while $k^1 = 4$.

[19] In the above model, $\alpha = 2$ means that if everyone else is kind, the country is willing to sacrifice one unit of material payoff for the satisfaction of being kind in return. Similarly, if everyone else is mean, the country is willing to give up exactly one unit of material payoff for the satisfaction of being mean in return.

cooperation dwindles fast (Ledyard 1995, Camerer 2003, Zelmer 2003, Barrett and Dannenberg 2012, 2014). The decline does not seem to be caused by learning or confusion, but rather by conditional cooperators being disappointed by others' contribution levels (Fishbacher and Gächter 2010). That is, conditional cooperation – which is consistent with reciprocity – is present in the lab, but it is rarely strong enough and/or widespread enough to sustain cooperation. To keep contributions to a public good high in the lab, additional institutions, like individual sanctioning mechanisms or endogenous sorting into groups, are typically required (see, e.g., Fehr and Gächter 2000b, Brekke et al. 2011).

One may distinguish between at least three types of obstacles. First, there is the question of whether reciprocal preferences are indeed strong and widespread enough. Second, even if they are, cooperation will not be achieved if countries expect others to pollute: low expectations will tend to be self-fulfilling.

To the latter point, one may object that in the course of climate negotiations, countries are communicating extensively. Thus, if the grand coalition is indeed stable, they might simply decide to coordinate on it. However, this is where the third problem enters – namely that preferences may be private information. If countries do not know the preferences of other countries, each can have a strategic interest in misrepresenting their true preferences. Then, countries *cannot know* in advance whether a majority coalition is stable at all - even if the others claim that they plan to support it.

5 Conclusions

In a simple three-stage climate coalition formation game, I have shown that reciprocal preferences could potentially play an important role.

First, the situation in which everyone pollutes is always stable. Reciprocal countries' unwillingness to abate is then even stronger than that of countries with standard preferences. If no-one else abates, a reciprocal country would *like* to repay others' meanness by polluting itself – a preference which adds to the disincentive represented by the economic cost of abatement.

With sufficiently strong and widespread reciprocity, the grand coalition, or a majority coalition, can be stable as well. Nevertheless, although the experimental literature indicates that reciprocity is indeed prevalent in several cultures, it also shows that, among individual participants in laboratory experiments at least, such preferences do not seem to be strong and widespread enough to sustain high cooperation levels. Thus, although the theoretical analysis does indicate that the grand or majority coalition may be stable, enforced by reciprocity preferences, the question of whether this is at all realistic in practice remains open.

References

[1] Andreoni, J. (1988): Privately provided public goods in a large economy: the limits of altruism, Journal of Public Economics 35 (1), 57–73.

[2] Andreoni, J. (1990): Impure altruism and donations to public goods: a theory of warm-glow giving, Economic Journal 100 (401), 464–477.

[3] Barrett, S. (1992): "International Environmental Agreements as Games", in R. Pethig (Ed.): Conflict and Cooperation in Managing Environmental Resources, Berlin: Springer, 11-37.

[4] Barrett, S. (1994): Self-Enforcing International Environmental Agreements, Oxford Economic Papers 46, 878-894.

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

[6] Barrett, S., and A. Dannenberg (2012): Climate negotiations under scientific uncertainty, Proceedings of the National Academy of Sciences of the United States of America 109 (43), 17372-17376.

[7] Barrett, S., and A. Dannenberg (2014): Second Best Agreements and the Prisoners' Dilemma Trap. Paper presented at the Fifth World Congress of Environmental and Resource Economists, Istanbul, June 28 - July 2.

[8] Bolton, G.E, and A. Ockenfels (2000): ERC – A theory of equity, reciprocity, and competition, American Economic Review 90(1), 166-93.

[9] Brekke, K.A., K.E. Hauge, J.T. Lind, and K. Nyborg (2011): Playing with the Good Guys: A Public Good Game with Endogenous Group Formation, Journal of Public Economics 95, 1111-1118.

[10] Burger, N.E., and C.D. Kolstad (2009): Voluntary Public Goods Provision, Coalition Formation, and Uncertainty, NBER Working Papers 15543, National Bureau of Economic Research.

[11] Camerer, C. (2003): Behavioral Game Theory. Experiments in strategic interaction, Princeton University Press/Russell Sage Foundation.

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

[13] Conconi, P., and C. Perroni (2002): Issue linkage and issue tie-in in multilateral negotiations, Journal of International Economics 57, 423–447.

[14] Cox, J.C., Friedman, D., and S. Gjerstad (2007): A tractable model of reciprocity and fairness, Games and Economic Behavior 59(1), 17-45.

[15] Croson, R., 2007. Theories of commitment, altruism and reciprocity: evidence from linear public goods games. Economic Inquiry 45, 199–216.

[16] Croson, R., Fatas, E., Neugebauer, T., 2006. Reciprocity, matching and conditional cooperation in two public goods games. Economics Letters 87, 95–101.

[17] Dufwenberg, M. (2008): "Psychological games". In S.N. Durlauf and L.E. Blume (Eds.): *The New Palgrave Dictionary of Economics Online,* Second Edition*,* Palgrave Macmillan, 28 January 2014, doi:10.1057/9780230226203.1358.

[18] Dufwenberg, M., and G. Kirchsteiger (2004): A Theory of Sequential Reciprocity, Games and Economic Behavior 47(2), 268-98.

[19] Falk, A., E. Fehr, U. Fischbacher (2003): On the Nature of Fair Behavior, Economic Inquiry 41(1), 20-26.

[20] Falk, A., and U. Fischbacher (2006): A Theory of Reciprocity, Games and Economic Behavior 54, 293-315.

[21] Fehr, E. and U. Fischbacher (2002): Why Social Preferences Matter - the Impact of Non-Selfish Motives on Competition, Cooperation and Incentives, Economic Journal 112, C1-C33.

[22] Fehr, E., and S. Gächter (2000a): Fairness and Retaliation: The Economics of Reciprocity, Journal of Economic Perspectives 14(3), 159-181.

[23] Fehr, E., and S. Gächter (2000b): Cooperation and punishment in public goods experiments, American Economic Review 90 (4), pp. 980–994.

[24] Fehr, E., and S. Gächter (2002): Altruistic Punishment in Humans, Nature 415, 137-140.

[25] Fehr, E., and K. Schmidt (1999): A Theory of Fairness, Competition, and Cooperation. Quarterly Journal of Economics 114, 817-868.

[26] Financial Times (2014): Xi and Obama revive hopes on climate. FT View, 12.11.14. http://www.ft.com/intl/cms/s/0/ec8da760-6a64-11e4-8fca-00144feabdc0.html#axzz3Q0t1dVlb (accessed 27.01.15).

[27] Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters, 71, 397–404.

[28] Fischbacher, U., Gächter, S., 2006. Heterogeneous social preferences and the dynamics of free riding in public goods. IZA Discussion Papers 2011.

[29] Fischbacher, U., Gächter, S., 2010. Social preferences, beliefs, and the dynamics of freeriding in public goods. American Economic Review 100, 541–556.

[30] Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters 71, 397–404.

[31] Froyn, C.B., and J. Hovi (2008): A climate agreement with full participation, Economics Letters 99, 317–319.

[32] Grüning, C., and W. Peters (2007): Can Justice and Fairness Enlarge the Size of International Environmental Agreements? European University Viadrina (http://www.wiwi.europa-uni.de/de/lehrstuhl/fine/fiwi/team/gruening/GrueningPeters_04_07.pdf).

[33] Hadjiyiannis, C., D. Iris, C. Tabakis (2012): International Environmental Cooperation under Fairness and Reciprocity, the B.E. Journal of Economic Analysis & Policy (Topics), 12(1), Article 33.

[34] Hauge, K. E. and O. Røgeberg (2014): Contributing to Public Goods as Individuals versus Group Representatives: Evidence of Gender Differences. Memorandum 16/14, Department of Economics, University of Oslo.

[35] Heitzig, J., K. Lessmann and Y. Zou (2011): Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games, PNAS 108 (38), 15739–15744.

[36] Herrmann, B., and C. Thöni (2009): Measuring conditional cooperation: A replication study in Russia. Experimental Economics, 12, 87–92.

[37] Hoel, M.O. (1992): International Environment Conventions: The Case of Uniform Reductions of Emissions. Environmental and Resource Economics 2, 141-159.

[38] Hoel, M.O., and K. Schneider (1997): Incentives to Participate in an International Environmental Agreement, Environmental and Resource Economics 9(2), 153–170.

[39] Kocher, M.G., T. Cherry, S. Kroll, R.J. Netzer, M. Sutter (2008): Conditional cooperation on three continents, Economics Letters 101, 175–178.

[40] Kocher, M., and M. Sutter (2007): Individual versus group behavior and the role of the decision making procedure in gift-exchange experiments, Empirica 34(1), 63-88.

[41] Kolstad, C.K. (2013): International Environmental Agreements with Other-Regarding Preferences, unpublished paper, Stanford University.

[42] Kratzsch, U., G. Sieg and U. Stegemann (2012): An international agreement with full participation to tackle the stock of greenhouse gases, Economics Letters 115 (3), 473–476.

[43] Kugler, T., G. Bornstein, M.G. Kocher, M. Sutter (2007): Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy, Journal of Economic Psychology 28(6), 646-657.

[44] Landler, M., and H. Cooper (2010): After a bitter campaign, forging an alliance. New York Times March 18, 2010, accessed 10.04.14 at http://www.nytimes.com/2010/03/19/us/politics/19policy.html?pagewanted=all&_r=0.

[45] Lange, A. (2006): The Impact of Equity-Preferences on the Stability of International Environmental Agreements, Environmental and Resource Economics 34, 247-267.

[46] Lange, A., and C. Vogt (2003): Cooperation in International Environmental Negotiations due to a Preference for Equity, Journal of Public Economics 87, 2049-2067.

[47] Ledyard, J.O. (1995): Public Goods: a Survey of Experimental Research. In: Kagel, J.H., Roth, A.E. (Eds.): The Handbook of Experimental Economics. Princeton University Press, Princeton, New Jersey, pp. 111–194.

[48] Levine, D.K. (1998). Modeling Altruism and Spitefulness in Experiments, Review of Economic Dynamics 1, 593-622.

[49] Martinsson, P., N. Pham-Khanh, C. Villegas-Palacio (2013): Conditional cooperation and disclosure in developing countries, Journal of Economic Psychology 34, 148–155.

[50] Nyborg, K. and M. Rege (2003): Does Public Policy Crowd Out Private Contributions to Public Goods? Public Choice 115 (3), 397-418.

[51] Ostrom, E. (1990): Governing the Commons: The Evolution of Institutions for Collective Action, Cambridge: Cambridge University Press.

[52] Rabin, M: Incorporating Fairness into Game Theory and Economics, American Economic Review 83, 1281-1302.

[53] Segal, U., and J. Sobel (2007): Tit for tat: Foundations of preferences for reciprocity in strategic settings, Journal of Economic Theory 136, 197-216.

[54] Sobel, J. (2005): Interdependent Preferences and Reciprocity, Journal of Economic Literature 43, 392-436.

[55] Tavoni, A., A. Dannenberg, G. Kallis, A. Löschel (2011): Inequality, communication, and the avoidance of disastrous climate change in a public goods game. Proceedings of the National Academy of Sciences of the United States of America, 108 (29), 11825-11829.

[56] Thöni, C., J. -R. Tyran, and E. Wengström (2012): Microfoundations of social capital, Journal of Public Economics 96(7-8), 635-643.

[57] Van der Pol, T., H.-P. Weikard, E. van Ierland (2012): Can altruism stabilise international climate agreements? *Ecological Economics* 81, 112-120.

[58] White House (2014): U.S.-China Joint Announcement on Climate Change, http://www.whitehouse.gov/the-press-office/2014/11/11/us-china-joint-announcement-climate-change (accessed 27.01.15).

[59] Zammit-Lucia, J. (2013): COP19: the UN's climate talks proved to be just another cop out. Guardian 02.12.13, http://www.theguardian.com/sustainable-business/cop19-un-climate-talks-another-cop-out (accessed 27.01.15). (http://www.theguardian.com/sustainable-business/cop19-un-climate-talks-another-cop-out).

[60] Zelmer, J. (2003): Linear public good games: a meta-analysis, Experimental Economics 6, 299–310.

6 Appendix: Proofs

Proof of Proposition 1:

Proof. i) For $Q = 0$ to be a Nash equilibrium, it must be the case that $u_i(0,0) \geq u_i(1,0)$ for every i. Since countries are identical, it is sufficient to demonstrate that this holds for one i. Using eq. (10), $u_i(0,0) \geq u_i(1,0)$ is equivalent to

$$\alpha \geq 2(b-c)$$

which will always hold with $\alpha > 0$, because $b - c < 0$.

ii) For $Q = N$ to be a Nash equilibrium, it must be the case that $u_i(1, N-1) \geq u_i(0, N-1)$ for every i. Using (**??**), and that $Q = N$ implies $Q_{-i} = N-1$, this gives

$$\begin{aligned} bN - c + \frac{3}{4}\alpha &\geq b(N-1) + \frac{1}{4}\alpha \\ \alpha &\geq 2(c-b). \end{aligned}$$

iii) Consider first the possibility of a Nash equilibrium in pure strategies in which a share p of countries, where $0 < p < 1$, plays the pure strategy Abate, a share $1-p$ plays the pure strategy Pollute, and where all i are indifferent between Abate and Pollute. This would require, first, that $\hat{Q}_{-i}$ is an integer, otherwise $Q_{-i} = \hat{Q}_{-i}$ is not possible (and if $Q_{-i} \lesseqgtr \hat{Q}_{-i}$, i is not indifferent between the pure strategies). Assume that $\hat{Q}_{-i}$ is an integer. However, if countries play different pure strategies, it cannot be the case that Q_{-i} is identical for all i. For a given Q, if $q_j = 1$ and $q_h = 0$, we must necessarily have $Q_{-j} = Q - 1$ and $Q_{-h} = Q$, hence $Q_{-j} < Q_{-h}$. Thus, the only possibility for all i to be indifferent is if they all play a mixed strategy.

Consider next the possibility that a share p play Abate, strictly preferring Abate, while a share $1-p$ play Pollute, strictly preferring Pollute. Define $\hat{Q}_{-i}$

such that $u_i(1, \hat{Q}_{-i}) = u_i(0, \hat{Q}_{-i})$. Then $Q_{-i} > \hat{Q}_{-i}$ is required for Abate to be strictly preferred by i, while $Q_{-i} < \hat{Q}_{-i}$ is required for Pollute to be strictly preferred. Hence we would need that for any j who Abates, $Q_{-j} > \hat{Q}_{-i}$, while for any h who Pollutes, $Q_{-h} < \hat{Q}_{-i}$. This implies $Q_{-j} > Q_{-h}$. But since, as demonstrated above, $Q_{-j} < Q_{-h}$, this cannot hold.

From eq. (12), we know that when $Q = \hat{Q}_{-i}$, the *share* of others playing Abate is $\frac{1}{2} + \frac{c-b}{\alpha}$. Consider now the possibility that every country i plays a mixed strategy such that $q_i = 1$ with probability $p = \frac{1}{2} + \frac{c-b}{\alpha}$ (and $q_i = 0$ with probability $p = \frac{1}{2} - \frac{c-b}{\alpha}$). Then, the expected number of others playing $q_i = 1$ equals $p(N-1) = \hat{Q}_{-i}$ for every i. In this situation, i is indifferent between Abate and Pollute. By the assumptions $c > b$ and $\alpha \geq 2(c-b)$, we know that $\frac{1}{2} < p < 1$. Hence, for every i, given that every other country plays Abate with probability $p = \frac{1}{2} + \frac{c-b}{\alpha}$, using the same strategy is a best response for i. The expected number of abating countries in this equilibrium is given by $N(\frac{1}{2} + \frac{c-b}{\alpha})$.

■

Proof of Proposition 2:

Proof. For $Q = 0$ to be a Nash equilibrium, it must be the case that $u_i(0,0) \geq u_i(1,0)$ for every i. For reciprocal countries with $\alpha_i = \alpha$, the proof is exactly as in Proposition 1, part i). For countries with $\alpha_i = 0$, this holds because the game is a Prisoners' dilemma and abate is strictly dominated by pollute, see footnote 1.

For $Q = A$ to be a Nash equilibrium, it must be the case that $u_i(0, A) \leq u_i(1, A)$ for A players and $u_i(0, A) \geq u_i(1, A)$ for the remaining $N - A$ players. The latter follows because abate is a strictly dominated strategy for all $N - A$ players who have $\alpha_i = 0$. What remains to be shown is that $u_i(0, A) \leq u_i(1, A)$ for the A players who have $\alpha_i = \alpha$. When $Q = A$, $Q_{-i} = A - 1$. By eq. (11), abate is preferred by i when $Q_{-i} = A - 1$ if

$$A \geq (\frac{c-b}{\alpha} + \frac{1}{2})(N-1) + 1$$

or equivalently,

$$\alpha \geq 2(c-b)\frac{N-1}{2A-N-1}. \tag{16}$$

This is feasible given that $2A - N - 1 > 0$, i.e.

$$A > \frac{N+1}{2}.$$

■

Proof of Proposition 5:

Proof. Assume $k = 0$. Then in Stage 3, all countries are non-signatories and thus play non-cooperatively. We can then use the results from the non-cooperative game. By Proposition 1, part i, we know that $u_i(0,0) \geq u_i(1,0)$ and that $Q = 0$ is a Nash equilibrium in the non-cooperative game; thus if $k = Q_{-i} = 0$ in the participation game, each non-signatory pollutes in Stage 3.

If $k = 0$, there is no coalition to decide in Stage 2 whether to Abate or not. If one country still joined in Stage 1, so that a "coalition" consisting of 1 country came into existence, such a coalition would decide the strategy of only one country and thus correspond to a non-cooperative player, whose best response to others' Pollution would be to Pollute (see the proof of Proposition 1, part i). Given this, there is no incentive to join in Stage 1.

External stability requires that $U_n(0,0) \geq U_s(1,1)$, which was verified above. Internal stability is not an issue here, since no country is a signatory and a coalition of -1 countries is not feasible. ■

Proof of Proposition 6:

Proof. Assume $k = N$. Then in Stage 3, there are by assumption no non-signatories.

In Stage 2, the coalition of N countries prefers to Abate if $U_s(1,N) \geq U_s(0,N)$. Consider first the case where a coalition of size $k = N-1$ would prefer to abate. No individual signatory is then pivotal in the sense that its participation is decisive for the coalition's policy, and every signatory's kindness can be expressed as in the non-cooperative case, by eq. (10). Thus, the coalition will abate if

$$bN - c \geq -\alpha$$

which always holds since, by assumption, $bN - c > 0$ and $\alpha > 0$.

In Stage 1, country i will then join if, given the expectation that everyone else joins, it can do no better than joining. Proposition 1, part ii) demonstrates that if $N-1$ others abate and countries play non-cooperatively, then country i can do no better than abate too, given that $\alpha > 2(c-b)$ (which is assumed in the current Proposition). Hence, with the expectation that $N-1$ others join and the coalition abates, country i can do no better than abating too, which is equivalent to joining in Stage 1.

What if a coalition of size $N-1$ is not expected to abate in Stage 2? Every individual signatory i would then be pivotal in the sense that given everyone else's strategy and beliefs, its participation is decisive for the coalition's policy. In that case, if i joins, everyone else gets a payoff of $bN - c$, while if it does not join, everyone else gets a payoff of 0. The equitable payoff is then, due to eq. (5),

$$\pi_{ij}^e = \frac{1}{2}(bN - c) \tag{17}$$

and according to eq. (4), i's kindness if joining is given by

$$f_{sj} = \frac{(bN-c) - \frac{1}{2}(bN-c)}{(bN-c)} = \frac{1}{2} \tag{18}$$

and if not joining

$$f_{nj} = \frac{0 - \frac{1}{2}(bN-c)}{(bN-c)} = -\frac{1}{2}$$

which means that even a pivotal country's kindness is given by eq. (7), i.e. $f_{ij} = q_i - \frac{1}{2}$, and utility can be expressed by eq. (10). Hence the grand coalition,

if it exists, will abate in Stage 2. The rest of the analysis above thus goes through as before.

Note that country i will be indifferent between being a signatory and being a non-signatory that abates. Thus, any situation in which a share x of the N countries are signatories to an abating coalition and a share $1-x$ are non-signatories who abate is also stable. However, $x < 1$ would not affect the coalition's decision to abate in Stage 2 (due to Proposition 1, part ii), hence any situation in which $x < 1$ is equivalent to the case where $x = 1$ both in terms of outcomes and utilities.

The above establishes internal stability. External stability is not an issue here, since no country is a non-signatory and a coalition of $N+1$ countries is not feasible. ■

Proof of Proposition 7:

Proof. In Stage 3, non-signatories play non-cooperatively and thus have the same influence on others as in the non-cooperative game. It follows that the kindness of a non-signatory i towards any other country j can be expressed as in eq. (7): $f_{ij} = q_i - \frac{1}{2}$.

Turn then to Stage 2. For a *given* abatement policy of the coalition, a signatory's influence on others' payoff goes solely through the country's own contribution to the coalition's abatement, chosen implicitly when deciding in Stage 1 whether to join. Hence, for a non-pivotal signatory i, kindness to any other country j is also given by $f_{ij} = q_i - \frac{1}{2}$ (where q_i is determined by the coalition's policy).

Consider now the case where a coalition of k members is abating, and where, given everyone's strategies and beliefs, the loss of one member would have made the coalition pollute. Every individual signatory i is then pivotal in the sense that given everyone else's strategy and beliefs, i's participation is decisive for the coalition's policy in Stage 2. Assume further that non-signatories are expected to pollute in Stage 3. In this case, if i joins, every other signatory gets a payoff of $bk-c$, while every non-signatory gets a payoff of bk. If i does not join, everyone else gets a payoff of 0. The equitable payoff for other signatories would then, according to eq. 5, be

$$\pi_{is}^{e} = \frac{1}{2}(bk - c) \tag{19}$$

and for non-signatories

$$\pi_{in}^{e} = \frac{1}{2}(bk) \tag{20}$$

Using this and eq. (4), i's kindness to another signatory if joining is thus given by

$$f_{ss} = \frac{(bk-c) - \frac{1}{2}(bk-c)}{(bk-c)} = \frac{1}{2} \tag{21}$$

and if not joining

$$f_{ns} = \frac{0 - \frac{1}{2}(bk-c)}{(bk-c)} = -\frac{1}{2}$$

Moreover, i's kindness to a non-signatory if joining is given by

$$f_{sn} = \frac{bk - \frac{1}{2}bk}{bk} = \frac{1}{2} \tag{22}$$

and if not joining,

$$f_{nn} = \frac{0 - \frac{1}{2}(bk)}{(bk)} = -\frac{1}{2}$$

Consequently, even for a pivotal signatory to an abating coalition, kindness can be expressed as $f_{ij} = q_i - \frac{1}{2}$. As a result, the reciprocity function (eq. 9) and utility function (eq. 10) can be applied as before.

In the situation described in the Proposition, non-signatories pollute in Stage 3. For a signatory, we will thus have $Q_{-i} = k - 1$ if the coalition abates and $Q_{-i} = 0$ if the coalition pollutes. A coalition of $k < N$ members will abate in Stage 2 if $U_s(1, k) \geq U_s(0, k)$.

Using eq. (10), this implies

$$\begin{aligned} bk - c + \frac{3}{2}\alpha(\frac{k-1}{N-1} - \frac{1}{2}) &\geq -\frac{1}{4}\alpha \\ k &\geq \frac{2c(N-1) + \alpha(N+2)}{2b(N-1) + 3\alpha} \end{aligned} \tag{23}$$

Define $\underline{k}$ as the coalition size making the coalition exactly indifferent between polluting and abating in Stage 2, i.e. $U_s(0, \underline{k}) = U_s(1, \underline{k})$, or

$$\underline{k} = \frac{2c(N-1) + \alpha(N+2)}{2b(N-1) + 3\alpha} \tag{24}$$

The coalition will abate in Stage 2 if $k \geq \underline{k}$. k^1 is defined as the smallest integer such that $k^1 \geq \underline{k}$. Thus, in Stage 2, a coalition of k^1 countries will abate, but a coalition of $k^1 - 1$ will not.

In Stage 1, a country will join if, given the expectation that $k - 1$ others join, it can do no better than joining; that is, $U_s(1, k) \geq U_n(0, k - 1)$. Consider a country that expects $k^1 - 1$ others to join. Since the coalition will abate when $k = k^1$, the utility of each signatory if it joins is

$$U_s(1, k^1) = bk^1 - c + \frac{3}{2}\alpha(\frac{k^1 - 1}{N-1} - \frac{1}{2}). \tag{25}$$

If the country does not join, the coalition will consist of $k^1 - 1$ signatories and will not abate, and each non-signatory's utility is

$$U_n(0, k^1 - 1) = -\frac{1}{4}\alpha \tag{26}$$

The country will thus join if $U_s(1, k^1) \geq U_n(0, k^1 - 1)$, i.e.

$$bk^1 - c + \frac{3}{2}\alpha(\frac{k^1 - 1}{N-1} - \frac{1}{2}) \geq -\frac{1}{4}\alpha \tag{27}$$

which is exactly the same problem as considered in eq. (23). Thus, the above inequality holds if $k^1 \geq k$, which holds by definition. That is, if i expects $k^1 - 1$ others to join, i can do no better than joining. Hence, a coalition of k^1 members is internally stable.

External stability requires that for $k = k^1$, no non-signatories want to join. The coalition abates regardless of whether $k = k^1$ or $k = k^1 + 1$. A country that expects k^1 others to join will join if $U_s(1, k^1 + 1) \geq U_n(0, k^1)$. Using eq. (10), this would imply

$$\begin{aligned} b(k^1+1) - c + \frac{3}{2}\alpha(\frac{k^1}{N-1} - \frac{1}{2}) &\geq bk^1 + \frac{1}{2}\alpha(\frac{k^1}{N-1} - \frac{1}{2}) \\ \alpha(\frac{k^1}{N-1} - \frac{1}{2}) &\geq c - b \end{aligned} \tag{28}$$

Since $c > b$ and $\alpha > 0$, the above can only hold if $\frac{k^1}{N-1} \geq \frac{1}{2}$, or $k^1 \geq \frac{N-1}{2}$. However, this cannot be the case, given the assumptions of the Proposition.

To see this, note that $\underline{k}$ can be characterized as follows. First, if $b(N+2) \geq 3c$ (or $c/b \leq (N+2)/3$), $\underline{k}$ is increasing in α:

$$\begin{aligned} \frac{\partial \underline{k}}{\partial \alpha} = & \frac{(N+2)(2b(N-1)+3\alpha) - 3(2c(N-1)+\alpha(N+2))}{(2b(N-1)+3\alpha)^2} \\ & \frac{(N+2)(2b(N-1)+3\alpha) - 3((2c+\alpha)(N-1)+3\alpha)}{(2b(N-1)+3\alpha)^2} \end{aligned} \tag{29}$$

i.e., $\frac{\partial \underline{k}}{\partial \alpha} > 0$ iff

$$\begin{aligned} (N+2)(2b(N-1)+3\alpha) - 3(2c(N-1)+\alpha(N+2)) &> 0 \\ b(N+2) &> 3c. \end{aligned} \tag{30}$$

Second, when α goes to infinity, $\underline{k}$ goes to $\frac{N+2}{3}$:

$$\lim_{\alpha \to \infty} \underline{k} = \lim_{\alpha \to \infty} \frac{2c(N-1)/\alpha + (N+2)}{2b(N-1)/\alpha + 3} = \frac{N+2}{3} \tag{31}$$

Thus, $\frac{N+2}{3}$ is an upper boundary for $\underline{k}$ under the given assumptions. Since k^1 is the smallest integer weakly larger than $\underline{k}$, the upper boundary for k^1 is $\frac{N+2}{3} + 1 = (N+5)/3$. The question is whether we can have $k^1 \geq \frac{N-1}{2}$. This is only possible if N is relatively small:

$$\begin{aligned} \frac{N+5}{3} &\geq \frac{N-1}{2} \\ 13 &\geq N \end{aligned}$$

Consequently, under the given assumptions, $k^1 < \frac{N-1}{2}$, which means that eq. (28) cannot hold. Thus, a coalition of size k^1 is internally and externally stable.

Finally, recall that k^0 is the smallest integer weakly larger than c/b. Since k^1 is the smallest integer such that $k^1 \geq \underline{k} > c/b$ (see above), we must have $k^1 \geq k^0$. ■

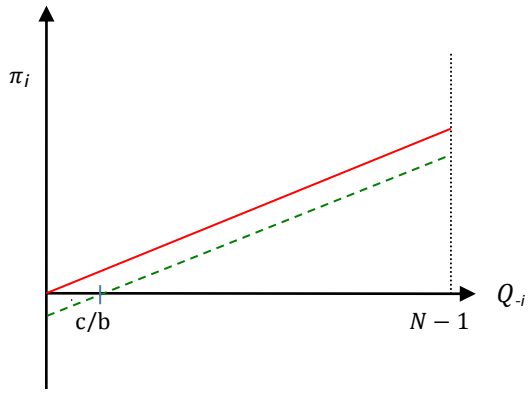

Figure 1: Payoff of country *i*, given that Q_{-i} others abate.
Red solid line: payoff if *i* pollutes.
Green dashed line: payoff if *i* abates.

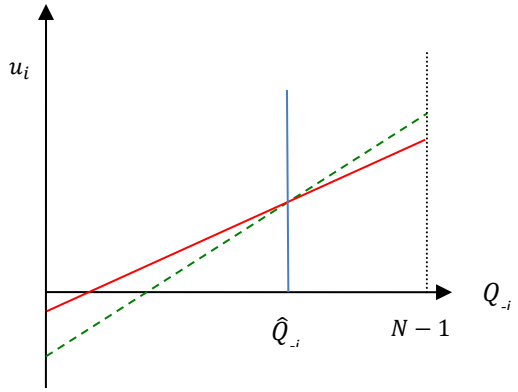

Figure 2: Utility of a reciprocal country *i*, given that Q_{-i} others abate.
Red solid line: utility if *i* pollutes.
Green dashed line: utility if *i* abates.

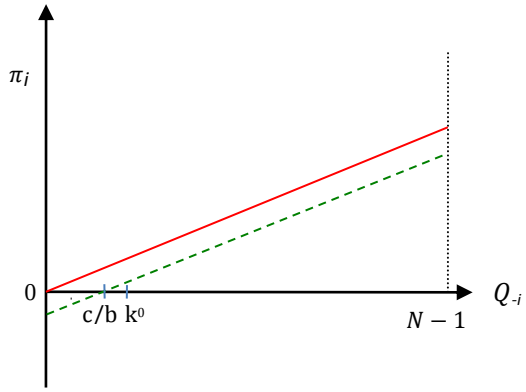

Figure 3: Payoff of country *i* (standard preferences), given that Q_{-i} others abate.
Red solid line: payoff if *i* pollutes.
Green dashed line: payoff if *i* abates.
c/b: the minimum *k* for which the coalition prefers to abate.
k^0: the smallest integer weakly larger than *c/b*.

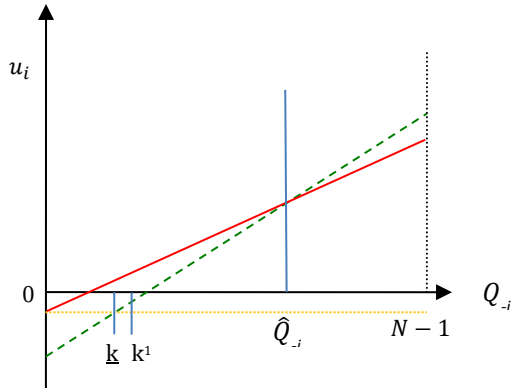

Figure 4: Utility of a reciprocal country *i*, given that Q_{-i} others abate.
Red solid line: Utility if *i* pollutes.
Green dashed line: Utility if *i* abates.
Orange dotted line: Utility if no one abates.
k : the minimum *k* for which the coalition prefers to abate.
k^1: the smallest integer weakly larger than *k*.